TRANSFORMATIONS

One of the assumptions of using ANOVA to test for significance is that the errors should be independently and normally distributed.

Randomization is used to break up any correlation of experimental units.

A problem that may influence this assumption is that the errors may be heterogeneous.

There are two types of heterogeneity.

1. Irregular: certain treatments possess considerably more variability that others. e.g. In insecticide trials, the checks may contain considerably more insects that the treated experimental units; therefore, the checks contribute to the Error MS to a larger degree than the treated units. Consequently, the standard deviation will be too large for comparisons among treated experimental units.

   This portion of the experiment is not under statistical control.

   The best procedure to compensate for this problem is to omit certain portions of the data from the analysis or use orthogonal contrasts.

2. Regular: arises from some type of non-normality of the data in the experiment.

   This non-normality is caused by a relationship between the variability of several treatments and the mean.

   To correct the problem, the data can be transformed such that the transformed errors are normally distributed.

<u>Ways the Mean and Variance Can Be Related</u>

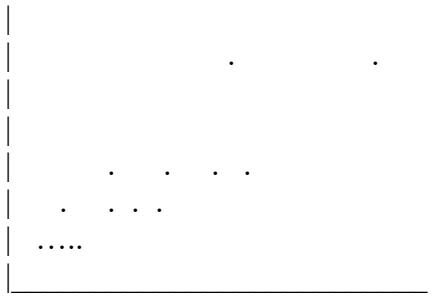
1. Count data: e.g. number of infested plants per plot, number of lesions per leaf, etc. These type of data may follow a Poisson distribution where the mean equals the variance.
2. Binomial data: data in which only two outcomes are possible. For example, susceptible vs. non-susceptible, present vs. not present, etc.

<u>Detecting the Presence of Variability Heterogeneity for a CRD</u>

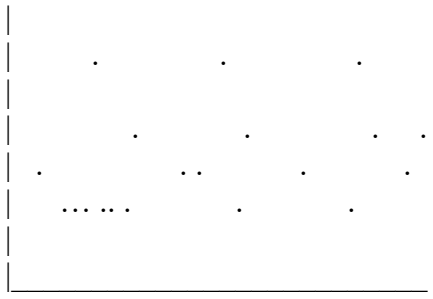
Step 1. For each treatment, compute the variance and the mean across replicates.

Step 2. Plot a scatter diagram of the treatment variances vs. the treatment means. The number of points in the scatter diagram equals the number of treatments.

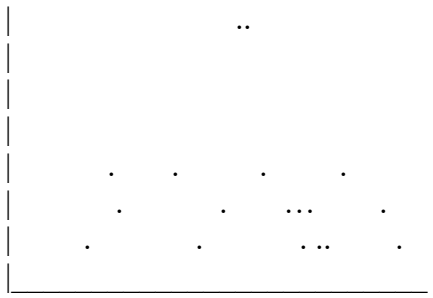
Step 3. Visually examine the scatter diagram to identify the patter of relationship if any.



Heterogeneous variance where the variance is proportional to the mean.



Homogeneous variance



Outliers – transformation will not work.

Causes of outliers

1. Mean(s) have high variability.
2. Errors in collecting data.

To determine if transformations are necessary for others designs, plot residual on the Y-axis and the predicted values on the X-axis.

Data Transformations

1. Logarithmic (Log_{10}) transformation

   Appropriate for data where the standard deviation is proportional to the mean.

   Helpful when the data are expressed as a percentage of change.

   These types of data may follow a multiplicative model instead of an additive model.

   If the data set includes small values (e.g. less than 10), use the transformation Log(Y+1) instead of Log Y (Y is the original data).

2. Square root transformation

   Useful for count data (data that follow a Poisson distribution).

   Appropriate for data consisting of small whole numbers.

   In both these cases the mean may be proportional to the variance.

   Examples are the number of infested plants per plot, the number of insects caught in a trap, the number of weeds per plot (i.e. data obtained in counting rare events).

   This transformation also may be appropriate for percentage data where the range is between 0 and 20% or between 80 and 100%.

   If most of the values in the data set are less than 10, especially if zeros are present, the transformation to use is $(Y+0.5)^{1/2}$ instead of $Y^{1/2}$.

3. Arc sine square root transformation - Arc Sine $(Y)^{1/2}$

   Appropriate for data on proportions, binomial data, and data expressed as percent of control.

   The value of 0% should be substituted by (1/4n) and the value 100% by (100-1/4n), where n is the number of units in which the percentage data were based (i.e. the denominator used in computing the percentage.

The following rules may be useful in choosing the proper transformation scale for the percentage data derived from count data.

Rule 1. For percentage data lying within the range of 20 - 80%, no transformation is needed.

Rule 2. For percentage data lying within a range of either 0 - 20% or 80 – 100%, but not both, the square root transformation could be useful.

Rule 3. For percentage data that do not follow the ranges specified in either Rule 1 or Rule 2 (e.g. percent control data), the Arc Sine square root transformation may be useful.

Determining if a Transformation is Needed

Perform the ANOVA on untransformed data.

Check the residual vs. predicted value plots to determine if a transformation is needed.

If a transformation is needed, transform the data using the appropriate method.

Determine if the transformation corrected the problem of non-normality of the errors.

If the transformation did not correct the problem, then analyze and discuss the nontransformed data.

Performing an ANOVA Using Transformed Data

Perform the ANOVA using the transformed data.

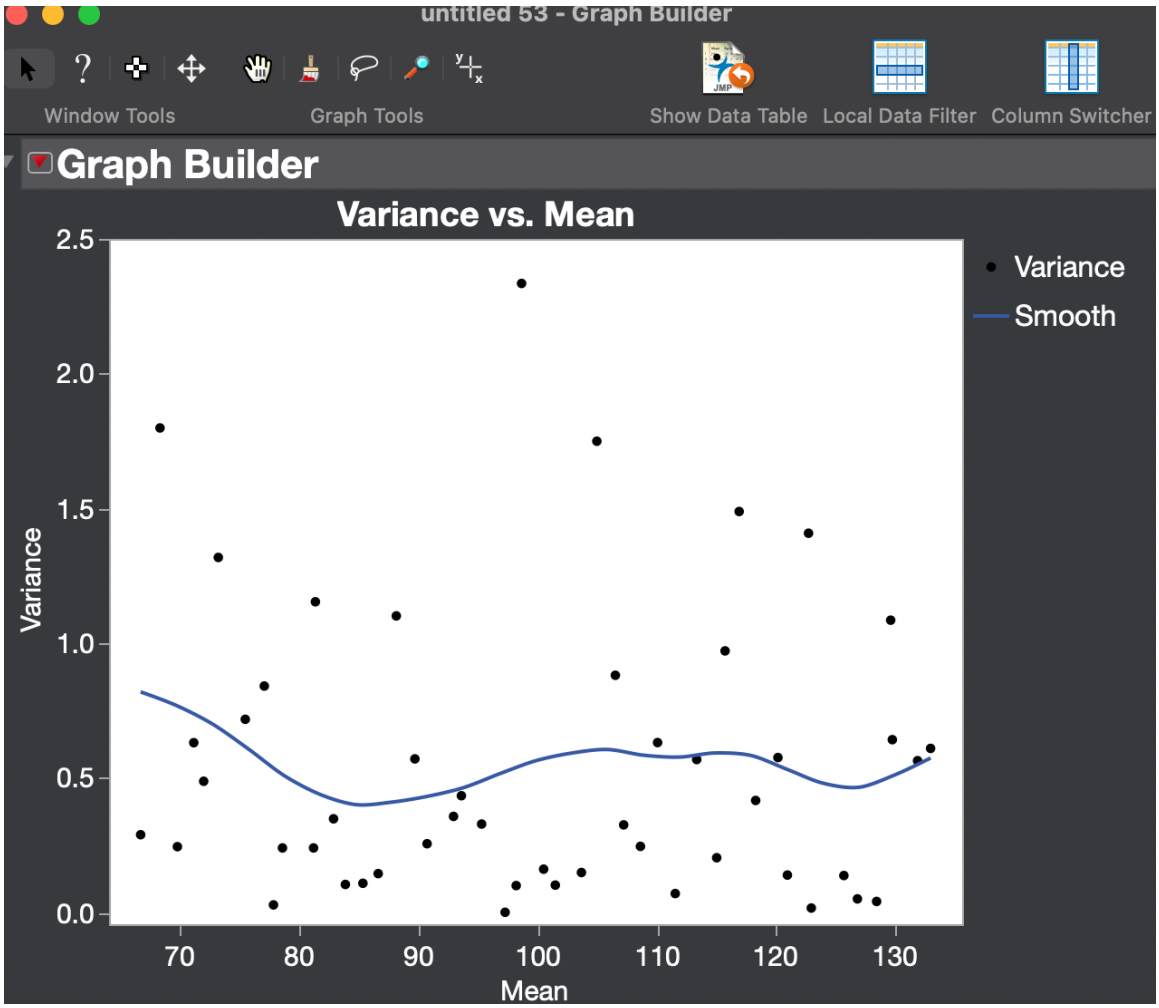
The LSD used to compare differences should be calculated using the transformed error mean square.

Mean separation should be done using the LSD calculated from the transformed data.

When presenting means, untransformed means can be used. However, somewhere in you presentation or paper, it should be mentioned that transformed data were used to perform the ANOVA.

JMP Example to Look at Residual Plots

- The example is based on the RCBD analysis of an experiment with 50 treatments.
- Plot the mean vs. the variance of the treatments to look for relationship between these two statistics.
- Data were first analyzed using *Distribution*, then the variance and the mean for each treatment was plotted using *Graph Builder.*

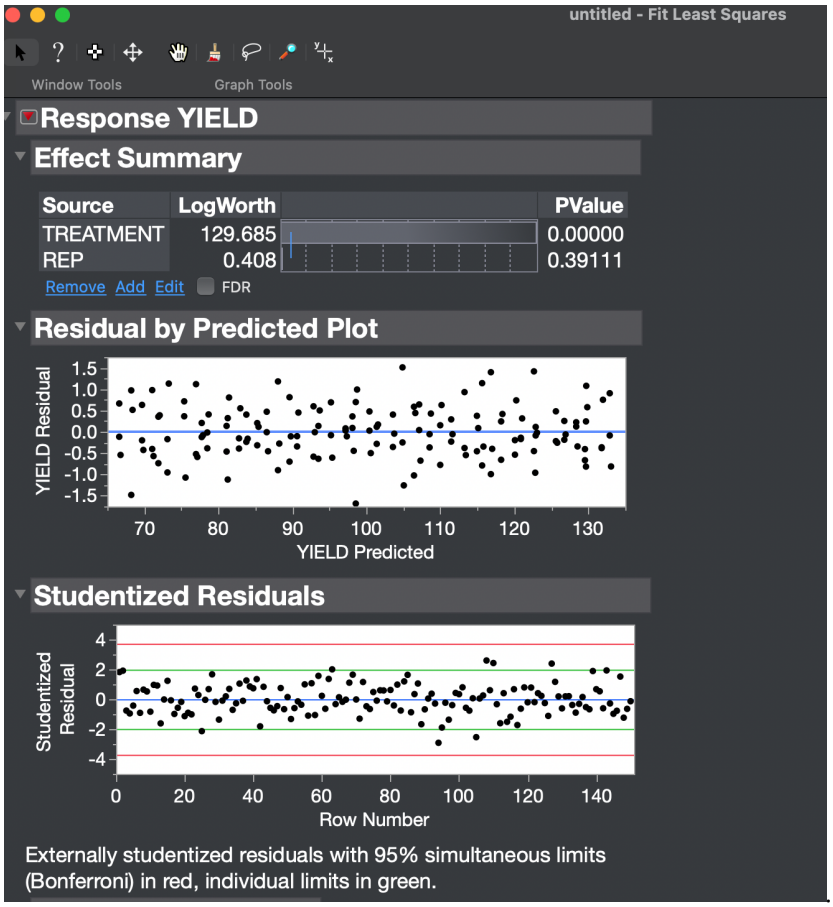


- **Conclusion: There is no obvious relationship between the mean and the variance.**

- The next plots that will be evaluated are the residual by the predicted value plot for each observation, and the studentized residuals for each observation.

- This plot will be generated using *Fit Model* with the Emphasis being *Effect Screening.*

- Note that there are 150 observations in this experiment (i.e. 50 treatments x 3 replicates)



- **Conclusion 1: There is no obvious relationship between the residuals and the predicted yield values.**

- **Conclusion 2: There are no obvious outliers in the Studentized Residual plot. Outliers would be values > |4|.**