**Missing data Mechanism**

**MCAR, MAR, MNAR**

1. Missing Completely at Random, MCAR

MCAR means there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others.

Little's chi-square statistic for testing whether values are missing completely at random (MCAR) can be used. If the p-value is less than 0.05, the data are not missing completely at random

1. Missing at Random, MAR

MAR means there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data.

Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual’s observed variables.

Example, if men are more likely to tell their weight than women, weight is MAR.

1. Missing Not at Random, MNAR

MNAR means there is a relationship between the propensity of a value to be missing and its values.

Example people with the lowest education are missing on education or the sickest people are most likely to drop out of the study.

There are no easy ways of dealing with this type of missing data. One have to check the data collection process. May be set some rules that not > 10% are missing…

**What do we do?**

Usually we do complete case analysis, a.k.a. listwise deletion.

This restricts analyses to individuals with observed data.

This is not good as this assumes missing is **MCAR**.

Hence it leads to biased and under-powered results

**Some Single Imputation Methods**

There are several single imputation methods in use.

1. One is using sample mean: the missing value as the average of the observed values.
2. Regression prediction: the missing value is imputed based on a trend value
3. “Hot-Deck”: For a missing case, find cases with the same observed values on other variables and then choose one randomly
4. Predictive matching: this combines methods (2) and (3)
5. Use a dummy variable, with values of 0 and 1 for missing and nonmissing and use this variable in the model.
6. Intention to treat (ITT): last observation carried forward.
7. Modified Intention to treat (mITT): at least two measurements

**Some weaknesses in Single Imputation**

This can be reasonable only if <5% of data are missing– (Graham , 2008)

Often this results in overly precise estimates.

 Simply treats all values as observed values and does not take into account the uncertainty in the imputation method.

 Hence the results will have more significance and narrower confidence intervals.

**Other appropriate imputation methods**

1. *Maximum likelihood*

In some cases, Maximum Likelihood approach is feasible and when ML exist, this method works very well.

It directly maximizes the likelihood function, f(X,Y).

This method uses observed values, taking into account the missing values as well. Cases are usually weighted by the inverse probability of response.

ML approach is often used to deal with attrition.

1. *Multiple imputation (MI)*

In MI, missing values are filled multiple times by creating multiple (like 5, 10 etc) “complete” data sets.

Then analyses are done separately on each data set the results of which are combined across the data sets

MI takes into account the uncertainty in the imputations.

Total variance as a function of the within and between imputed variance

Only the missing values are imputed, the same data can be used for many analyses.

The mathematics behind combining



1. *Multiple imputation by Chained Equations (MICE)*

MICE, also known as “Fully conditional specification” or “Sequential regression multiple imputation”, fits model of each variable, conditional on all other variables

Model used depends on type of variable (continuous/binary/ ordinal)

Steps: Variables: X1 (Continuous), X2 (Binary) , X3 (Ordinal)

1. Do simple imputations to fill in missing values for X1, X2, X3
2. Using cases with observed X1, fit linear regression model of - predicting missing X1 using X2 and X3
3. Using cases with observed X2, fit logistic regression model of -predicting missing X2 using X1 and X3
4. Using cases with observed X3, fit proportional odds regression model of - predicting missing X3 using X1 and X2
5. Iterate Steps 2-4
6. Repeat step 5

MICE can easily work with large datasets and the models can more accurately reflect distribution of each variable. It allows bounds and incorporates restrictions on subpopulation.

**Example 1: DM patients data (n=16).sav**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | Age | Education | Income | Gender | No Drugs | Control |
| 1 | 38 | Secondary | Low | Female | >1 | Poor |
| 2 | 47 | High school | . | Male | >1 | Poor |
| 3 | 23 | . | High | Male | 1 | Good |
| 4 | 27 | High school | High | Male | >1 | . |
| 5 | 22 | High school | Low | Female | 1 | Poor |
| 6 | 34 | . | High | Male | . | Poor |
| 7 | 38 | Secondary | High | . | >1 | Good |
| 8 | . | Secondary | Medium | Male | 1 | . |
| 9 | 29 | High school | Medium | . | 1 | Poor |
| 10 | 27 | College | . | Female | >1 | Poor |
| 11 | 31 | Secondary | Medium | Male | 1 | . |
| 12 | . | College | Medium | Male | >1 | Poor |
| 13 | 36 | High school | Medium | Male | 1 | Poor |
| 14 | 42 | College | Medium | Male | >1 | . |
| 15 | 26 | . | High | Female | 1 | Poor |
| 16 | 54 | High school | Low | Female | >1 | Poor |

‘.’ indicates missing values



Data with missing values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | Age | Education | Income | Gender | No Drugs | Control |
| 1 | 38 | Secondary | Low | Female | >1 | Poor |
| 2 | 47 | High school | . | Male | >1 | Poor |
| 3 | 23 | . | High | Male | 1 | Good |
| 4 | 27 | High school | High | Male | >1 | . |
| 5 | 22 | High school | Low | Female | 1 | Poor |
| 6 | 34 | . | High | Male | . | Poor |
| 7 | 38 | Secondary | High | . | >1 | Good |
| 8 | . | Secondary | Medium | Male | 1 | . |
| 9 | 29 | High school | Medium | . | 1 | Poor |
| 10 | 27 | College | . | Female | >1 | Poor |
| 11 | 31 | Secondary | Medium | Male | 1 | . |
| 12 | . | College | Medium | Male | >1 | Poor |
| 13 | 36 | High school | Medium | Male | 1 | Poor |
| 14 | 42 | College | Medium | Male | >1 | . |
| 15 | 26 | . | High | Female | 1 | Poor |
|  |  |  |  |  |  |  |
| 566 | 54 | High school | Low | Female | >1 | Poor |



Some imputed cases

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | Imputation | Age | Education | Income | Gender | No Drugs | Control |
| 1 | 1 | 38 | Secondary | Low | Female | >1 | Poor |
| 2 | 1 | 47 | High school | Low | Male | >1 | Poor |
| 3 | 1 | 23 | College | High | Male | 1 | Good |
| 4 | 1 | 27 | High school | High | Male | >1 | Good |
|  |  |  |  |  |  |  |  |
| No | Imputation | Age | Education | Income | Gender | No Drugs | Control |
| 1 | 2 | 38 | Secondary | Low | Female | >1 | Poor |
| 2 | 2 | 47 | High school | Medium | Male | >1 | Poor |
| 3 | 2 | 23 | College | High | Male | 1 | Good |
| 4 | 2 | 27 | High school | High | Male | >1 | Poor |
|  |  |  |  |  |  |  |  |
| No | Imputation | Age | Education | Income | Gender | No Drugs | Control |
| 1 | 3 | 38 | Secondary | Low | Female | >1 | Poor |
| 2 | 3 | 47 | High school | High | Male | >1 | Poor |
| 3 | 3 | 23 | College | High | Male | 1 | Good |
| 4 | 3 | 27 | High school | High | Male | >1 | Good |

Results from analysis

Dependent: Control

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Original | Input 1 | Input 2 | Input 3 | Pooled |
| Age | .000 | -.006 | -.016 | -.019 | -.014 |
| College | 1 | 1 | 1 | 1 | -.741 |
| Secondary | -.113 | -.540 | -.774 | -.908 | -.164 |
| High school | .502 | .034 | -.144 | -.382 | .351 |
| High | 1 | 1 | 1 | 1 | 1 |
| Low | .778 | .514 | .314 | .224 | .351 |
| Medium | -.665 | -.668 | -.433 | -.450 | -.5171 |
| Female | 1 | 1 | 1 | 1 | 1 |
| male | .832 | 1.107 | 1.288 | 1.045 | .147 |
| 1 drug | 1 | 1 | 1 | 1 | 1 |
| > 1 drug | 1.869 | 1.902 | 1.920 | 1.820 | 1.881 |
| Constant | -.193 | 0.315 | -.016 | .755 | .348 |