Multiple Imputation for Missing Data

Benjamin Cooper, MPH
Public Health Data & Training Center
Institute for Public Health

Outline

- Missing data mechanisms
- What is Multiple Imputation?
- Software Options
  - SAS, Stata, IVEware, R, SPSS
- Syntax example
- Imputation issues and problems

Missing data mechanisms (MCAR)

- **Missing Completely At Random (MCAR)**
  - A variable is missing completely at random, if neither the variables in the dataset nor the unobserved value of the variable itself predict whether a value will be missing.
  - Likelihood of occurrence? RARE
  - Example: A subset of survey respondents are randomly selected to answer additional questions. The remaining respondents will be missing (completely at random) on those additional questions

Missing data mechanisms (MAR)

- **Missing At Random (MAR)**
  - A variable is said to be missing at random if other variables (but not the variable itself) in the dataset can be used to predict missingness on a given variable.
  - Likelihood of occurrence? Common
  - Example: Males may be less likely to answer certain survey questions than females. Gender predicts missingness

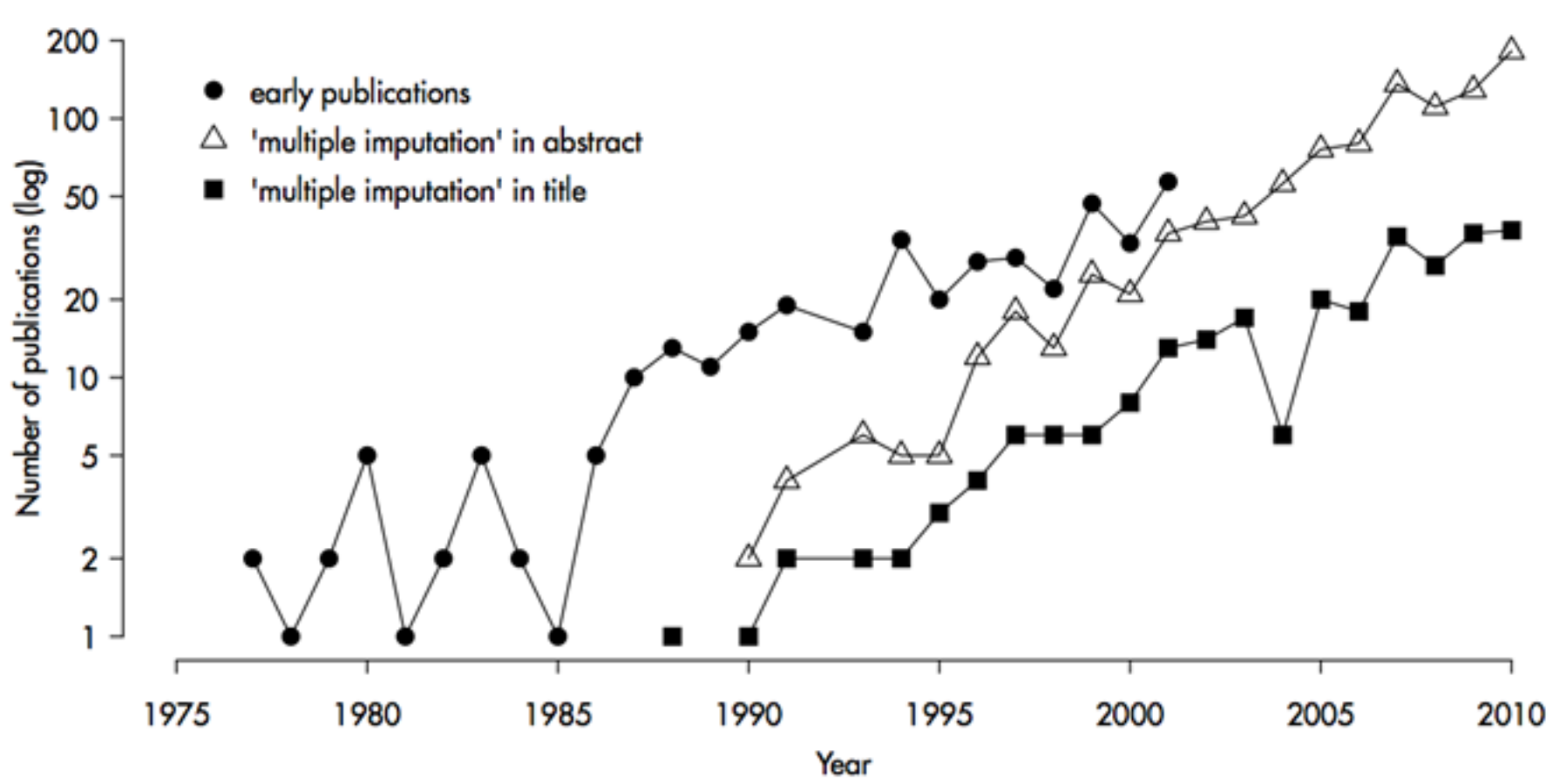
Missing data mechanisms (NMAR)

- **Not Missing at Random (NMAR)**
  - Data are said to be missing not at random if the value of the unobserved variable itself predicts missingness.
  - Likelihood of occurrence? Common
  - Example: Very high income individuals may be less likely to report their income than others with moderate income

Missing Data Mechanisms: Summary

- Different types of missing data require different treatment
- Data missing completely at random will not bias estimates, yet a reduced sample size from complete case analysis can increase standard errors
- In general, MI and other methods assume data are missing at random (MAR)

Multiple Imputation on the rise



Source: www.multiple-imputation.com

What is Multiple Imputation?

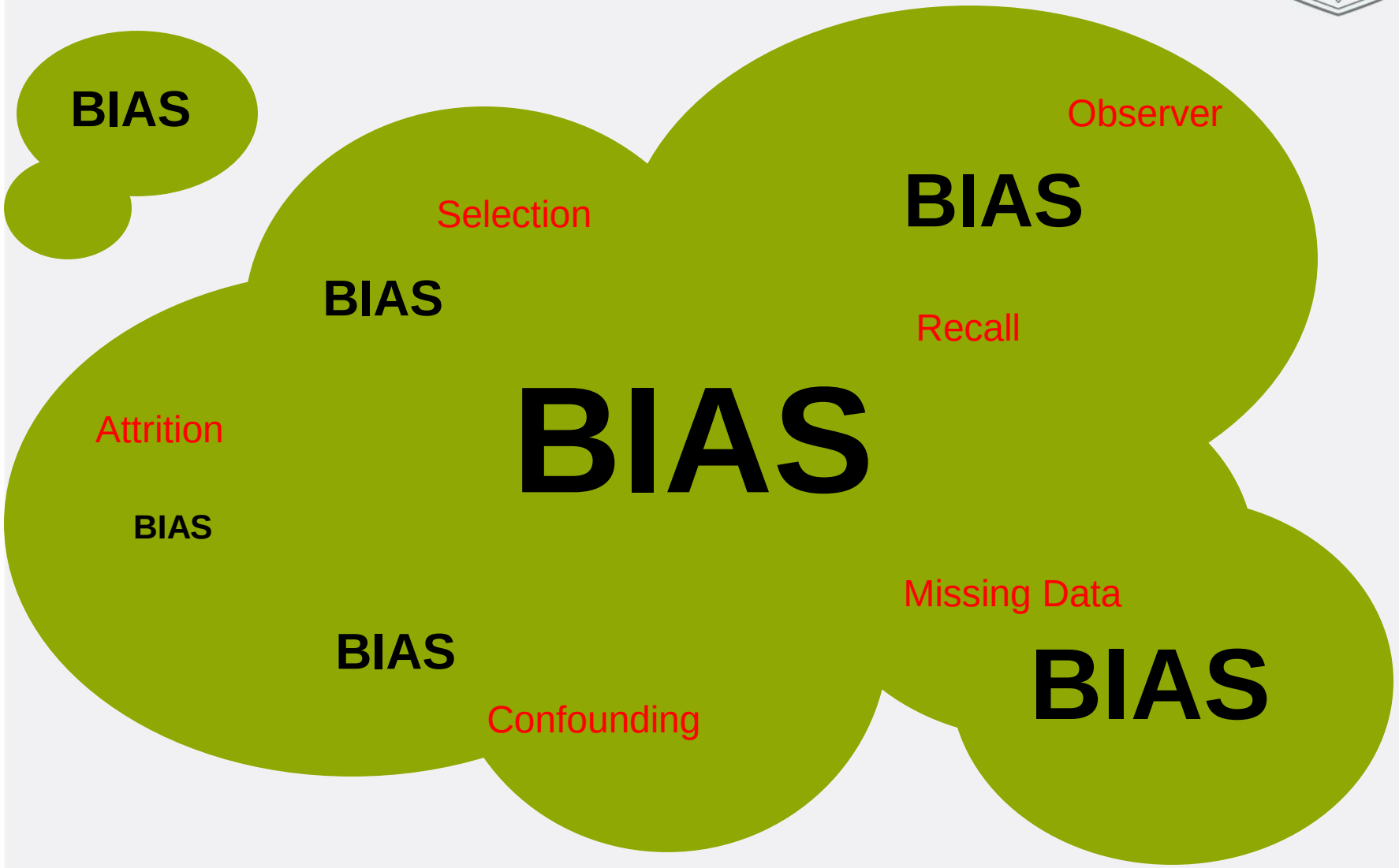
To impute values generally means to *replace* missing values with some other value.

There are a variety of methods to impute the value, such as mean imputation, nearest neighbor, hot deck, etc.

Multiple imputation creates multiple imputed values for any one given missing value.

The analysis takes place “separately” on each imputed dataset and then results are combined using Rubin’s combining rules.

Why Impute?


BIAS
Observer
BIAS
Selection
BIAS
Recall
BIAS
Attrition
BIAS
Missing Data
BIAS
BIAS
Confounding

Imputation: Multiple Complete Copies of the Dataset

Original Data

id	age	gender	height	weight
1	23	M	501	210
2	45	M	506	195
3	67	F	510	180
4	.	M	508	89
5	55	.	600	145
6	78	F	.	156
7	39	.	504	130
8	27	F	501	.
9	61	M	601	.
10	48	M	511	178

Implicate 1

id	age	gender	height	weight
1	23	M	501	210
2	45	M	506	195
3	67	F	510	180
4	46	M	508	89
5	55	M	600	145
6	78	F	511	156
7	39	F	504	130
8	27	F	501	155
9	61	M	601	165
10	48	M	511	178

Implicate 2

id	age	gender	height	weight
1	23	M	501	210
2	45	M	506	195
3	67	F	510	180
4	42	M	508	89
5	55	M	600	145
6	78	F	508	156
7	39	F	504	130
8	27	F	501	142
9	61	M	601	171
10	48	M	511	178

Three basic steps

1. ***Imputation***
   - Make *M*=2 to 50 copies (implicates) of original data set filling in with conditionally random values
2. ***Analyses***
   - Of each data set separately
3. ***Pooling***
   - *Point estimates.* Average across *M* analyses
   - *Standard errors & Confidence Intervals*. Combine variances.

Before you begin…

Certain variables should be created *before* imputation
- Example, multiple mutually exclusive binary variables for one construct (race)

If working with multiple discrete groups of observations, consider imputing separately and combine afterward.

Ensure all missing data is <null> or represented by a period. Alpha missing value codes may not get imputed.

Know your data! Check for skip patterns and other issues that could allow data to be imputed that shouldn't exist in the first place (e.g. number of cigarettes smoked/day for non-smokers)

MI software comparison

- STATA
  - based on each conditional density
  - chained equations
- SAS
  - joint distribution of all the variables
  - assumed multivariate normal distribution
- IVEware (SAS-callable or standalone)
  - same as Stata, more options for complex survey data.
- R
  - Multiple packages (mi, Amelia, MICE, etc.)
- SPSS (ver.17 or greater)
  - (offers MI but only through the add-pm Missing Values module

IVEware: IMPUTE command

IMPUTE – Multiple options without being labor intensive to setup

IMPUTE produces imputed values on a variable-by-variable basis for each individual in the data set conditional on all the values observed for that individual.

Imputations are created through a sequence of multiple regressions, varying the type of regression model by the type of variable being imputed.

Support for five data types with specific regression models for each: (1) Continuous (linear); (2) Binary (logistic); (3) categorical (polytomous with more than two categories); (4) Count (Poisson); and (5) mixed (a continuous variable with a non-zero probability mass at zero, generalized logit or mixed logistic/linear).

Covariates include all other variables observed or imputed for that individual.

The sequence of imputing missing values can be continued in a cyclical manner, each time overwriting previously drawn values, building interdependence among imputed values and exploiting the correlational structure among covariates.

Example: NHANES survey

Sub-sample from NHANES data (2009-2010)

- race, income, marital status, education level, age, gender, number in household
- n=10537
- Missing: Income 87(<1%), Marital Status 4319(59%), Education Level 4319(59%)

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
id	Respondent sequence number	10537	56892.00	3041.91	51624.00	62160.00
gender	Gender	10537	1.5041283	0.5000067	1.0000000	2.0000000
race	Race/Ethnicity - Recode	10537	2.7477460	1.1742422	1.0000000	5.0000000
adultedu	Education Level - Adults 20+	6218	3.2883564	1.3174066	1.0000000	9.0000000
marstat	Marital Status	6218	2.5561274	2.7399461	1.0000000	99.0000000
numhouse	Total number of people in the Household	10537	3.8294581	1.7620541	1.0000000	7.0000000
income	Annual Household Income	10450	11.5588517	16.8852457	1.0000000	99.0000000
RefPerAge	HH Ref Person Age	10537	45.5576540	16.0207168	18.0000000	80.0000000

PROC MI & IVEware: Comparing Syntax

IVEware Syntax:

```
%IMPUTE(name=ive, dir=C:\Users\John\SAS\MI, setup=new);
DATAIN example.MI_data;
DATAOUT final.MI_data_out ALL;

DEFAULT CATEGORICAL;
CONTINUOUS  adultedu ;
TRANSFER   id gender race adultedu marstat numhouse income ;

BOUNDS  age(>=6, <=17)  ;

ITERATIONS 10;
MINRSQD 0.01;
MULTIPLES 10;
SEED 54321;
PRINT  coef;

RUN;
```

PROC MI Syntax:

```
PROC MI=DATA.IN  NIMPUTE=5 OUT=DATA.OUT seed=54321;
var id gender race adultedu marstat numhouse income;
run;
```

A Few Issues

- Can I impute the dependent variable?
- Is there an upper limit to the amount of missing data to be imputed?
- How many implicates do I need?
- Can I impute in one software and analyze in another?

References

Stata (Windows & MacOS)

http://www.stata.com/capabilities/multiple-imputation/

http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt1.htm

SAS (Windows)

http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect038.htm

http://www.ats.ucla.edu/stat/sas/seminars/missing_data/mi_new_1.htm

IVEware (Windows / MacOS / Linux)

http://www.isr.umich.edu/src/smp/ive/

R (Windows / MacOS / Linux)

http://www.stat.ucla.edu/~yajima/Publication/mipaper.rev04.pdf

http://www.stat.columbia.edu/~gelman/arm/missing.pdf