Sensitivity Analysis for Missing Data

Sara Geneletti & Jose Pina-Sánchez

Introduction

- Most research datasets (outside lab conditions) are affected by missing data

Introduction

- Most research datasets (outside lab conditions) are affected by missing data
- Failing to sample sections of the population, i.e. coverage error
  - Common in big data, e.g. assessing changes in beauty standards using pictures in social media
  - Study on urban biodiversity from google map images (missing animals living indoors or in the wast system)

Introduction

- Most research datasets (outside lab conditions) are affected by missing data
- Failing to sample sections of the population, i.e. coverage error
  - Common in big data, e.g. assessing changes in beauty standards using pictures in social media
  - Study on urban biodiversity from google map images (missing animals living indoors or in the wast system)
- Failing to recorded selected cases
  - Incomplete forestry surveys because of running out of funding
  - Members of the libertarian party who do not provide their household details for the Census

Introduction

- Most research datasets (outside lab conditions) are affected by missing data
- Failing to sample sections of the population, i.e. coverage error
  - Common in big data, e.g. assessing changes in beauty standards using pictures in social media
  - Study on urban biodiversity from google map images (missing animals living indoors or in the wast system)
- Failing to recorded selected cases
  - Incomplete forestry surveys because of running out of funding
  - Members of the libertarian party who do not provide their household details for the Census
- Cases that drop from a sample across time, i.e. attrition
  - Longitudinal study of twins
  - Study of whale migration patterns using attached GPS trackers

Introduction

- Implications of missing data
  - Smaller samples, i.e. loss of precision
  - Potentially biased samples, when missing cases are different from those observed

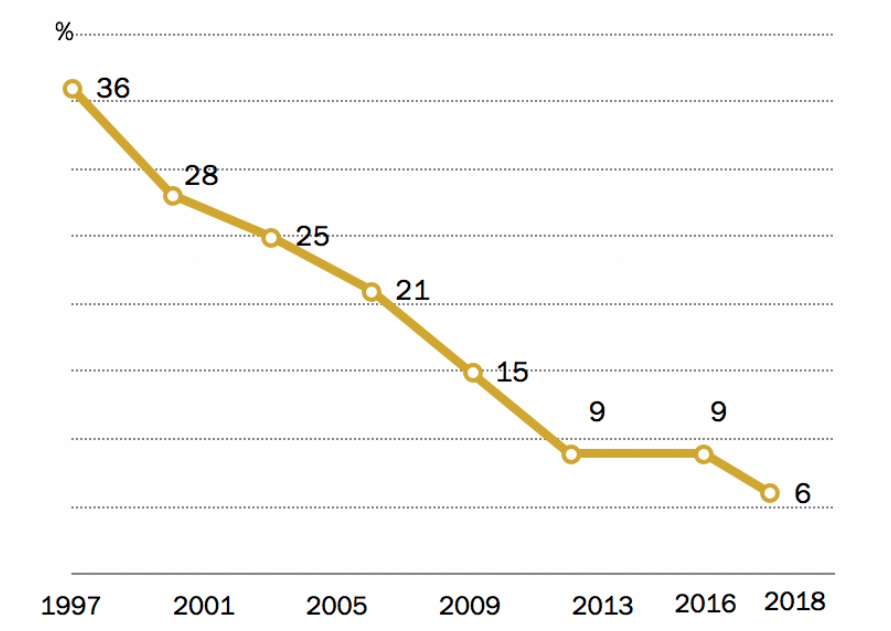
Introduction

- Implications of missing data
  - Smaller samples, i.e. loss of precision
  - Potentially biased samples, when missing cases are different from those observed
- A much bigger problem when we deal with human participants
  - People have agency and might prefer not to be studied
  - For reasons of: anonimity (e.g. census), social desirability (e.g. self-reported xenophobic attitudes), mistrust (ethnicity), lack of spare time (e.g. self-completed questionnaires)
  - Samples will often be not representative of the population

Introduction

After brief plateau, telephone survey response rates have fallen again

Response rate by year (%)



Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates are typical for surveys conducted in each year.
Source: Pew Research Center telephone surveys conducted 1997-2018.

PEW RESEARCH CENTER

Implications

- The impact of missing data on our findings will depend on:
  - The type of missing data mechanism
  - Its prevalence
  - What we do about it

Implications

- The impact of missing data on our findings will depend on:
  - The type of missing data mechanism
  - Its prevalence
  - What we do about it
- Missing data mechanisms
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Missing not at random (MNAR)
- What we do about it
  - Nothing (listwise deletion)
  - Use auxiliary data (e.g. probability weights, multiple imputation, full information maximum likelihood)
  - Sensitivity analysis

Implications

- The impact of missing data on our findings will depend on:
  - The type of missing data mechanism
  - Its prevalence
  - What we do about it
- Missing data mechanisms
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Missing not at random (MNAR)
- What we do about it
  - Nothing (listwise deletion)
  - Use auxiliary data (e.g. probability weights, imputation, full information maximum likelihood)
  - Sensitivity analysis

Missing Data Mechanisms

- Missing completely at random (MCAR)
  - The missing values are just a random subset of our data
  - E.g. some of the data was lost by accident, such as a random back up problem
  - Implications: loss of precision because of the smaller sample

Missing Data Mechanisms

- Missing completely at random (MCAR)
  - The missing values are just a random subset of our data
  - E.g. some of the data was lost by accident, such as a random back up problem
  - Implications: loss of precision because of the smaller sample
- Missing data at random (MAR)
  - The probability of data being missing is only random within groups defined by the observed data
  - E.g. In a court information system where the name of presiding judge is recorded, male judges forget to submit their sentence report more commonly than female judges
  - If left unadjusted will bias our estimates, if adjusted becomes 'ignorable' (the missing data can be considered MCAR)

Missing Data Mechanisms

- Missing completely at random (MCAR)
  - The missing values are just a random subset of our data
  - E.g. some of the data was lost by accident, such as a random back up problem
  - Implications: loss of precision because of the smaller sample
- Missing data at random (MAR)
  - The probability of data being missing is only random within groups defined by the observed data
  - E.g. In a court information system where the name of presiding judge is recorded, male judges forget to submit their sentence report more commonly than female judges
  - If left unadjusted will bias our estimates, if adjusted becomes 'ignorable' (the missing data can be considered MCAR)
- Missing data not at random (MNAR)
  - When neither MCAR nor MAR hold; the probability of the data being missing varies for reasons that are unknown to us
  - E.g. Offenders who believe to have been subjected to discrimination by the criminal justice system will be less likely to report their ethnicity
  - It is not easily adjustable; we can rely on sensitivity analysis to assess the extent of the bias under different scenarios

Listwise Deletion

- Drop all missing cases
  - The default approach
  - Assumes MCAR

Listwise Deletion

- Drop all missing cases
  - The default approach
  - Assumes MCAR

ID	Offender ethnicity	Offence type	Prev. convictions	Sentence outcome
1	White	Assault with harm	5	Custody
2	-	Common assault	10	Custody
3	-	Common assault	5	Non-custody
4	Black	Assault with harm	1	Custody
5	-	-	-	-
6	White	Assault with harm	3	Non-custody

Listwise Deletion

- Drop all missing cases
  - The default approach
  - Assumes MCAR

ID	Offender ethnicity	Offence type	Prev. convictions	Sentence outcome
1	White	Assault with harm	5	Custody
2	-	Common assault	10	Custody
3	-	Common assault	5	Non-custody
4	Black	Assault with harm	1	Custody
5	-	-	-	-
6	White	Assault with harm	3	Non-custody

ID	Offender ethnicity	Offence type	Prev. convictions	Sentence outcome
1	White	Assault with harm	5	Custody
4	Black	Assault with harm	1	Custody
6	White	Assault with harm	3	Non-custody

Adjustments Based on Auxiliary Data

- Use the information we observe to predict the missing cases
  - Assumes MAR
  - Can be undertaken in many different ways (examples below of single and multiple imputation)

Adjustments Based on Auxiliary Data

- Use the information we observe to predict the missing cases
  - Assumes MAR
  - Can be undertaken in many different ways (examples below of single and multiple imputation)

ID	Offender ethnicity	Offence type	Prev. convictions	Sentence outcome
1	White	Assault with harm	5	Custody
2	-	Common assault	10	Custody
3	-	Common assault	5	Non-custody
4	Black	Assault with harm	1	Custody
5	-	-	-	-
6	White	Assault with harm	3	Non-custody

Adjustments Based on Auxiliary Data

- Use the information we observe to predict the missing cases
  - Assumes MAR
  - Can be undertaken in many different ways (examples below of single and multiple imputation)

ID	Offender ethnicity	Offence type	Prev. convictions	Sentence outcome
1	White	Assault with harm	5	Custody
2	-	Common assault	10	Custody
3	-	Common assault	5	Non-custody
4	Black	Assault with harm	1	Custody
5	-	-	-	-
6	White	Assault with harm	3	Non-custody

ID	Offender ethnicity	Offence type	Prev. convictions	Sentence outcome
1	White	Assault with harm	5	Custody
2	Black	Common assault	10	Custody
3	White	Common assault	5	Non-custody
4	Black	Assault with harm	1	Custody
6	White	Assault with harm	3	Non-custody

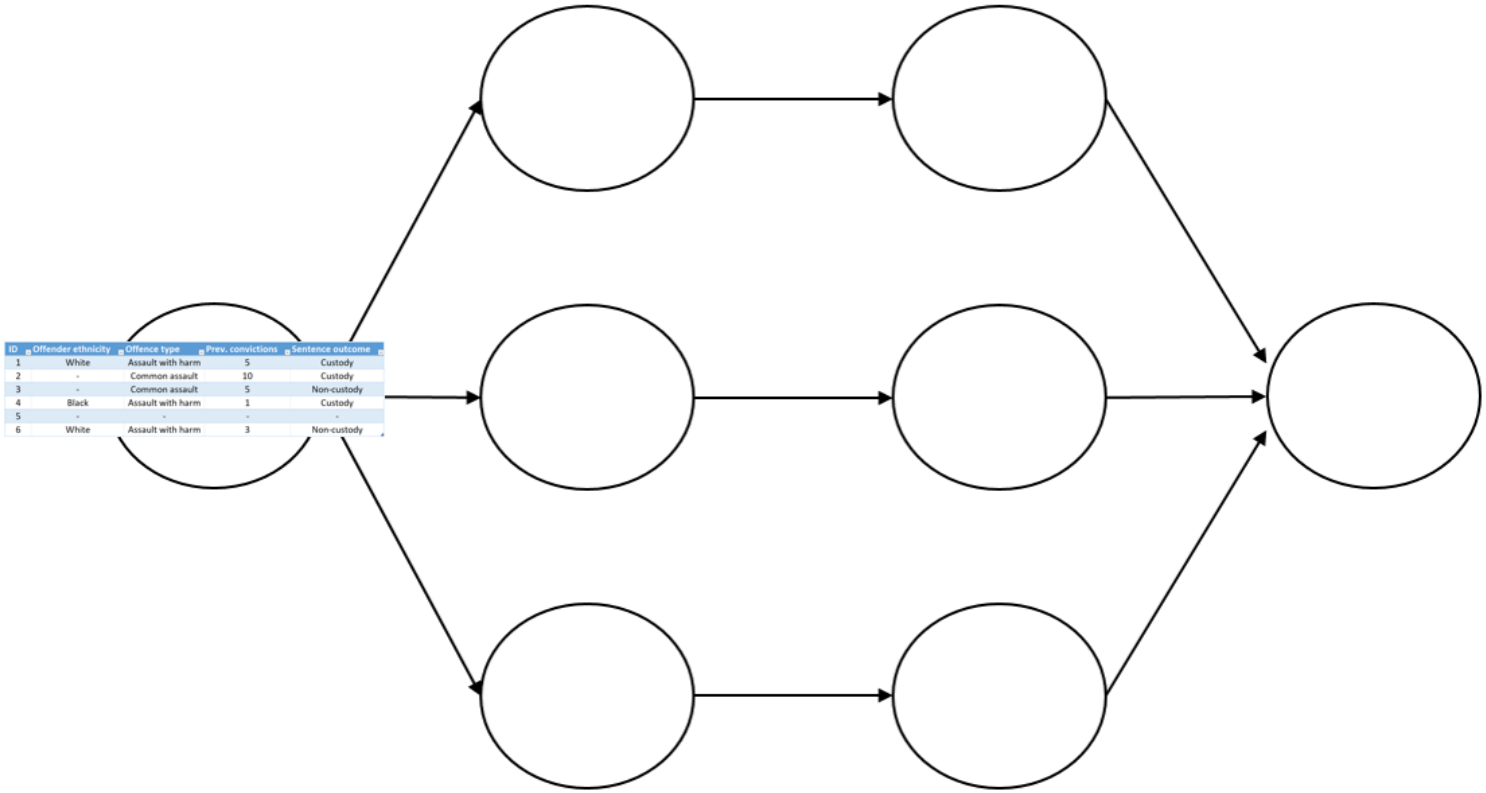
Imputation

- Hot-deck & regression imputation
  - Each missing case replaced with a value from a similar observation in the dataset
  - Uses other variables and cases for which there is complete information to make predictions about the missing values
  - Hot-deck imputation if the prediction is made using matching, regression imputation if using regression

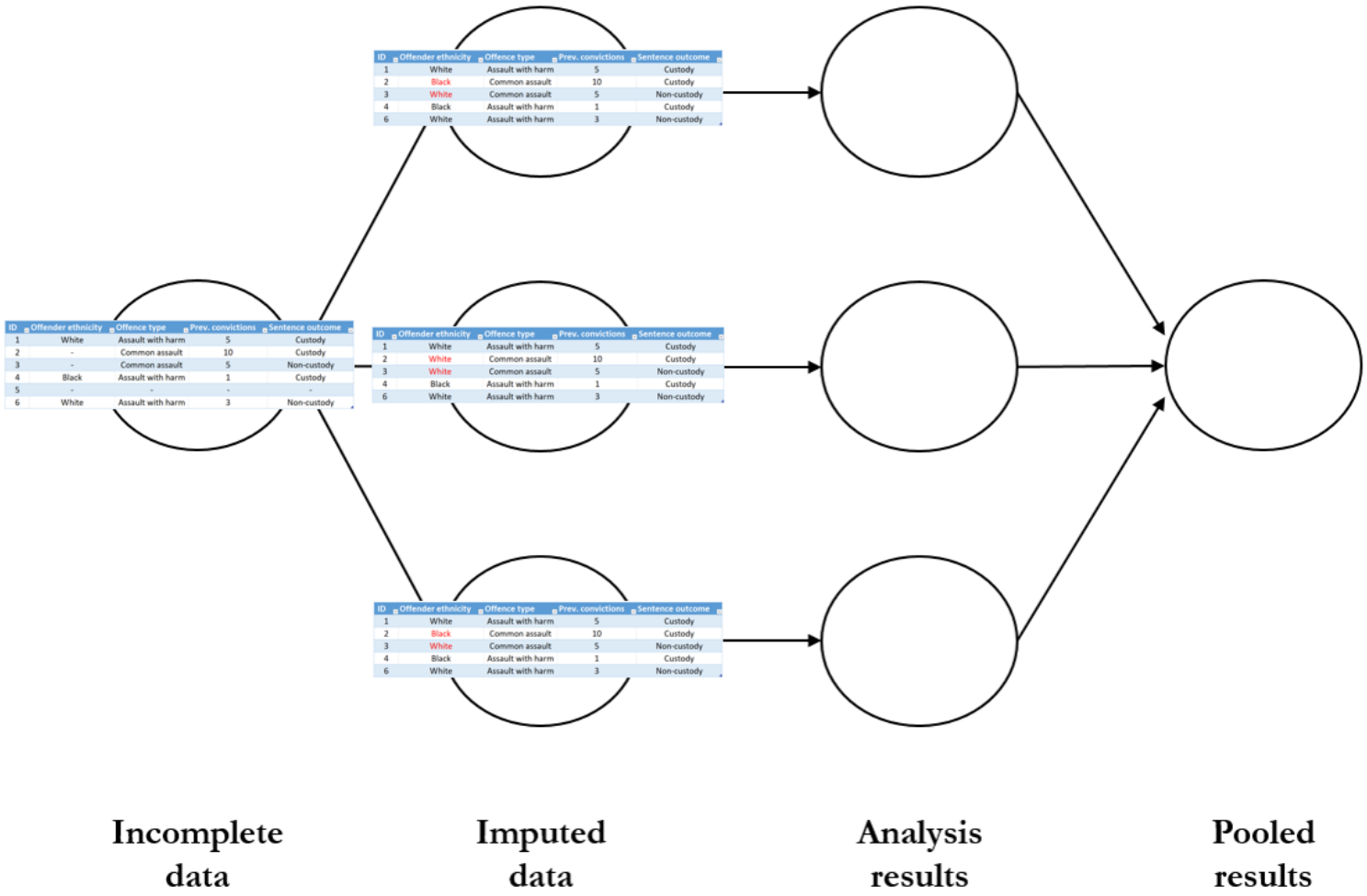
Imputation

- Hot-deck & regression imputation
  - Each missing case replaced with a value from a similar observation in the dataset
  - Uses other variables and cases for which there is complete information to make predictions about the missing values
  - Hot-deck imputation if the prediction is made using matching, regression imputation if using regression
- Multiple imputation
  - Each missing value is replaced with multiple plausible values to generate multiple complete data sets
  - Having multiple values eliminates the problem of treating imputed cases as real data, i.e. accounts for the uncertainty of the imputation process
  - The analysis is conducted in each of those datasets, results from each analysis are saved and pooled into an average of estimates

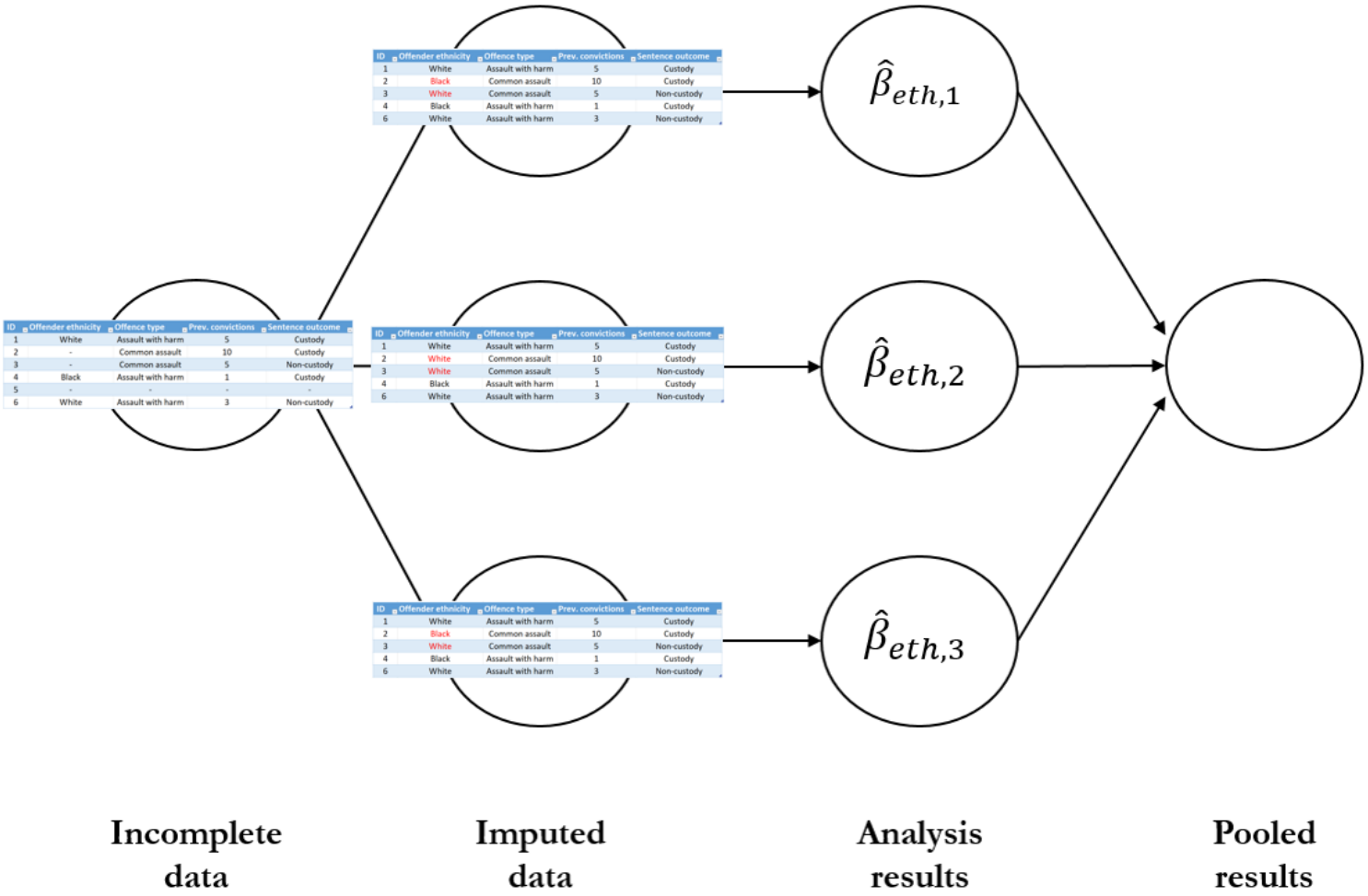
Multiple Imputation



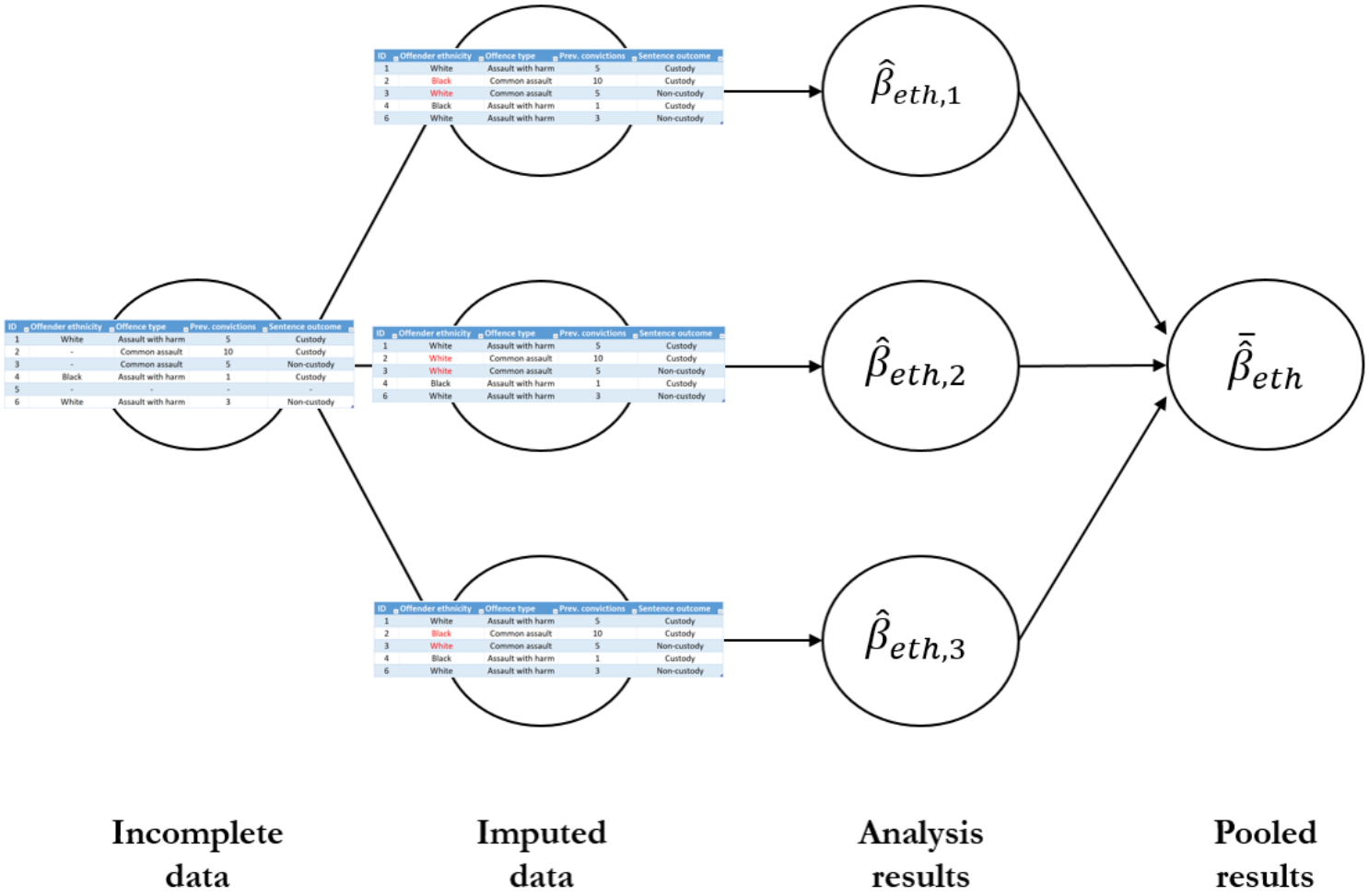
Multiple Imputation



Multiple Imputation



Multiple Imputation



Sensitivity Analysis for MNAR

- Often we do not have auxiliary data to predict the missingness, but cannot assume MCAR, so we have MNAR

Sensitivity Analysis for MNAR

- Often we do not have auxiliary data to predict the missingness, but cannot assume MCAR, so we have MNAR
- We can still rely on subjective or imperfect insights about the mechanisms behind the missingness
- For example, insights based on ...
  - Administrative data exploring the same population
  - Missing data analyses undertaken on different populations (i.e. literature review)
  - Interviews with experts, or individuals from the target population
  - Our own educated guess

Sensitivity Analysis for MNAR

- Examples of MNAR where we can still have an educated guess about the missingness

Sensitivity Analysis for MNAR

- Examples of MNAR where we can still have an educated guess about the missingness
  - Estimates of poverty from a survey based on sampling postal address, missing homeless people, of which we might have an educated guess of their prevalence and poverty level based on interviews with charity volunteers

Sensitivity Analysis for MNAR

- Examples of MNAR where we can still have an educated guess about the missingness
  - Estimates of poverty from a survey based on sampling postal address, missing homeless people, of which we might have an educated guess of their prevalence and poverty level based on interviews with charity volunteers
  - Missing data on monthly income from a survey on academic integrity in the UK, which we anticipate high earning academics might be more likely to miss, as we can see the distribution of academic wages in their pension fund data

Sensitivity Analysis for MNAR

- Examples of MNAR where we can still have an educated guess about the missingness
  - Estimates of poverty from a survey based on sampling postal address, missing homeless people, of which we might have an educated guess of their prevalence and poverty level based on interviews with charity volunteers
  - Missing data on monthly income from a survey on academic integrity in the UK, which we anticipate high earning academics might be more likely to miss, as we can see the distribution of academic wages in their pension fund data
  - Ethnic minority offenders who perceive they have been discriminated (e.g. set to remand when a similar offence committed by a White British offender would have been granted bail) are less likely to report their ethnicity, which we suspect based on the literature pointing at lower trust and compliance with the criminal justice system amongst ethnic minorities

Sensitivity Analysis for MNAR

- Examples of MNAR where we can still have an educated guess about the missingness
  - Estimates of poverty from a survey based on sampling postal address, missing homeless people, of which we might have an educated guess of their prevalence and poverty level based on interviews with charity volunteers
  - Missing data on monthly income from a survey on academic integrity in the UK, which we anticipate high earning academics might be more likely to miss, as we can see the distribution of academic wages in their pension fund data
  - Ethnic minority offenders who perceive they have been discriminated (e.g. set to remand when a similar offence committed by a White British offender would have been granted bail) are less likely to report their ethnicity, which we suspect based on the literature pointing at lower trust and compliance with the criminal justice system amongst ethnic minorities
- Question: Have you encountered similar MNAR problems?

Sensitivity Analysis

- Using sensitivity analysis we can explore the bias in our findings under different scenarios

Sensitivity Analysis

- Using sensitivity analysis we can explore the bias in our findings under different scenarios
- One way to do this is by simulating the unobserved cases according to our educated guesses
  - Not the only way, could also use informative priors in Bayesian statistics, or modified forms of maximum likelihood estimation
  - We like simulations as they are more intuitive and transparent than other approaches

Sensitivity Analysis

- Using sensitivity analysis we can explore the bias in our findings under different scenarios
- One way to do this is by simulating the unobserved cases according to our educated guesses
  - Not the only way, could also use informative priors in Bayesian statistics, or modified forms of maximum likelihood estimation
  - We like simulations as they are more intuitive and transparent than other approaches
- Then we replicate our analysis using the simulated data and note discrepancies in our estimates of interest

Sensitivity Analysis

- To undertake effective sensitivity analysis need to rely on two key skills
- Develop a good intuition about the missing mechanism
  - Subject specific knowledge, our key strength as social scientists
- Turn that 'soft' knowledge into statistical simulations
  - We will see how in the practical
  - Where we will explore different forms of missing ethnicity data in a study of disparities in sentencing

Workshop

- We will structure the practical as follows
  - Present the dataset (a pre-formatted version of Crown Court data from the Sentencing Council)
  - Estimate our naive model showing the effect of offenders' ethnicity on the probability of receiving a custodial sentence after controlling for key case characteristics
  - Simulate MCAR, MAR, and MNAR to assess their impact the effect of ethnicity that we have naively assumed when we disregard the missing data
  - Repeat this process for sentence length (if we have time)

Workshop

- The main take away points from this workshop

Workshop

- The main take away points from this workshop
  - The impact of missing data can vary largely
  - In some instances it can be very damaging, even when its prevalence is low

Workshop

- The main take away points from this workshop
  - The impact of missing data can vary largely
  - In some instances it can be very damaging, even when its prevalence is low
  - Doing nothing is not an option, neither is using multiple imputation or similar in autopilot

    Question: What would be the problem of relying just on multiple imputation in our example where offender's ethnicity is MNAR?

Workshop

- The main take away points from this workshop
  - The impact of missing data can vary largely
  - In some instances it can be very damaging, even when its prevalence is low
  - Doing nothing is not an option, neither is using multiple imputation or similar in autopilot

    Question: What would be the problem of relying just on multiple imputation in our example where offender's ethnicity is MNAR?
  - Undertaking sensitivity analysis to assess the robustness of our findings should be the default