

Missing Value Analysis

Dr. Tariq Mahmood

Associate Professor – IBA Karachi

Head of BDA Laboratory

Missing Values

- Relatively common
- Can have a significant effect on the conclusions
- Missing due to various reasons: Equipment failure, data lost in transit, unsatisfactory data, data entry errors
- Concept of a probability distribution
- Normal distribution
- Assumptions/hypotheses
- Observed data vs. Missing data

Missing Values

- Missing completely at random (MCAR) – prob. that the missing data is not related to either the value that we want to fill in, or the observed responses which are already there
 - Probability of a missing data value is independent of any observation in the data set.
 - Analysis remains completely un-biased
 - We have to analyze our business problem to understand whether the data is MCAR or not

Missing Values

- Missing at Random (MAR): prob. that the missing data is not related to the value that we want to fill in, but is depending on the observed responses which are already there
 - Analysis remains un-biased to some extent
- Missing Not at Random (MNAR): If the above two situations are not applicable, then MNAR
 - Analysis will be quite biased towards the data

Handle Missing Data

- List-wise or case deletion
 - Most frequent, even in statistical packages
 - May introduce bias
 - If MCAR, then this is best. Otherwise, biasness may occur
 - More the data, the easier it becomes to delete
- Pair-wise deletion
 - When data required to test some business assumption is missing, delete that
 - Fill-up the other missing data using some technique
 - Less biased for MCAR and MAR

Handle Missing Data

- Mean Substitution
 - Mean is a reasonable estimate of the average behavior
 - May not work with values that are not completely random
 - May not work if different situations lead to different number of missing values
 - Not preferred
- Regression Imputation
 - Replace with estimated values
 - Replacing with a probable value estimated by other available information.
 - Imputation retains a great deal of data over the list-wise or pair-wise deletion
 - Imputation avoids significantly altering the standard deviation or the shape of the distribution.

Handle Missing Data

- Last Observation Carried Forward
 - Applicable for time-series data
 - strongly assumes that the value of the outcome remains unchanged by the missing data → but this is not very likely
 - Therefore, not recommended
- Maximum Likelihood and Expectation Maximization
 - Assume that the observed data are a sample drawn from a multivariate normal distribution
 - After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.
 - Much preferred

Handle Missing Data

- Multiple Imputation
 - Replace with a set of plausible missing value
 - Multiple datasets, each of them analyzed separately
 - Caters for uncertainty and gives more breadth to solving the problem
- Sensitivity Analysis
 - How the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs
 - When analyzing the missing data, additional assumptions on the reasons for the missing data are made, and these assumptions are often applicable to the primary analysis.
 - However, the assumptions cannot be definitively validated for the correctness.