STA 3180 Statistical Modelling: Missing Data

Missing Data Lecture Notes for STA 3180 Statistical Modelling

Key Concepts:

Missing data is a common issue in statistical analysis and can be caused by a variety of factors such as respondent refusal, survey design, or data entry errors. It is important to understand the different types of missing data and the implications of each type in order to accurately analyze data sets.

Definitions:

Missing Completely at Random (MCAR): Missing data that occurs completely at random, meaning that the probability of a value being missing is unrelated to any other variables in the dataset.

Missing at Random (MAR): Missing data that occurs randomly, but is related to other variables in the dataset.

Missing Not at Random (MNAR): Missing data that is not random and is related to the values of other variables in the dataset.

Rules:

1. When dealing with missing data, it is important to identify the type of missing data (MCAR, MAR, or MNAR) in order to determine the best approach for analysis.

2. If the missing data is MCAR, then the data can be analyzed without any special considerations.

3. If the missing data is MAR or MNAR, then special considerations must be taken in order to accurately analyze the data.

4. If the missing data is MAR, then multiple imputation can be used to fill in the missing values.

5. If the missing data is MNAR, then maximum likelihood estimation can be used to estimate the missing values.

Examples:

Example 1:

A researcher is conducting a survey on student performance in school. The survey includes questions about the student's gender, age, and grade level. Some students do not answer the question about their gender, which is considered missing data. In this case, the missing data is MCAR since the probability of a student not answering the question about their gender is unrelated to any other variables in the dataset. Therefore, the data can be analyzed without any special considerations.

Example 2:

A researcher is conducting a survey on student performance in school. The survey includes questions about the student's gender, age, and grade level. Some students do not answer the question about their

grade level, which is considered missing data. In this case, the missing data is MAR since the probability of a student not answering the question about their grade level is related to other variables in the dataset (e.g. age). Therefore, multiple imputation must be used to fill in the missing values in order to accurately analyze the data.