

MISSING DATA in medical research

How to handle it?



Outline

1. Importance of understanding why the data are missing, investigating the extent and patterns of missing data, and the implications of this for the analysis.

2. Simple solutions that sound attractive, but usually are not.

3. Recommended approaches.

Understand the reasons behind missing data

- No study of any complexity manages to collect all the intended data
 - From a general standpoint, missing values may cause two problems: **loss of efficiency or information**, and **bias**
 - The extent of information loss is not directly linked to the proportion of incomplete records
- **eg. when estimating the prevalence of a rare disease, a small proportion of missing observations could have a disproportionate impact on the resulting estimate**

Understand the reasons behind missing data

- It is important to understand the reasons behind missingness, albeit the mechanism causing the missing data can rarely be definitely established
- Why? For handling the remaining data correctly. If values are missing completely at random, the data sample is likely still representative of the population. But if the values are missing systematically, analysis may be biased.

1. Missing completely at random (MCAR)

- The probability of a missing value neither depends on any observed data, nor unobserved data
- **If respondent or interviewer *accidentally* skips a question on alcohol consumption, the missing data on alcohol consumption data are MCAR.**
- Under MCAR, cases with missing data are a simple random subsample from all cases in the data, and analysis remains unbiased. However, information has been lost.

2. Missing at random (MAR)

- Missingness is not random, but it can be fully accounted for by variables where there is complete information, ie. missing data depend on observed data but not on unobserved data
- **If females skip a question about alcohol consumption more frequently than males, and gender is observed for all respondents, then missing data on alcohol consumption are MAR.**
- **Males are less likely to fill in a depression survey but this has nothing to do with their level of depression, after accounting for gender.**

3. Missing not at random (MNAR)

- The probability of missingness depends on data that are not observed
- If respondents with higher alcohol consumption are more inclined to skip a question about average number of doses per week compared to those with normal / low consumption, the missing data on alcohol consumption are MNAR.
- Men failed to fill in a depression survey *because* of their level of depression.
- If the sickest people are most likely to drop out of the study.

Diagnosing the mechanism

- **MAR vs MNAR**

- Measure some of that missing data: follow-up on a paper survey with phone calls to a group of the non-respondents and ask a few key survey items. This allows you to compare respondents to non-respondents. If their responses on those key items differ by very much, that's good evidence that the data are MNAR.
- Examine patterns in the data get an idea of what's the most likely mechanism. Use *your substantive scientific knowledge of the data and your field*. The more sensitive the issue, the less likely people are to tell you.

Diagnosing the mechanism

- **MAR vs MCAR**

- Testing: create dummy variables for whether a certain variable is missing (1 = missing, 0 = observed). Run t-tests and chi-square tests between this variable and other variables in the data set to see if the missingness on this variable is related to the values of other variables
- For example, if women really are less likely to tell you their weight than men, a chi-square test will tell you that the percentage of missing data on the weight variable is higher for women than men
- The SPSS Missing Data module has a very nice procedure for doing this automatically

Main approaches to handle missing data

1. **Omission** - where samples with incomplete data are discarded from further analysis

2. **Analysis** - by directly applying methods unaffected by the missing values

3. **Imputation** - where values are filled in the place of missing data

1. Listwise deletion

- Default in many software
- Whether or not results will be biased after listwise deletion is dependent on the underlying mechanism that caused the missing data
- **Under MCAR, listwise deletion reduces the sample size and thus power, but does not give any biased results**
- **Under MAR, the leftover sample may not be representative of the total population anymore, consequently leading to biased results** (eg. in the alcohol consumption example, females are underrepresented in the leftover sample). As a result, statistical inferences from analyses that include gender or variables that are related to gender may be biased.
- **Under MNAR, listwise deletion has the same problem as it has under MAR,** because there are systematic differences between the deleted cases and the cases that are left. However, under MAR the causes of this systematic dropout can be traced, whereas under MNAR they cannot.

1. Listwise deletion

- Even if MCAR, can easily decrease the effective sample size
- This happens when an analysis includes many variables, and each is missing for a few unique cases.



Say you have a data set with 200 observations and use 10 variables in a regression model. If each variable is missing on the same 10 cases, you end up with 190 complete cases, 5% missing.



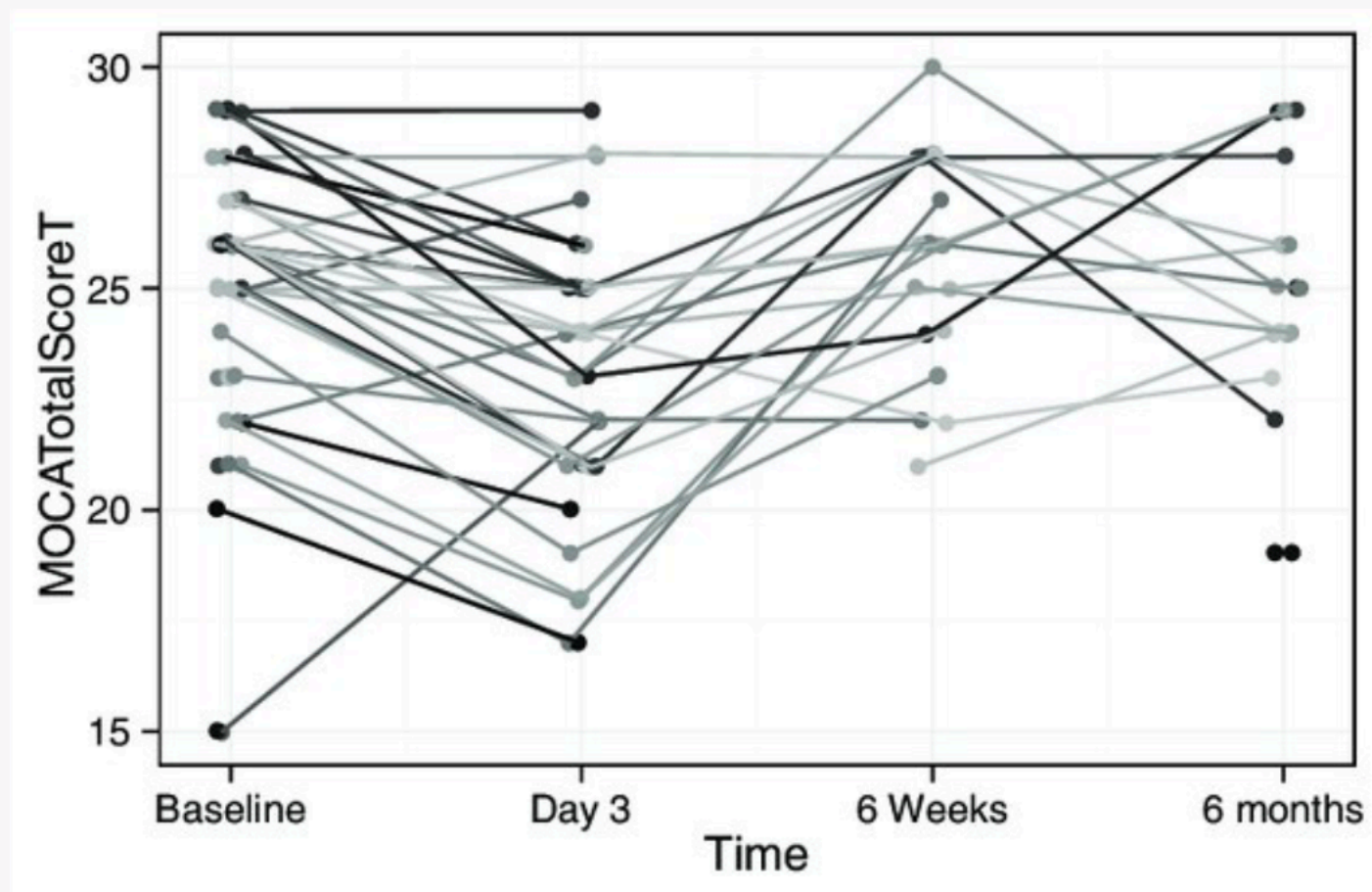
But if you have a *different* 10 cases missing on each variable, you will lose 100 cases (10 cases by 10 variables). With only 5% missing data, you end up with 100 complete cases, 50% missing. Not so good.

2. Analysis

- The ultimate aim of the missing data analysis is never to predict the exact values themselves, but rather to facilitate revealing the most important findings from an incompletely observed data at hand as precisely as possible
- Thus, methods that can handle missing values without any pre-processing procedures, are ideal
- Maximum likelihood based methods

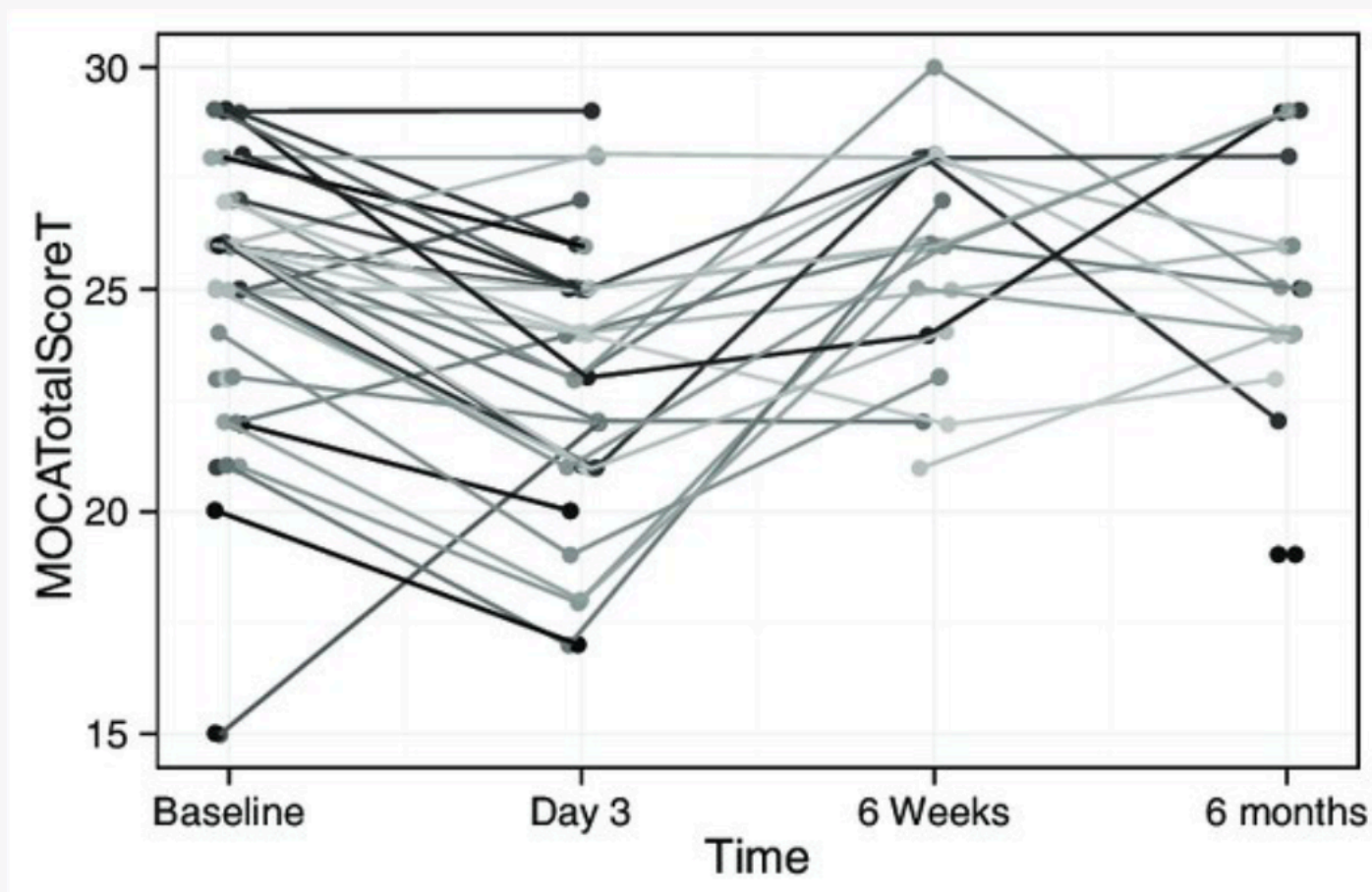
2. Analysis

- Example: missing data in longitudinal studies



Longitudinal changes in individual patients' MoCA scores (MoCA indicates Montreal Cognitive Assessment).

- **Repeated measures ANOVA** would handle missing data through listwise deletion, meaning keeping only patients with complete data = problems with power and bias
- **Linear mixed model** uses all available data. If the data are missing at random, this will give you unbiased estimates of each of these means.



Longitudinal changes in individual patients' MoCA scores (MoCA indicates Montreal Cognitive Assessment).

3. Imputation

“The idea of imputation is both seductive and dangerous.”

– *R.J.A Little & D.B. Rubin.*

3. Imputation

- Fill in plausible values for missing data
- A common misconception of missing data methods is the assumption that imputed values should represent “real” values
- The purpose when addressing missing data is to correctly reproduce the variance/covariance matrix we would have observed had our data not had any missing information

3. Mean imputation

- Unconditional mean imputation: missing values for an individual variable are replaced with its overall estimated mean from the available cases.

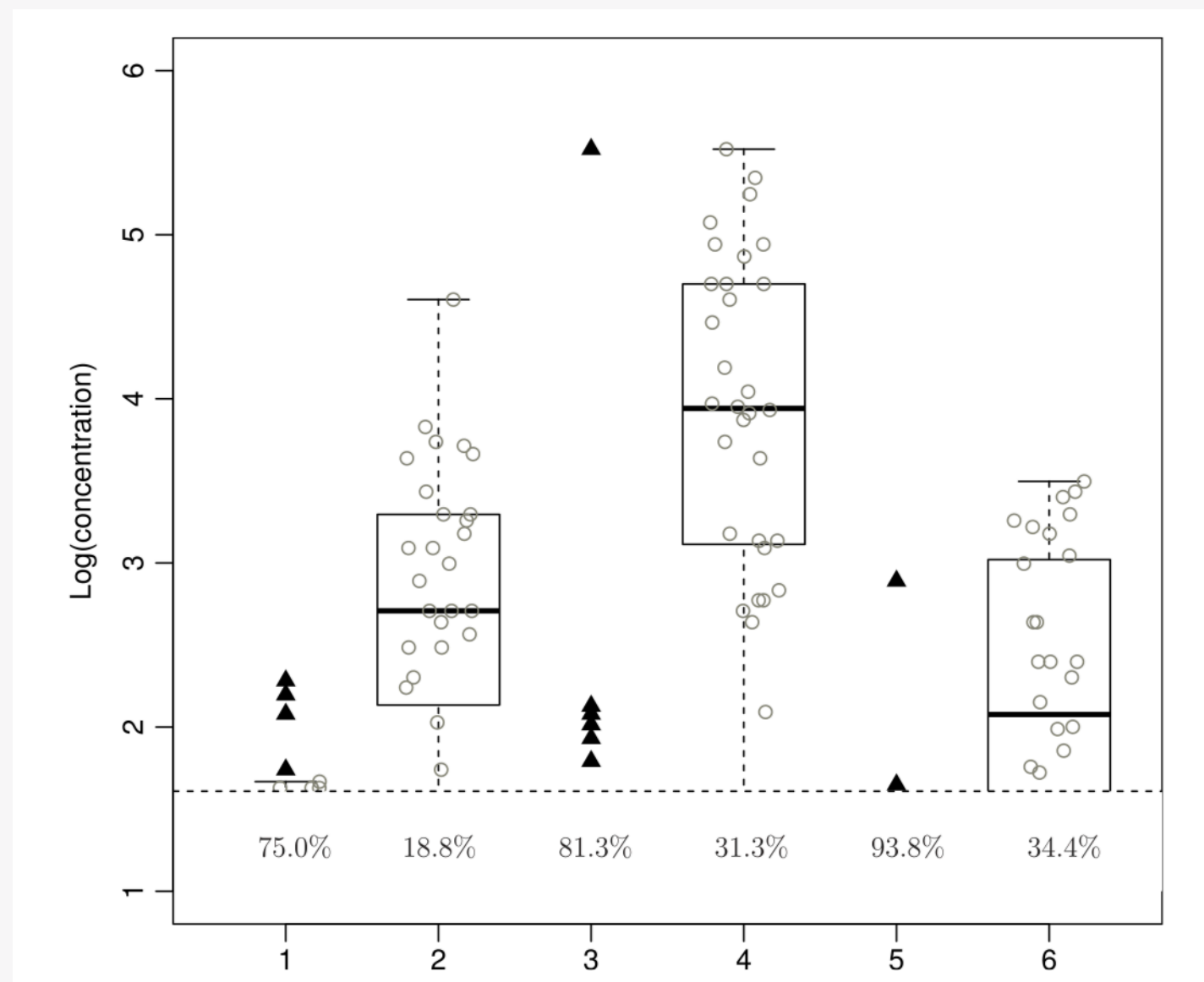


Simple and easily implemented method



Will result in an artificial reduction in variability due to the fact you are imputing values at the center of the variable's distribution. This also has the unintended consequence of changing the magnitude of correlations between the imputed variable and other variables.

Special case: values below lower limit of detection



Distributions of the six log-transformed toxic chemical compound concentrations. The dashed line is the LLOD set at $\log(5) = 1.61$. Gray dots are the uncensored concentrations and black triangles are the outliers. The percentages of the left-censored values for each compound are given below the boxplots.

Special case: values below lower limit of detection

- One commonly used approach is to substitute the left-censored values with a suitable constant and then analyse the resulting data as it was complete
- Potential substitution values include the sample mean or median of the uncensored values for the corresponding variable, zero, $LLOD/2$, or a minimum of the observed values.
- No matter how advanced the substitution method, single imputation does not reflect the full uncertainty created by missing data

3. Regression imputation

- Regression/conditional mean imputation, which replaces missing values with predicted scores from a regression equation.



Uses complete information to impute values

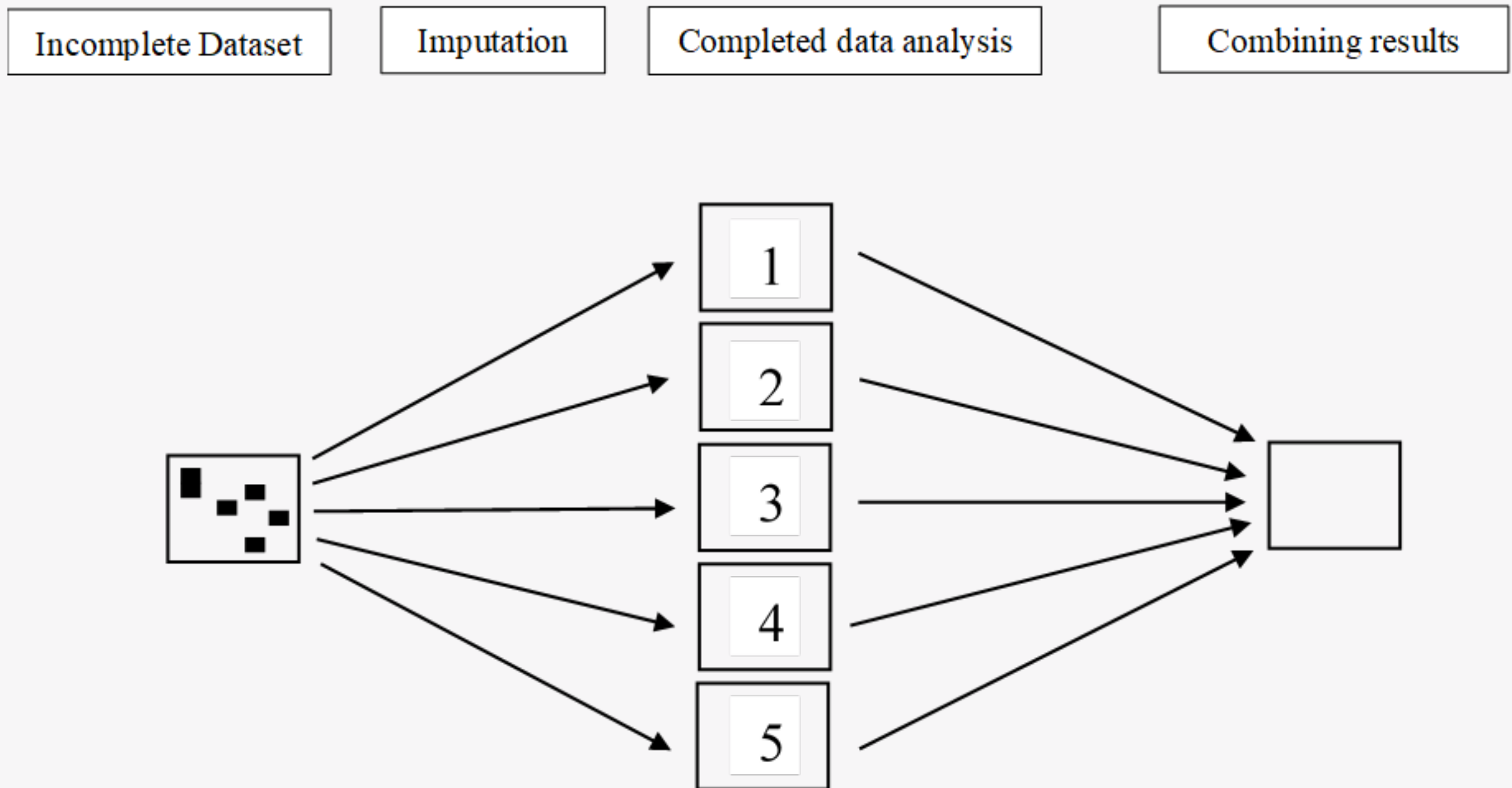


All predicted values will fall directly on the regression line once again decreasing variability, just not as much as with unconditional mean imputation

3. Multiple imputation

- MI estimates the values multiple times to capture the uncertainty
- It uses an imputation method with error built in, the multiple estimates should be similar, but not identical
- The statistical analysis of interest, such as ANOVA or logistic regression, is performed separately on each data set, and the results are then combined
- Because of the variation in the imputed values, there should also be variation in the parameter estimates, leading to appropriate estimates of standard errors and appropriate p-values

3. Multiple imputation



3. Under which missingness mechanisms does MI work?

- **MCAR:** Like listwise deletion, MI will give unbiased results under MCAR
- **MAR:** MI is capable of correcting the bias when the missing data are MAR. Consider the example where females tend to leave a question measuring someone's alcohol consumption open more often than males. If in the imputation model gender is used as a (dummy) predictor for the imputation of consumption, then the dependence of the missingness on gender is incorporated in the imputed values for income.

3. Under which missingness mechanisms does MI work?

- **MNAR:** When missingness depends on information that is not observed, unbiased results are no longer guaranteed. However, MI still picks up dependencies of the missing data on observed information, and dependencies on unobserved data can partly be accounted for by relations with observed variables that may also be related to the unobserved variables on which the missingness depends (auxiliary variables). **The more variables are included in the imputation model, the less residual dependence of the missingness on unobserved variables or information remains.**

3. MI software

- In SPSS: Analyze > Multiple Imputation > Impute Missing Data Values...
- Many procedures support pooling of results from analysis of multiply imputed datasets
- When imputation markings are turned on, a special icon is displayed next to procedures that support pooling
- List of supported methods:

<https://www.ibm.com/docs/en/spss-statistics/23.0.0?topic=imputation-analyzing-multiple-data>

3. MI software

			Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Imputation	Model		B	Std. Error	Beta		
0	1	(Constant)	.965	.091		10.564	.000
		X	.910	.074	.880	12.298	.000
1	1	(Constant)	.949	.066		14.409	.000
		X	.960	.056	.890	17.198	.000
2	1	(Constant)	1.080	.072		14.907	.000
		X	.868	.061	.848	14.133	.000
3	1	(Constant)	.950	.071		13.303	.000
		X	.972	.060	.876	16.071	.000
4	1	(Constant)	.994	.064		15.509	.000
		X	.896	.054	.882	16.489	.000
5	1	(Constant)	.929	.063		14.793	.000
		X	.925	.053	.891	17.366	.000
Pooled	1	(Constant)	.980	.095		10.371	.000
		X	.924	.074		12.418	.000

SPSS output of a regression analysis with a single predictor (X) and an outcome variable (Y) on a multiply imputed data set. The pooled results are shown in bold.

3. MI software

- In Stata: procedure **mi**
- Multiple impute data but also explore the patterns of missingness present in the data
- Tutorial:

https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/