## SOKO Vegetation Period and Pod Shattering Data Analysis

330 plant genotypes were tested for their vegetation period and pod shattering at 2 different locations and data was collected for the years 2018,2019,2020 and 2021.

However, the data is highly unbalanced due to a lot of missing values. To normalize the data, two different methods can be used.

1. **Impute the data using one single value across all the years.**

This can be done but are a lot of missing values and some genotypes don’t have a data at all and it would greatly affect the analysis and we don’t know the experimental conditions for each year.

1. **Complete case analysis**

In this analysis, genotypes with NAs are removed completely from the data frame which created a highly unbalanced data.

To normalize the data and to meet the assumptions of linear model, log transformation was applied to vegetation period data and arcsine square root transformation was applied to pod shattering data.

For pod shattering, percentages were noted which ranges from 0-100% thus log transformation, or any other transformation isn’t suitable for this type of data that’s why arcsine square root transformation is used which can deal with zero values.

**Data formatting**

Data was formatted and merged together in a long format and all NAs were removed. Formatted data files were then used for normalization.

1. **Vegetation Period**

Plots obtained with log transformed vegetation period values are attached below.

## Diagnostic plot for normality of residuals (Q-Q plot)

A graph of a normal q-q plot

Description automatically generated

## Diagnostic plot for constant variance (Residuals vs. Fitted values plot)

A graph of a number of dots

Description automatically generated

## Diagnostic plot for linearity (Observed vs. Fitted values plot)

A graph of a graph of a number of dots

Description automatically generated

**Linear model**

Linear model “log\_veg\_period ~ year\_location + (1|genotype)” was used to extract BLUPs.

1. **Pod shattering**

## Diagnostic plot for normality of residuals (Q-Q plot)

A graph with numbers and lines

Description automatically generated

## Diagnostic plot for constant variance (Residuals vs. Fitted values plot)

A graph of a graph with black and white dots

Description automatically generated

## Diagnostic plot for linearity (Observed vs. Fitted values plot)

**A graph of a graph with a line

Description automatically generated**

These graphs don’t look okay with the transformation that’s done but no other transformation works on this data as well.