# MODULE SECTION: STATISTICAL MODELLING

## Discussion Questions: Set B

**1. When should you use a model and a hypothesis test?**

* Statistical models are used to describe, test, or predict a relationship between variables, while a hypothesis test is used when you want to test an hypothesis (claim) about the relationship between variables. Most statistical models are embedded with some form of hypothesis even though not explicitly stated, and that is why there all the models run significance tests through p-values to determine whether there is enough evidence to justify or predict the relationship between variables. Therefore, hypothesis tests can be conducted within statistical models, however, depending on the nature of data you have and your data analysis goal, for example having two treatment groups and goal is to compare the yield between the treatment groups, your analysis could require a simple hypothesis test such as t-test. While in cases where your data has a more complex structure, such as blocking and more than two treatment groups, you will need to describe a model for how your yield data is influenced by the variables in your data and then run a test in line with the described model.

**2. Examples of each category of data in agriculture**

a) Binary Data (only two possible outcomes)

* Example: Whether a plant is diseased (1) or healthy (0).
* Example: Whether a farmer adopts a new technology (Yes/No).

b) Proportional Data (fractions or percentages)

* Example: The proportion of land covered by weeds in a farm.
* Example: The percentage of seeds that germinate under different treatments.

c) Binomial Data (counts of successes in repeated trials)

* Example: The number of successful crop harvests in 10 years.
* Example: The number of pest-resistant plants in a batch of 50.

d) Count Data (whole numbers, often representing frequencies)

* Example: The number of pests found per plant.
* Example: The number of crop failures per season.

e) Ordinal Data (ranked data with a meaningful order but unknown differences between levels)

* Example: Soil fertility classified as low, medium, or high.
* Example: Farmers’ satisfaction with an irrigation system rated as poor, fair, good, or excellent.

f) Multinomial Data (more than two categories without a natural order)

* Example: Crop variety chosen by a farmer (e.g., maize, wheat, rice, beans).
* Example: Types of pests found in a field (e.g., aphids, caterpillars, beetles).

**3. Roles and Underlying Assumptions of Statistical Models**

a) Simple Linear Regression Model

* Role: Measures the relationship between two continuous variables (e.g., how rainfall influences maize yield).
* Assumptions:
  1. Linearity – The relationship between independent (e.g., rainfall) and dependent (e.g., yield) variables is linear.
  2. Independence – Observations are independent of each other.
  3. Homogeneity of variances – Variability of residuals is constant across all levels of the independent variable.
  4. Normality – Residuals (differences between observed and predicted values) should be normally distributed.

b) Analysis of Variance (ANOVA) Model

* Role: Compares means across multiple groups (e.g., testing whether three different fertilizers produce different yields).
* Assumptions:
  1. Independence – Data points are independent within and across groups.
  2. Normality – Data in each group follow a normal distribution.
  3. Homogeneity of Variance (Homoscedasticity) – Variability within each group should be similar.

c) Chi-Square Test

* Role: Tests for associations between categorical variables (e.g., is there a relationship between soil type and crop disease presence?).
* Assumptions:
  1. Independence – Each observation falls into only one category.
  2. Expected frequencies – Each category should have a sufficiently large expected count (usually >5).