# MODULE SECTION: STATISTICAL MODELLING

## Discussion Questions Set A

**1. What is a statistical model, and how is it different from a conceptual model?**

* A statistical model is a mathematical representation of relationships between variables in a dataset based on some statistical assumptions on how the data was generated. On the other hand, a conceptual model is a theoretical framework that outlines the relationships or processes in a system based on prior knowledge or hypotheses. It is more abstract and does not involve specific equations or statistical tools. Instead, it highlights key variables and their interactions.

**Key Difference**:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Statistical models** | **Conceptual models** |
| Nature | Quantitative, based on data | Qualitative or mixed, based on theory |
| Purpose | Predicting, testing, and analyzing data relationships | Understanding system components and interactions |
| Representation | Equations based on probability assumptions | Diagrams, flowcharts, descriptive frameworks |
| Relationship | Quantitatively tests or fits the relationships defined by the conceptual model using data. | Guides the data collection/generation using the defined relationship between variables |
| Example | Linear regression predicting crop yield | Agroecosystem framework showing nutrient flows |

**2. How do you choose the right statistical model for your data?**

1. Selecting the appropriate statistical model could involve the following steps:
2. Understand the Data:
   * Determine the type of variables (categorical, continuous, ordinal) in your dataset.
   * Assess the data distribution (e.g., normal, skewed, count data) through data exploration techniques
3. Define Your Data Analysis Goal:
   * Are you predicting an outcome, estimating relationships, or testing a hypothesis?
4. Examine the Relationships between Variables:
   * For linear relationships, models like linear regression might work.
   * For non-linear or complex relationships, consider generalized linear models, non-linear regression, or machine learning methods.
5. Start Simple:
   * Start with simpler models (e.g., linear regression) before exploring more complex ones.
6. Check Assumptions:
   * Match the model assumptions (e.g., normality, homoscedasticity, independence) to your data. If the model assumptions are not met by your data, you need to work on adjusting your variables or change the model.
7. Use Model Selection Tools:
   * Alternative to point 5 above, you can employ criteria like Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or cross-validation to compare models.
8. Compare Model with Domain Knowledge:
   * Ensure the model aligns with theoretical expectations and conceptual understanding.

**3. What do you do if the statistical model you have chosen is incorrect for your data?**

If the chosen model is unsuitable, for example, your data does not match the model assumptions, you can use the following steps to address the problem:

1. Identify the Issue:
   * Diagnose the problem using residual plots or goodness-of-fit tests, or cross-validation errors.
   * Common issues include poor fit, violation of assumptions, or overly complex models.
2. Reassess Assumptions:
   * Check if assumptions (e.g., linearity, normality) are valid.
   * If assumptions are violated, consider transforming variables or using alternative models (e.g., generalized linear models).
3. Simplify or Adjust:
   * Try simpler models or remove unnecessary predictors.
   * If multicollinearity is an issue, consider dimensionality reduction techniques like PCA.
4. Use Robust or Flexible Models:
   * Consider non-parametric methods if assumptions about data distribution are too restrictive.
5. Iterate:
   * Modify the model based on diagnostic findings and re-fit it to the data.
6. Consult an Expert:
   * If unsure, collaborate with a statistician for guidance.