Survey Inference with Incomplete Data

Trivellore Raghunathan

Chair and Professor of Biostatistics, School of Public Health

Research Professor, Institute for Social Research

University of Michigan

Presented at the National Conference on Health Statistics, August 16-18, 2010

Goals

- Importance of dealing with missing data in national surveys
- Weighting and Imputation as a general purpose solutions for missing data
- Why we need multiple imputation?
- Enhance the use of all available information for the creation of public-use datasets
- Design implications and future directions
- Other two talks provide several applications and software related issues

Missing Data

- A pervasive problem and is getting worse
  - Response rates are generally declining in all surveys (unit nonresponse)
  - Subjects who are willing to participate in surveys hesitate to provide all information (item nonresponse)
- Threat to quintessential notion of a representative sample from the population
  - Leading to bias of unknown direction and magnitude
  - Loss of efficiency

What is the reasons for missing data? (Missing Data Mechansim)

X	Y_{obs}
X	?

Missing Completely at random (MCAR)

$$Y_{obs} \overset{distribution}{=} ?$$

Missing at random (MAR)

$$Y_{obs} \mid X = x \overset{distribution}{=} ? \mid X = x$$

Not Missing At random (NMAR)

$$Y_{obs} \mid X = x \overset{distribution}{\neq} ? \mid X = x$$

Analysis

- Most complete-case (available case) analyses are valid under MCAR assumption
  - Default in most software packages
  - Unreasonable assumption
- MAR assumption is much weaker
  - Depends on how good are the X as predictors of Y
  - Non-testable assumption
- NMAR
  - Need explicit formulation of differences between respondents and non-respondents
  - Need External data
  - Non-testable assumption

Weighting (Unit Nonresponse)

- MAR assumption
- Group respondents and non-respondents based on X (Adjustment Cells)
- Attach weights to respondents in each group to compensate for non-respondents in the same group
  - Example : White females aged 25-35 living in Southwest Region

100 in sample

80 nonrespondents

20 nonrespondents

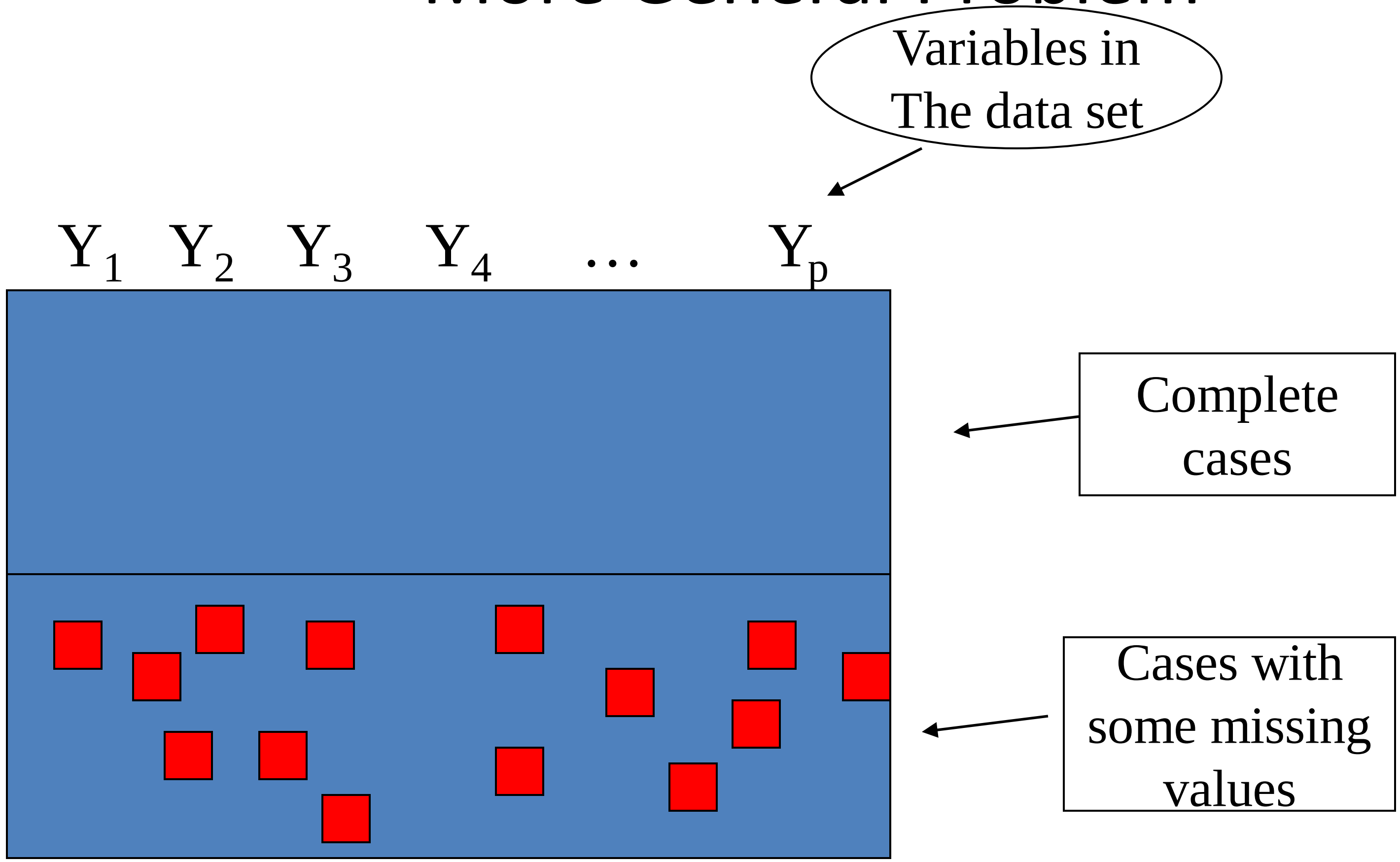
pr(response in cell) = 0.8

response weight = 1.25

Formation of Adjustment Cells

- Need X that are predictive of Y (or a collection of Y's in a multi-purpose survey)
- Using X's that are not predictive of Y will not reduce bias but will increase variance
- Current survey practice focuses too much on finding X's that differentiate respondents from non-respondents but predictive power of X for Y is more critical
- Need to think proactively in collecting X's that are related to multiple Y's through design modification

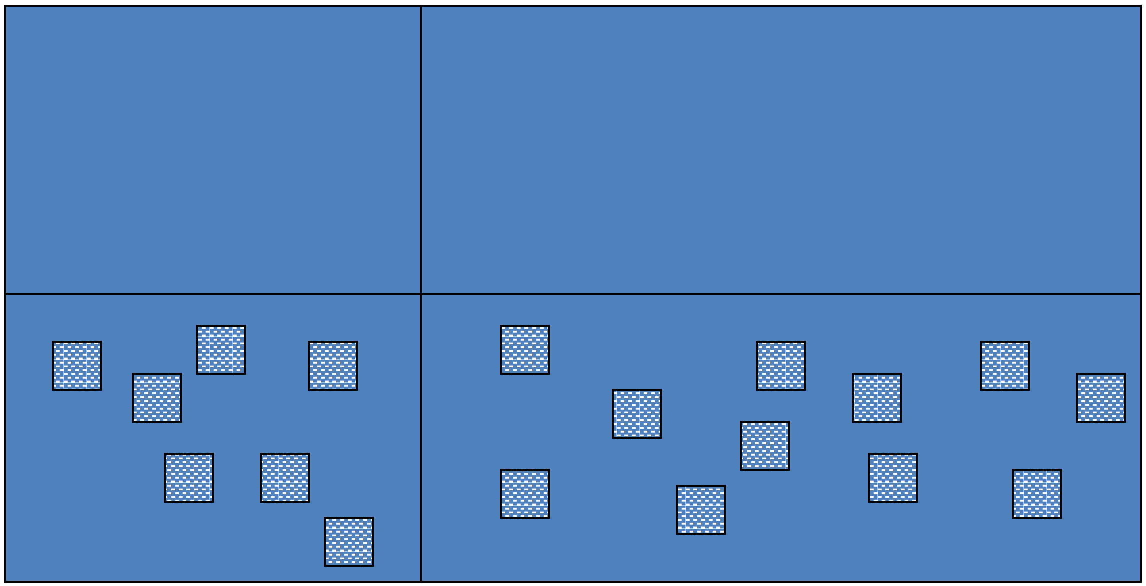
More General Problem



D_{obs} = Observed data:

D_{miss} = Missing data:

Imputation



Imputation refers to filling in a value for each missing datum based on other information (e.g., a model and observed data)

$Imputation:$

$Draws\ from\ predictive\ distribution\ \Pr(D_{miss} \mid D_{obs})$

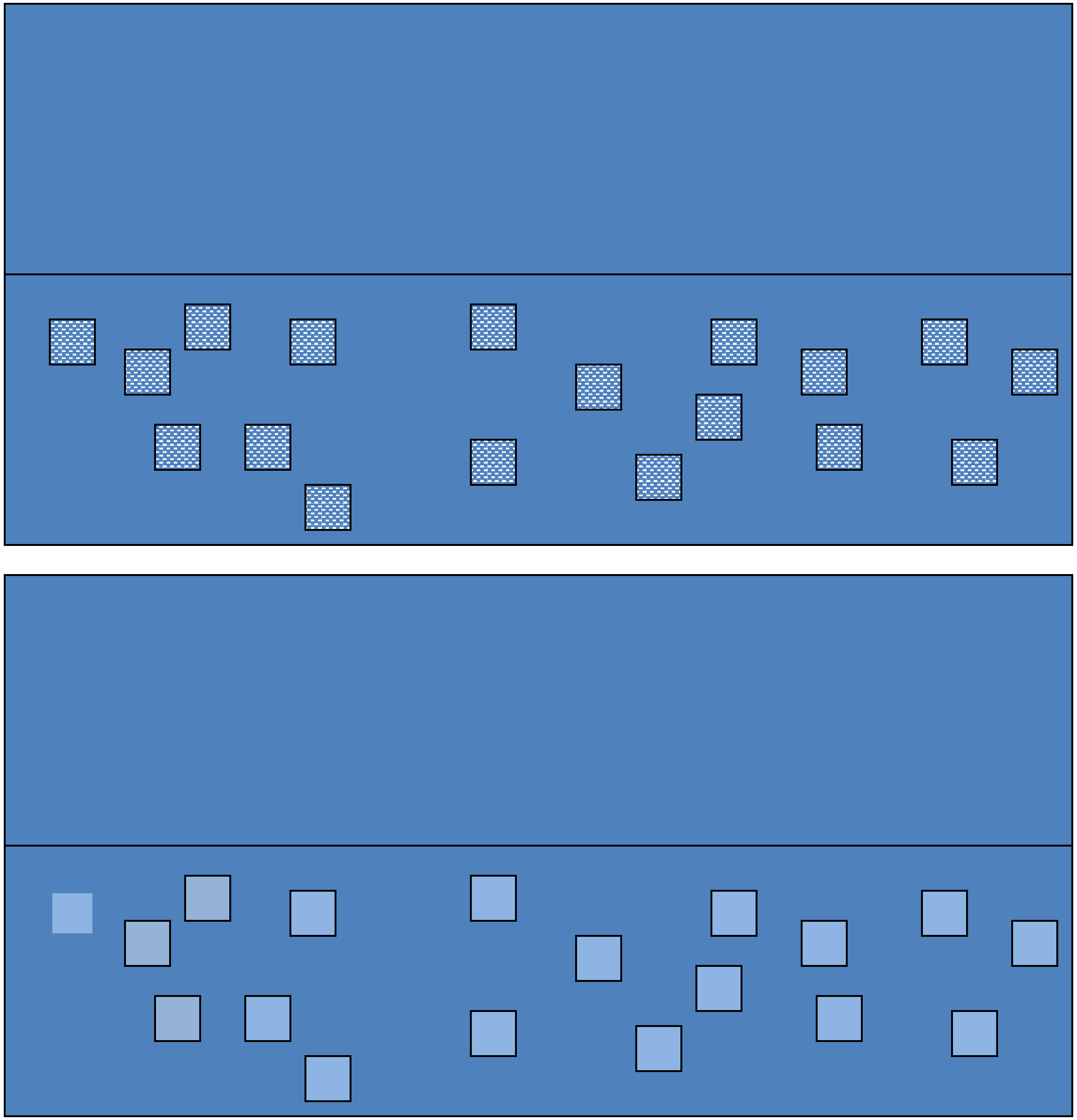
Imputation

- **Typically used for item nonresponse**
- **Benefits of imputation**
  - **Completes the data matrix**
  - **If imputation is performed by a producer of public-use data:**
    - **Missing data are handled comparably across secondary data analyses**
    - **Information available to the data producer but not the public can be used in creating imputations**

Imputation

- Important issues:
  - Imputations are not real values
  - Single imputation fed into standard software package treats the imputed values as real values
  - Underestimates the variance estimates due to ignoring uncertainties associated with imputes
- Goes against our “culture” where approximations of the sample designs (collapsing, combing PSUs, strata etc) avoid underestimation

Multiple Imputation



Repeat Imputation process several times (say M times)

Uncertainty due to imputation is captured by the "between Imputed Data" Variation

Analysis of Multiply Imputed Data

- Analyze each imputed data separately

$$Estimate: e_1, e_2, \ldots, e_M$$

$$Variance(= SE^2): v_1, v_2, \ldots, v_M$$

- Combine Estimates

$$\overline{e} = (e_1 + e_2 + \ldots + e_M)/M$$

- Combine variances

$$\left.\begin{aligned} \overline{v} &= (v_1 + v_2 + \ldots + v_M)/M \\ b &= \mathrm{var}(e_1, e_2, \ldots, e_M) \end{aligned}\right\} \quad T = \overline{v} + (1 + 1/M)b$$

Software for Creating Imputations

- SAS
  - PROC MI
  - User-developed IVEWARE (www.isr.umich.edu/src/smp/ive)
- Stata
  - ICE
- R
  - MICE
  - MI
- SOLAS
- AMELIA
- SPSS
- Stand-Alone
  - SRCWARE (www.isr.umich.edu/src/smp/ive)
  - NORM
  - PAN (www.stat.psu.edu/~jls)
  - CAT

Another good source:
www.multiple-imputation.com

Software for Analysis of Imputed Data

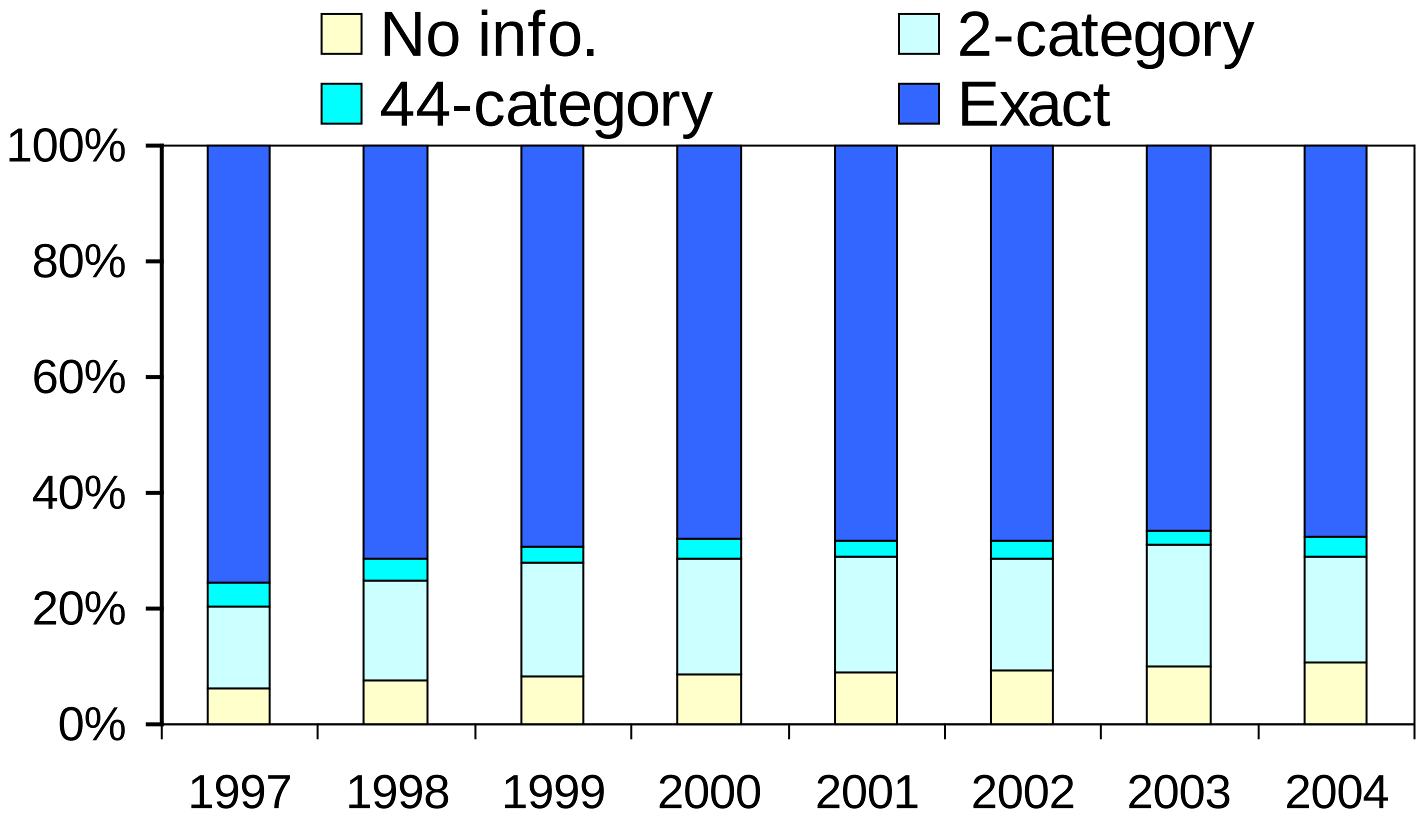
- SAS
  - MIANALYZE
  - IVEWARE
- SUDAAN
- STATA
  - MICOMBINE
  - MI
    - Newest version has excellent interface
- R (user defined macros)
- SRCWARE (Stand alone)

MULTIPLE IMPUTATION FOR MISSING INCOME DATA IN THE NATIONAL HEALTH INTERVIEW SURVEY

- Schenker, Raghunathan, Chiu, Makuc, Zhang, and Cohen (2006, JASA)
- National Health Interview Survey (NHIS)
  - Principal source of information on the health of the civilian non-institutionalized population
  - Data collected at both family and person levels
  - Contains items on health, demographic, and socioeconomic characteristics (e.g., income)
  - Allows the study of relationships between health and other characteristics

NHIS

- **Percent distribution of types of family income responses by year for the NHIS in 1997 – 2004**



Missingness appears to be related to several other characteristics, such as health, health insurance, age, race, country of birth, and region of residence

- **Missing income data multiply imputed for NHIS beginning with 1997**
  - ***M* = 5 sets of imputations of:**
  - **employment status for adults (< 4% missing)**
  - **personal earnings for adults who worked for pay**
  - **family income (and ratio of family income to Federal poverty threshold)**
- **Imputed income files since 1997, with documentation, available at NHIS Web site: www.cdc.gov/nchs/nhis/2008imputedincome.htm**
- **Used adaptation of IVEware**

- Complicating issues handled during imputation
  - Hierarchical structure of data
    - Families and persons
    - Sometimes, one variable (e.g., personal earnings) restricted based on another variable (e.g., whether worked for pay), but both variables missing
    - Imputation within bounds
      - e.g., families for which categories rather than exact dollar values reported for income
- Several variables used as predictors (including design variables)
- Different types (continuous, categorical, count)
  - Small amounts of missingness (mostly $< 2\%$)

Results

- **Estimated percentage of persons of ages 45-64 in fair or poor health, by ratio of family income to Federal poverty threshold: 2001 NHIS**

Ratio to Poverty Threshold	No Imp. (NI)		Single Imp. (SI)		Mult. Imp. (MI)		Ratio of SEs	
	Est.	SE	Est.	SE	Est.	SE	NI ÷ MI	SI ÷ MI
< 1.00	45.6	1.68	39.4	1.34	39.9	1.54	1.09	0.87
1.00 – 1.99	32.7	1.32	29.8	1.03	29.3	1.11	1.19	0.93
2.00 – 3.99	16.1	0.63	16.0	0.51	15.9	0.55	1.15	0.94
4.00+	5.9	0.34	6.1	0.27	6.2	0.30	1.11	0.90

Summary of Multiple Imputation

- Retains advantages of single imputation
  - Consistent analyses
  - Data collector's knowledge
  - Rectangular data sets
- Corrects disadvantages of single imputation
  - Reflects uncertainty in imputed values
  - Corrects inefficiency from imputing draws
    - estimates have high efficiency for modest *M*, e.g. 5
- For this approach to be successful, we need to collect good correlates of variables that are expected to have large amounts of missing values