

Nonresponse, Editing and Imputation in Surveys



Eric Rancourt
Statistics Canada
May 22-24, 2006
Bilbao, Spain

Instructor



- Studies
- Research work
- Position
- Interest

Content



1. Introduction
2. Nonresponse
3. Editing
4. Imputation methods

Content



5. Imputation principles and approaches
6. Variance
7. Software and Quality Assessment
8. Examples

Timetable



9:30 – 11:00 Lecture

11:00 – 11:30 Break

11:30 – 13:30 Lecture

13:00 – 15:00 Lunch

15:00 – 17:00 Lecture

1. Introduction



Outline



A. Overview

B. Prevention of nonresponse

C. Full response framework

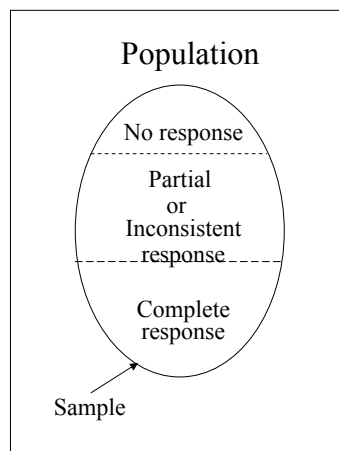
A. Overview



- Surveys
(also census or administrative data)
- Nonresponse
(and missing data)

What can we do?

Graphical representation





Examples of nonresponse

- Census (long form)
 - Everything but revenue answered
 - Respondent is 8 and married

- Business survey (monthly)
 - Responses are obtained quarterly
 - Data not available
 - Retired employees included in number of employees



Examples of nonresponse

- Agriculture Survey
 - A type of crop is missed
 - Small livestock forgotten

- Household Survey
 - Interview is long and some questions are too quickly dealt with (or skipped) near the end
 - Components of spending are not declared

Issues



- Nature of data
 - Establishment vs. Household
 - Skewed distributions vs. equally important units
- Parameter of interest (total or proportion)
- Nonrespondents different from respondents
- Impact of nonresponse (bias)
- Cannot “do nothing”
- Users

Definition of the problem



- How to perform estimation using a sample which contains missing data?
- Can we use only the complete response?
- How to use the information obtained from the partial respondents?
- How to create a usable data set for software designed for complete data sets?
- How to draw a correct inference to a population from a sample with missing data?

Approaches



- Using respondents only
(Do nothing approach)
- Re-weighting
- Imputation
- Unit substitution

Using respondents only



- Simple
- Complete file
- Does not invent data
- Many software packages available
- Not very efficient (some information is discarded)
- Risk of bias
- No design-based inference possible

Re-weighting



- Often simple
- Does not invent data (explicitly)
- Many software packages available
- Weights adjusted to eliminate or reduce bias
- Efficiency varying with the information used to compute weight adjustments
- Difficult to implement in cases of partial nonresponse

Imputation



- Complete data file
(allows for the use of complete data software)
- Utilizes all data
- Consistent for different analysts
- “Invents” data
- Data after imputation may be misleading
- Bias, variance, distorted relationships

Unit substitution



- Extra operations
- Apparent complete response
- Often leads to bias
- Inclusion probabilities / weights hard to compute

B. Prevention



- Determining objectives and taking potential nonresponse into account
- Elaborating and implementing survey and collection methods to maximize the amount of information obtained
- Developing and applying treatment and correction measures
- Measuring the impact of nonresponse to know the data quality and to better treat nonresponse on subsequent occasions

Steps of prevention



- Development
- Creation of the frame
- Elaboration of the design
- Questionnaire design
- Collection

Development



- Objectives of the survey
 - Realistic
 - Managing client's expectations
- Concepts
 - Clear definitions
- Resources
 - Sufficient for collection and follow-up (Time, Personnel)



Creation of the frame

- Coverage
 - Complete
 - No duplicates
 - Definition of units
- Contact information
 - Available
 - Correct
- Classification information
 - Language
 - Area of activity



Elaboration of the design

- Sub-sampling of nonrespondents
- Method of randomized response
- Sample common to several surveys
 - Burden
 - Collocated samples
- More efficient sample design
 - Auxiliary information
- Over sampling
 - Taking stratum response rates into account



Questionnaire design

- Development of questions
 - Simple questions
 - Appropriate length
 - Avoid abbreviations
 - Take collection mode into account
 - Good translation
 - Reduce instructions (guide)
 - Personalized questionnaire
 - Closed and interval questions



Questionnaire design

- Development of the questionnaire
 - Involve all parties
 - Evaluate previous surveys
 - Use cognitive research
(ex. Focus groups)
 - Tests and pilot survey

Note: Repeating questions is not wrong

Collection



- Refusal conversion
- Evaluation of previous results
- Measures for future use
- Good recruitment program
- Training
- Supervision

Collection



- Send an information letter
- Establish a good rapport
- Inform about the confidential aspect
- Draw interest for results
- Offer various collection modes
- Consider giving incentives
- Distribute cards for changes of address
- Explain why questions are asked

Treatment



- Follow-up
- Correction
- Imputation
- Estimation

Follow-up



- Allow for enough time in collection period
- Develop tighter edits on new variables
- Prioritize (e.g. score function)
- Set flags

Correction



- At collection time
- In the field
- Better than follow-up
- Not after collection

Imputation



- Auxiliary information
- Thorough modelling exercise
- Editing flags
- Enough time
- Evaluation before estimation

Estimation



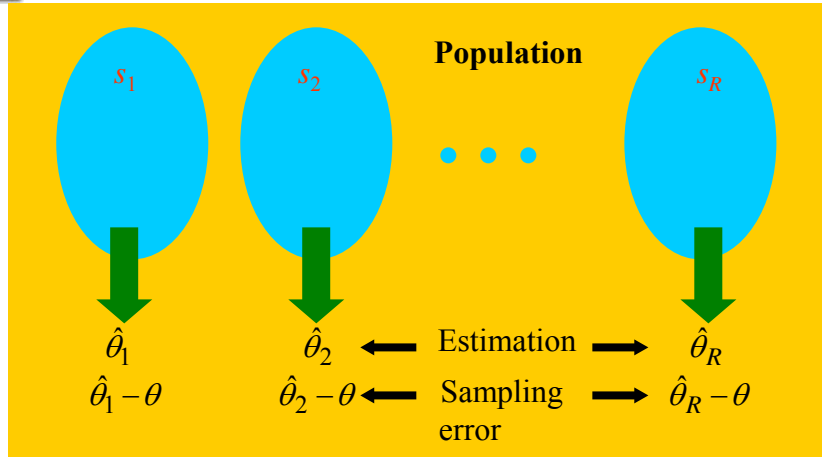
- More efficient
- Take adjustment / imputation into account
- Evaluate portion coming from treated nonresponse

C. Full response framework



- Let U be a finite population of possibly unknown size N
 $\{1, 2, \dots, i, \dots, N\}$
- Let y be a variable of interest
- The goal is to estimate parameters of interest of the finite population. A parameter of interest is a function of y_i such as:
 - The total
 - A domain mean.
 - A ratio of two population means.
 - Or others.

Full response framework



Full response framework



	Auxiliary Variables			Variables of Interest			Selection Indicator	Response Indicators			
	1	...	q	1	...	p		1	...	p	
Units in the population	1	z_{11}	...	z_{1q}	y_{11}	...	y_{1p}	I_1	a_{11}	...	a_{1p}

	i	z_{i1}	...	z_{iq}	y_{i1}	...	y_{ip}	I_i	a_{i1}	...	a_{ip}

	N	z_{N1}	...	z_{Nq}	y_{N1}	...	y_{Np}	I_N	a_{N1}	...	a_{Np}

Point Estimation



- Two types of estimators:
 - The **Horvitz-Thompson** estimator (HT)
 - The **generalized regression** estimator (GREG).

The Horvitz-Thompson Estimator



Parameter of interest is the population mean $\bar{Y} = \frac{1}{N} \sum_{i \in P} y_i$

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i \in S} w_i y_i$$

Example: For simple random sample, it coincides with the sample mean,

$$\bar{y}_{HT} \equiv \bar{y} = \frac{1}{n} \sum_{i \in S} y_i$$

The Generalized Regression Estimator



- Relationship of the form

$$m: y_i = \mathbf{z}'_i \boldsymbol{\beta} + \varepsilon_i$$

$$E_m(\varepsilon_i) = 0, V_m(\varepsilon_i) = \sigma_i^2, E_m(\varepsilon_i \varepsilon_j) = 0 \text{ if } i \neq j,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_q)$ is a vector of unknown parameters and σ_i^2 is an unknown parameter.

- Then,

$$Y = \sum_{i \in P} y_i = \sum_{i \in P} (\mathbf{z}'_i \boldsymbol{\beta} + \varepsilon_i) = \sum_{i \in P} \mathbf{z}'_i \boldsymbol{\beta} + \sum_{i \in P} \varepsilon_i, \text{ where } \varepsilon_i = y_i - \mathbf{z}'_i \boldsymbol{\beta}.$$

The Generalized Regression Estimator



- The generalized regression estimator (GREG):

$$\hat{Y}_{GREG} = \sum_{i \in P} \mathbf{z}'_i \hat{\boldsymbol{\beta}} + \sum_{i \in S} w_i e_i,$$

where $e_i = y_i - \mathbf{z}'_i \hat{\boldsymbol{\beta}}$.

- The GREG estimator may be written in many forms, such as:

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + (\mathbf{Z} - \hat{\mathbf{Z}}_{HT})' \hat{\boldsymbol{\beta}},$$

where $\hat{\mathbf{Z}}_{HT} = \sum_{i \in S} w_i \mathbf{z}_i$.

The Generalized Regression Estimator



- Let \mathbf{B} be the estimator of β

$$\mathbf{B} = \left(\sum_{i \in U} \mathbf{z}_i \mathbf{z}_i' / \sigma_i^2 \right)^{-1} \sum_{i \in U} z_i y_i / \sigma_i^2 .$$

- An estimator of \mathbf{B}

$$\hat{\mathbf{B}} = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / \sigma_i^2 \right)^{-1} \sum_{i \in s} w_i z_i y_i / \sigma_i^2 .$$

Estimation of Variance



- Why estimate variance?
 - To measure the quality (accuracy) of estimations.
 - To provide correct information to users.
 - To help draw the right conclusions.

Estimation of Variance



The sampling variance of $\hat{\theta}$ is

$$V_p(\hat{\theta}) = \sum_{s \in S} [\hat{\theta} - E_p(\hat{\theta})]^2 p(s)$$

If $\hat{\theta}$ is unbiased for θ , it becomes:

$$V_p(\hat{\theta}) = \sum_{s \in S} [\hat{\theta} - \theta]^2 p(s)$$

Variance of the HT Estimator



Example: Simple Random Sampling

In this case, $\pi_i = \frac{n}{N}$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ and

$$V_p(\hat{Y}_{HT}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$$

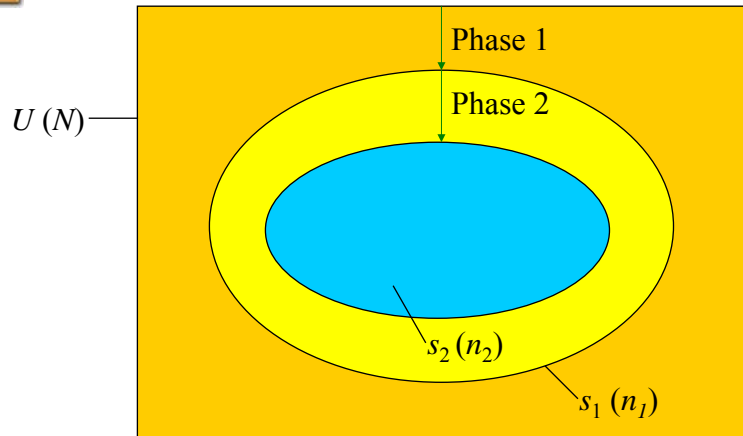
where $S_y^2 = \frac{1}{N-1} \sum_{i \in P} (y_i - \bar{Y})^2$ is the variance of the y

variable in the population and $\bar{Y} = \frac{1}{N} \sum_{i \in P} y_i$.

- S_y^2 is unknown and must be estimated.



Two-Phase Sampling



Two-Phase Sampling

- Parameter of interest: $Y = \sum_{i \in P} y_i$
- We have $\pi_{1i} = P(i \in s_1)$ and $\pi_{2i} = P(i \in s_2 | i \in s_1)$
- Let $w_{1i} = 1/\pi_{1i}$ and $w_{2i} = 1/\pi_{2i}$

$$\hat{Y}_{TP} = \sum_{i \in s} w_{1i} w_{2i} y_i$$

$$E_p(\hat{Y}_{TP}) = E_1 E_2(\hat{Y}_{TP} | s_1) = Y$$



Two-Phase Sampling

- The variance of \hat{Y}_{TP} is obtained as follows:

$$V(\hat{Y}_{TP}) = \underbrace{V_1 E_2(\hat{Y}_{TP} | s_1)}_{\text{Variance due to the first phase}} + \underbrace{E_1 V_2(\hat{Y}_{TP} | s_1)}_{\text{Variance due to the second phase}}$$

2. Nonresponse



Outline



- A. Definition
- B. Causes
- C. Types
- D. Classes

A. Definition



Cochran: Failure to measure some units of the selected sample.

Särndal:
Swensson Form of non-observation present in most surveys.
Wretman

Enlarged:
definition Failure to obtain a usable value in surveys.

Nonresponse

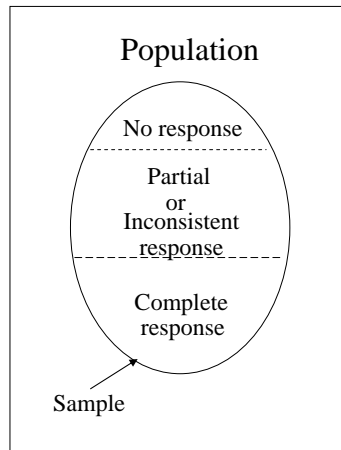


- Total nonresponse (unit)
 - No information obtained

- Partial nonresponse (item)
 - Some variables obtained

- Inconsistent or unusable response
 - Information obtained but not usable (e.g. out of scope units)
 - (→ Item nonresponse)

Graphical representation



Surveys



Censuses

Population Respondents

Surveys

Population → Sample → Respondents

Similar to:

Population → Phase I & II sample

→ Stage I & II sample

Others (combining data)

Files → Incomplete match

Nonresponse Mechanism



- Not controlled
- Not unique
- Key to solution
 - Causes
 - Types

Causes of nonresponse



- Total nonresponse
- Wrong contact information
 - Respondent is absent
 - Refusal
 - Move
 - Language problem
 - Closure
 - Lost questionnaire
 - Response burden too high
 - Survey perceived not to be important
 - Tight budget
 - Timeliness
 - Mandatory vs voluntary
- (Swain and Dolson 97)
(Panel on Incomplete Data 83)

Causes of nonresponse



- Question not understood
- Refusal
- Don't know
- Question forgotten by interviewer
- Data not available

Causes of nonresponse



Inconsistent or unusable response

(→ Item nonresponse)

- Impossible response
- Question wrongly understood
- Question wrongly asked
- Missing component in answer
- Response cannot be read
- Edits not satisfied
- Lost data

Other Causes



In longitudinal surveys

Censored data

- a) The measured duration started before the beginning of the study
- b) The measured duration will end after the end of the study

Truncated data

- a) The event happened before the study
- b) The event will happen after the study

Other causes



Planned nonresponse

Two-phase sampling

Apparent nonresponse

Response « Don't know » to a question on vote intention

→ The « true » value could effectively be « don't know »!

C. Types of nonresponse



1. Random (does not depend on a variable)

- Uniform mechanism

$$P(\text{answer} \mid X, Y) = P(\text{answer})$$

Also called :

MCAR (Missing completely at random)

Types of nonresponse



2. Depends on a variable
- Non-uniform mechanism

- 2.1 Depends on an auxiliary variable

$$P(\text{answer} \mid X, Y) = P(\text{answer} \mid X)$$

Also called :

MAR (Missing at random)

- 2.2 Depends on the variable of interest

$$P(\text{answer} \mid X, Y) = P(\text{answer} \mid X, Y)$$

Also called :

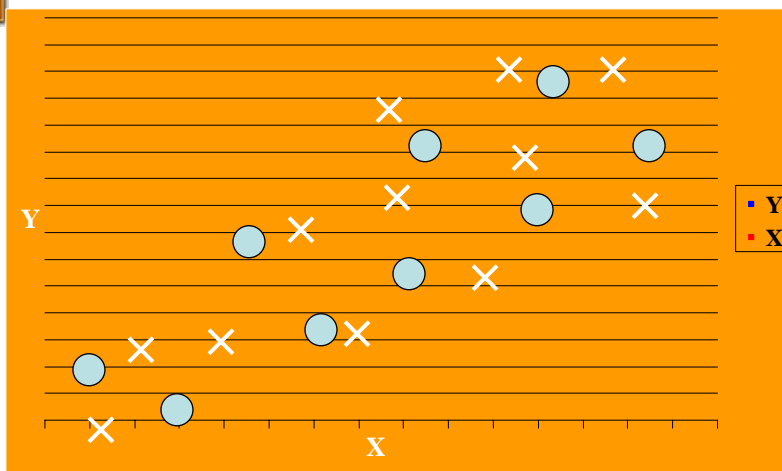
NMAR (Not missing at random)

Types of nonresponse

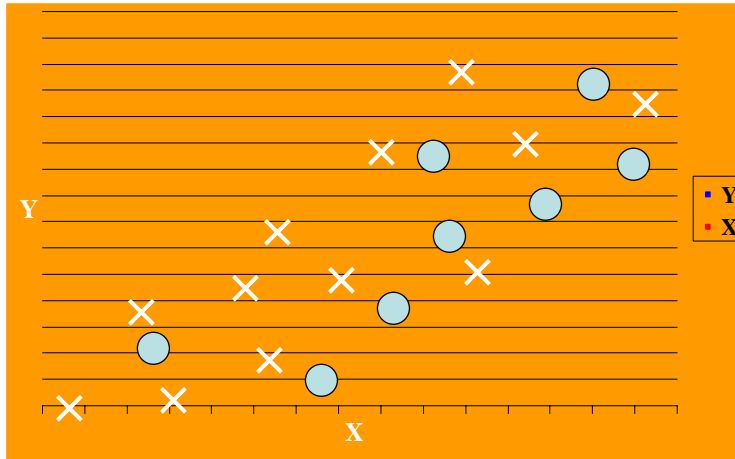


- Random (does not depend on a variable)
(Uniform mechanism)
- Depends on an auxiliary variable
- Depends on the variable of interest

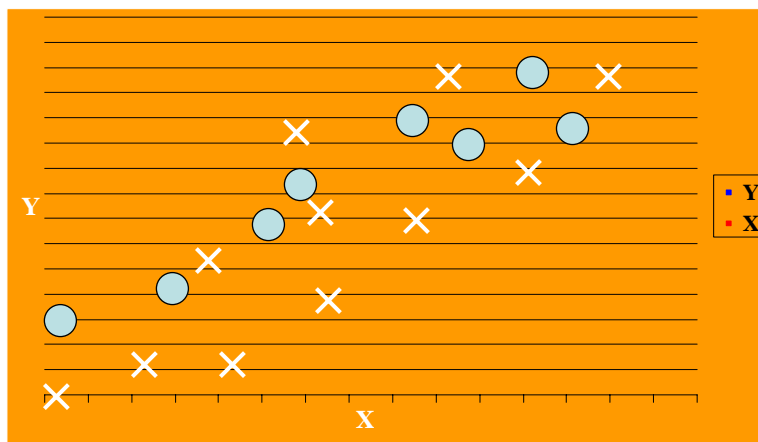
Uniform



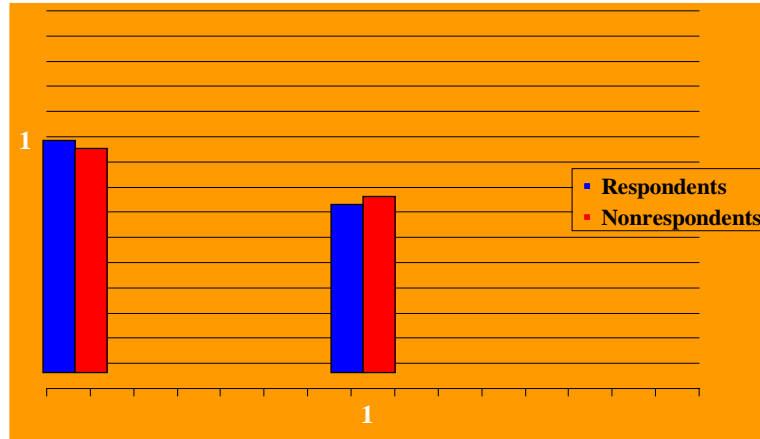
X- dependant



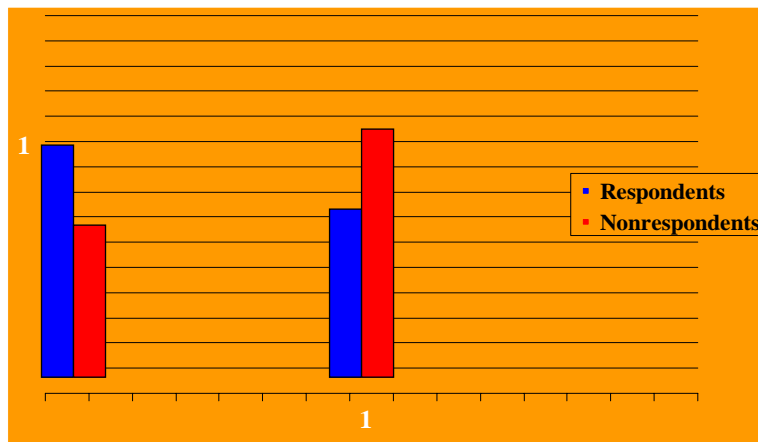
Y - dependant



Categorical variable (Uniform?)



Categorical variable (Uniform?)



Identifying nonresponse types



- Causes
- Experts
- Auxiliary information
- Comparing respondents and nonrespondents

Identifying nonresponse types



Modelling!

Effects Nonresponse



- Usually, the objective is to estimate totals and means. The effects nonresponse on the estimators include:
 - 1) Bias of the point estimators
 - 2) Increase of the variance of the point estimators
 - 3) Bias of the standard (naïve) variance estimators

Minimizing impact



1. Modelling
2. Classes

D. Classes



- Stratum
- Model group
- Domain
- Edit class
- Weighting class
- Imputation class
- Analysis class

Edit classes



Definition :

Partitions of the sample within which groups of edit rules are applied.

(Data Groups)

Examples :

- ➡ Regional offices
- ➡ Provinces
- ➡ Batch lots

Weighting classes



Definition :

Partitions of the sample within which weighting adjustments are computed.

Examples :

- Strata
- Model groups
- Homogeneous response groups

Imputation classes



Definition :

Partitions of the sample within which imputations are made.

Examples :

- Edit classes
- Domains
- Groups of strata

Analysis classes



Definition :

Partitions of the sample within which the analysis is performed.

- Equivalent to domains
- Often implicit to the model

Characteristics



1. Often based on socio-demographic or economic criteria;
(Age * sex), (Industry * region)
2. Should be close to publication domains (potential bias);

Ex : Imputation class : Age
Domain of interest : Age * sex

If there is a difference between men and women for the variable studied, it will be affected by the imputation.

Characteristics



3. Construct homogeneous classes;
 - a) according to observed averages;
 - b) according to relations between variables; (fit of the model);
 - c) according to the estimated response probabilities.

Characteristics



4. Do not use all combinations for categorical variables (over-specification of the model where all possible interactions are included);
5. Should be large enough

→ « (Example) » ← : 20 units or more

Construction



Definition: Group of units formed at the editing and imputation stage

Methods:

- Subject matter specialist
- Domains
- Classification techniques
- Response probability

Why using classes?



- **To reduce or eliminate the bias due to nonresponse**

• U : Population of size N ;

- Parameter: $\bar{Y} = \frac{1}{N} \sum_U y_i$

where y is the variable of interest

- Random Sample s of size n
- Assume that each unit responds independently with probability p_i

Why using classes?



- An imputed estimator of \bar{Y} is given by

$$\bar{y}_{I,1} = \frac{1}{n} \left[\sum_{s_r} y_i + \sum_{s_m} \hat{y}_i \right]$$

where \hat{y}_i denotes the imputed value for missing y_i .

- Under mean imputation, $\hat{y}_i = \frac{1}{r} \sum_{s_r} y_i$

Why using classes?



- $\bar{y}_{I,1}$ is biased:

$$Bias(\bar{y}_I) = \frac{1}{NP} \sum_U (p_i - \bar{P})(y_i - \bar{Y})$$

- Bias is 0 if the covariance between the variables p and y is 0

Recipe for class creation



Recipe for group creation:

similar \hat{p}_i
and/or
similar \hat{y}_i } Modelling

If well performed: Mean or hot-deck imputation
is sufficient

3. Editing



Outline



- A. Definition and objectives
- B. Finding errors
- C. Impact of editing
- D. Selective editing
- E. Macro editing



A. Definition and Objectives

- All procedures aiming at detecting wrong or suspicious values
- Applied at various levels of aggregation
- Can be manual or automated
- Not a correction tool for data but rather a quality control tool



Goals of editing

- Provide the basis for future improvement of the survey vehicle
- Provide information about the quality of the data
- Tidy up the data

Granquist (1984)

Granquist and Kovar (1997)

Why?



- Improve quality (over time)
- Availability of computers allows for more editing (not always good)
- Should enhance scope not volume of checks (must analyze failures)
- Savings can be redirected into more respondent follow-up
- Moving “on line” allows for more checks, more changes while with the respondent

Continuous improvement



- Impact throughout - not an isolated process
- => Seek optimal combination of data collection, editing, imputation
- => Appropriate mix of manual and automated procedures

Continuous Improvement



- Tracking/monitoring essential
- Audit trails, diagnostics, performance measures must be kept and studied
 - => Identify best practices
- Rethink objectives, scope...
- Reengineering, not conversion (incremental improvements)
- Prevention, not correction

B. Finding errors



- Types of edit rules
- Edit sets
- Error localisation principles

Types of edit rules



- Validity
- Consistency
- Distribution

Validity rules



- About the format of the expected answer

Examples:

- Capture error

Q: What is your age?

A: 331 (instead of 31)

- Invalid code

Q: What is your favorite activity?

a) work b) reading

c) sport d) other

A: e

Consistency rules



- Based on socio-economic laws or mathematical expression about relations that are known or assumed to be true

- Absolute rule

- Q1: # children in household: N1

- Q2: # adults in household: N2

- Q3: # people in household: N3

$$N1 + N2 = N3$$

If N1 = 2, N2 = 2, N3 = 4 → Satisfied

If N1 = 2, N2 = 2, N3 = 5 → Not satisfied

Consistency rules



- Non absolute rule

Example 1:

Q1: Sales: S

Q2: Expenses: E

Q3: Profit: P

$$S - E = P$$

Example 2:

Q1: Marital status: M

Q2: Age: A

Rule:

If M = married and A < 15 → Not satisfied

Note: It is possible to find real cases not satisfying the rules.



Distribution rules (statistical rules)

- These rules pay attention to the values of the variables and their inter-relationship
 - Absolute
Fixed limits are established
($y \geq 0$)
 - Boundaries (univariate)
Using the distribution, the 5th and 95th percentiles are obtained (or another measure)
Example
hours worked in a week by full time employees
 $25 < \text{\# hours} < 50$



Distribution rules

- Distance to the centre

$$d_k = \frac{|y_k - m|}{s}$$

m : measure of central tendency

Examples: Average, Median

s : measure of dispersion

Examples: Standard-error

If $d_k > c$ edit not satisfied



Distribution rules

- Sigma gap method
 - Calculation of σ : standard-error
 - Data are sorted in increasing order
 - The first y_k greater than the median for which $y_k - y_{k-1} \geq \alpha \sigma$ is looked for
 - All units greater than y_k do not satisfy the edit rule
- Any outlier detection method



Edit types

- Fatal edits (point to certain errors)
 - invalid entries
 - missing values
 - inconsistent responses
- Query edits (high probability of error)
 - data outside of subjective bounds
 - relatively high (low) values
 - “suspicious” entries

Fatal edits



- Fatal errors must be removed (user confidence)
 - editing well suited
 - not the costly task
 - judgment needed in fixing inconsistencies

Query edits



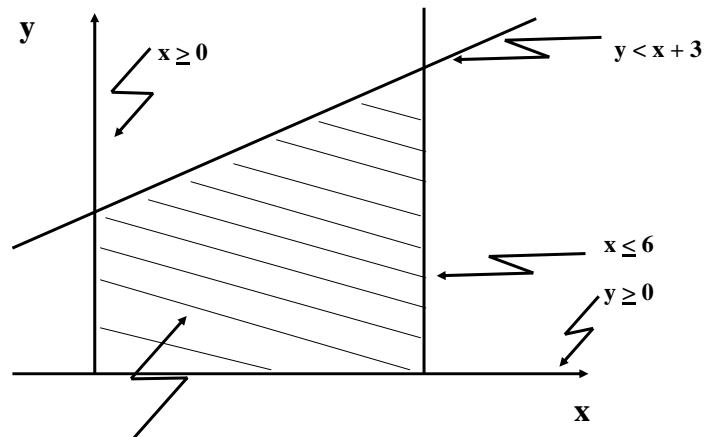
- Responsible for unacceptable costs (too many follow-ups)
- Must be defensible in light of benefits
- Balance must be struck (follow-up or not)

Edit sets



- Ensure consistency
- Remove redundancy
- Edit the edit set

Example of an edit set



ACCEPTANCE REGION

Consistency



- There may be some edits which are contradicting each other, thereby defining an empty region

Example

$y < x + 3$ and $x < 6$
are specified, then
 $y > 10$ leads to inconsistency

Redundancy



- Edits that are implied by the other edits and which need not be specified or verified

Example as above

$y < 9$ is redundant



Editing the edit sets

- Syntax verification
- Edit set complements (change conflict to validity)
- Consistency checks
- Redundancy checks
- Implied edits
- Hidden equalities
- Extreme values / Acceptable ranges
- Decision Logic Tables



Application

- Rules are either
 - Satisfied (no treatment required) or not satisfied (then...)
- Unit to be verified (query)
 - Manual treatment
 - Follow-up of the respondent
 - Kept but not used in the options for treatment
- Unit becomes missing (fatal)
 - Treated according to one of the options for treatment
- Two “boundaries”

Error localization



- Complex problem
 - Mathematical
 - Practical
- Fail vs. pass “region”
- Based on principles (e.g. minimum change)

Error localization



- Minimum change principle
(Fellegi and Holt, 76)

Description:

Changing the fewest amount of variables in order for the unit to satisfy the edit rules

Ex: Rule $A + B = C$
 $A=2, B=2, C=5$

A or B or C (only 1 change) is changed, but not (A and B) or (A and C) or (B and C) or (A, B and C).

Error localization



- NIM approach
(Bankier et al, 1996)

Description

Subject to available donors, changing the fewest amount of variables to satisfy the edit rules

Ex:

Relationships	Status	Age
son	single	35
daughter	single	32
mother	---	38

Error localization



- Sequential change principle

Description:

Question by question verification where consistency is achieved with respect to the previous questions

R1: $A > 2$

R2: $B = 1$ or $B = 2$ or $B = 3$

R3: $C = A + B$

We have $A = 1$, $B = 3$ and $C = 4$

- A is set to 3
- C is set to $3 + 3 = 6$

C. Impact of Editing



- When and where
- Cost of Editing
- Impact on Quality
- Issues

When and where to edit?



- During collection
 - By the interviewer
 - By the supervisor
- During capture
 - Automated rules in the system
- Before re-weighting
- Before imputation
- After imputation (post-editing)
- During analysis (macro-editing)

Cost of editing



- 20 - 40% of total survey cost
- Execution costs
 - Salaries
 - Computers and software
- Respondent re-contact most costly (for both parties)
- Bad-will costs (over burden)
- Opportunity costs (higher pay-off elsewhere)

Impact of editing



- Data changes (studies from Australia, Canada, U.S., Sweden...)
 - few large changes
 - many insignificant changes
 - e.g. 5% of changes results in a 90% overall change
- “Raw” to final comparisons
 - low percentage change (2, 10, 18)
- Manual review leaves many suspicious values unchanged
 - (20 - 30% hit rates)

Impact of editing



- Changes considered as “corrections”
 - editors’ differences of opinion
 - editing can actually be counter productive
 - point in time exists when just as many errors are introduced as are removed

Impact of editing



- Ability of editors to “fit reported data to models imposed by the edits”
 - spurious changes to “please the computer”
- Query edits only useful in verifying potential problems
 - editing 5 - 10% of values likely enough (somewhat more records)

Impact of editing



- Increased knowledge of survey data
- More control on the survey process
- Monitoring (catching) - problems
- changes

Issues



- Ability of current edits to detect errors
 - small systematic errors undetectable (concepts problems)
- Ability of respondents to report
 - different aggregations
 - memory limitations
 - not worth the effort for respondent
 - difference in concept (with some wording)

Editing and quality



Over-editing

- Changing too many data points can lead to massaging the data set up to an artificially “clean” status

- It could happen that hypotheses tested to be true from the survey data are such solely because the data were changed to satisfy these very hypotheses

Opportunity



- Editing is a high cost activity (20-40%)
- Time consuming
 - lost opportunities
 - timeliness and relevance “cost”
- Overediting - too much “double checking”
 - low hit rates
 - relatively few changes
- Creative editing
 - changes not always corrections
 - after a point as many new errors are introduced as corrected
- Differential (non-linear) impact of errors (Influential observations)

Opportunity



For total and query edits

- Small impact errors need to be removed (preserve agency reputation)
- Processing convenience
- Must be made quickly and objectively
=> fatal edits => automated methods
- Follow up with respondents
=> query edits => selective editing

D. Selective editing



- Complete editing is performed only for a sub-set of the survey data (subset of records or variables)
- Data are split between two groups: critical and non-critical units
- Critical units are subjected to all edit rules
- Non-critical units are subjected to a restricted number of edit rules (or no rules at all)

Can selective editing be put into practice?



- Greenberg & Petkunas (US)
 - manual review of large changes only
 - few errors responsible for majority of changes
 - few rudimentary edits sufficient
 - close to “final data” quickly

- Boucher, et al.; Kozak (Canada)
 - two streams
 - significant time gains with no quality losses
 - 20% of resources saved

Ordering errors by impact



- Latouche and Berthelot (Canada)
 - score function (static or real-time)
 - only 20% of units followed up (estimates within 2%)
 - complete process rethought
 - respondent burden minimized
- van de Pol (The Netherlands)
 - reduced editing to 25% of original effort
 - estimates within original confidence limits
- McDavitt et al. (Australia)
 - edit only 40% of failed records
 - “Significance Editing” => terminology (output editing, macro editing, aggregate editing...)
- More examples in Granquist and Kovar (1997)

F. Macro-editing

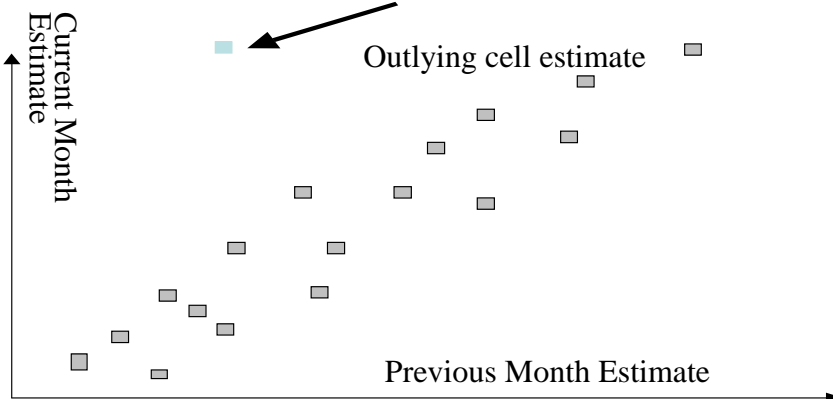


Definition

Editing of estimates or aggregated statistics (rather than micro-data)

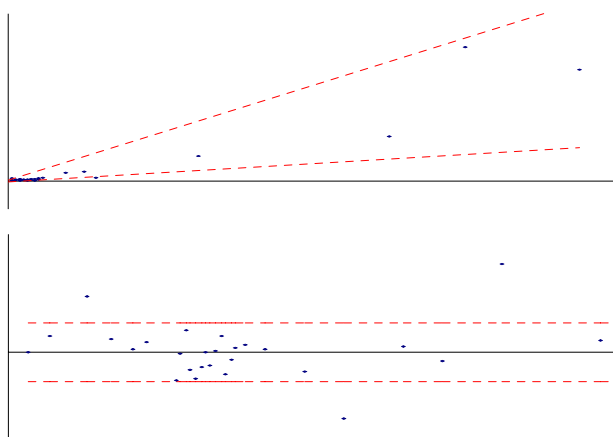
- Top-down method
- External micro-match
- Historical micro-match
- Hidiroglou - Berthelot
- Graphical methods
- Macro-editing often leads to micro-editing of selected data subsets

Macro-editing example Graphical editing



Macro-editing example

Hidiroglou - Berthelot



4. Imputation



Outline



- A. Definition
- B. Methods
- C. Issues

A. Definition



Description

For each missing value (or identified as such), a **replacement** value is found.

The process is carried out to the best of the knowledge (according to the availability of the auxiliary information).

Imputation may be carried out manually or using a computer.

Imputation



Characteristics

- Complete sample
(allows for the use of complete data software)
- Utilise all data
- Consistent for different analysts
- « Invents » data
- Data after imputation may be misleading

B. Imputation Methods



- Classifications
- Methods

Classification of Imputation Methods (unsettled terminology)



- Deterministic or Stochastic
- Deductive or Current data or Both
- Hot deck or Cold deck
- Multiple or Single
- Proper or not
- etc. ...

Kalton and Kasprzyk 1986) Framework



$$\hat{y}_{mi} = b_{ro} + \sum b_{rj} z_{rj} + \hat{e}_{mi}$$

- Ratio, regression
 - Mean imputation (within class)
 - Random hot deck
- Deterministic vs. stochastic =>
$$\hat{e}_{mi} = 0, \neq 0$$
- Residuals \hat{e}_{mi} selected
 - Randomly from model
 - Randomly from respondents
- Deterministic methods can be made stochastic (e.g. Regression with residuals)

Deductive methods



- Logical (deductive) imputation
 - Direct result of edits
 - Known systematic biases
 - Use when exact relationships exist
- Historical imputation
 - Repeated economic surveys
 - Stable variables
 - Trend adjustments
 - Use when correlation over time stronger than between similar units



Current data methods

- Mean value imputation
 - Within imputation classes
 - Destroys distributions
 - Use as a last resort
- Hot deck imputation
 - Random (within class) hot deck
 - Sequential hot deck
 - Sorted hot deck
 - Use when little is known about nonrespondents
- Nearest neighbour imputation
 - Good for ignorable nonresponse
 - Use strong x-y relationships exist



Hot or Cold?

- “Deck” methods
- Nearest neighbour methods
- Current data or previous data?



Model based methods

- Ratio imputation
- Regression imputation
 - Abundant relationships
 - Nonresponse bias
 - Nonrandom but ignorable nonresponse
- Note: Parameters derived from current data
- Use with quantitative data when strong relationships exist



Multiple Imputation

- Why impute multiply?
- Proper imputation
- Details in section 6



Imputation methods

- Logical / Deductive
 - Mean
 - Ratio
 - Regression
 - Model
 - Probability imputation
 - Previous value / Historical
 - Trend (unit and group)
 - Cold-deck
 - Hot-deck
 - Nearest neighbour
 - Nearest neighbour's trend
 - Imputation with residuals



Statistics
Canada

Statistique
Canada

Canada



Logical or deductive imputation

- The missing value to impute is deduced from the edit rules.
Example: $X = Y + Z$
 $X = 10, Z = 8$ and Y is missing
 $\Rightarrow Y = 2$
- Sometimes called deterministic imputation
- Exact method
- Simple
- Often not considered to be imputation



Statistics
Canada

Statistique
Canada

Canada

Mean imputation



- The missing value is replaced by the mean of the respondents

$$\hat{y}_k = \frac{\sum y_k}{m}$$

- Easy to compute
- Does not require auxiliary information
- Assumes uniform nonresponse
- Destroys distributions
- Useful if performed within sub classes

Ratio imputation



- The missing value is replaced by the adjusted value of another variable

$$y_k = \frac{\sum_r y_k}{\sum_r z_k} z_k = \hat{R} z_k$$

- Requires an auxiliary variable (which may be another variable on the survey)
- Simple (but assumes no intercept)
- Robust to nonresponse which depends on Z



Regression imputation

- The missing value is replaced by other variables' adjusted value (using respondents)

$$\hat{y}_k = \hat{B}_0 + \hat{B}_1 z_1 + \dots + \hat{B}_J z_J$$

- Requires auxiliary variables
- Robust to nonresponse which depends on one or many Z variables



Model imputation

- The missing value is replaced by a value predicted using a model based on the respondents

$$\hat{y}_k = \hat{f}_r(k)$$

Example: Non-linear regression
Exponential model

- Requires auxiliary variables
- May produce impossible values

Probability imputation



- In the case of (0,1) variables, the missing value is replaced by the probability of obtaining a value of 1

$$\hat{y}_k = \hat{P}(y_k = 1)$$

- More precise than randomly choosing between 0 and 1 according to observed frequencies
- Reduces the variance (removes stochastic process)
- Yields impossible values (so usually flip a coin)

Previous value / Historical



- The missing value is replaced by the value declared at the previous occasion

$$\hat{y}_{k,t} = y_{k,t-1}$$

$$\Rightarrow \hat{y}_k = z_k$$

- Equivalent to ratio with $\hat{R} = 1$
- Requires the previous value
- Assumes no trend
- Requires files matching



Unit-trend imputation

- The missing value is replaced by the value declared at the previous occasion, but adjusted according to the trend of the unit

$$\hat{y}_{k,t} = \frac{z_{k,t}}{z_{k,t-1}} y_{k,t-1}$$

- Requires the previous value
- Requires an auxiliary variable
- Requires file matching
- Equivalent to ratio with only one record within the imputation class



Statistics Canada
Statistique Canada

Canada



Group-trend imputation

- The missing value is replaced by the value declared at the previous occasion, but modified according to a group trend

$$\hat{y}_{k,t} = \frac{\sum y_{k,t}}{\sum y_{k,t-1}} y_{k,t-1} = \hat{t} y_{k,t-1} = \hat{R} z_k$$

- Requires the previous value
- Equivalent to ratio
- Requires file matching
- Note: It is also possible to obtain the trend from an external source



Statistics Canada
Statistique Canada

Canada



Cold-deck imputation

- The missing value is replaced by a (randomly chosen) value from another file

$$\hat{y}_k = y_{l(k)}^{CD}$$

- “Another file” may be
 - subset of respondents on previous occasion
 - artificial data
 - other externally obtained data
 - any fixed data set
- Provides a plausible value
- Preserves the structure of respondents
- May introduce outliers
- Auxiliary information not required
- Assumes no difference between the two sources



Statistics
Canada

Statistique
Canada

Canada



Hot-deck imputation

- The missing value is replaced by a (randomly chosen) value from the “clean” respondents in the current file

$$\hat{y}_k = y_{l(k)}^{HD}$$

- Provides a plausible value
- Preserves the structure of respondents
- May introduce outliers
- Auxiliary information not required
- Can be
 - Random
 - Sequential
 - Sorted



Statistics
Canada

Statistique
Canada

Canada



Nearest neighbour imputation

- The missing value is replaced by the nearest neighbour's value (according to a distance function based on one or more auxiliary variables)

$$\hat{y}_k = y_{l(k)}^{PP}$$

- Provides a plausible value
- Requires auxiliary variables



Nearest neighbour's trend

- The missing value is replaced by the value reported at a previous occasion modified according to the trend of the nearest neighbour

$$\hat{y}_{k,t} = \frac{z_{l(k),t}^{NN}}{z_{l(k),t-1}^{NN}} y_{k,t-1}$$

- Requires the previous value
- More likely to preserve post-imputation edit rules for partial donor imputation



Imputation with residuals

- The missing value is replaced by a predicted value to which a randomly selected residual is added

$$\hat{y}_k = \hat{f}_r(k) + e_k^*$$

Example:

$$\hat{y}_k = \hat{R} z_k + e_k^*$$

- May require an auxiliary variable
- Increases the variability of the data
- Preserves the distribution (subject to having chosen the residuals wisely)
- Choice of residuals
 - Based on respondents
 - From a selected distribution



Chain imputation

- Nearest neighbour of prediction (Predictive mean matching)
- Prediction of nearest neighbour
- Logistic followed by model
- Other

Logistic imputation followed by model imputation



Description

First, logistic regression is used to determine the category and the missing value is replaced by a value predicted using a model adjusted on the respondents

1) Prediction of category c

(ex. 0 or >0)

2) $\hat{y}_{k,c} = \hat{f}_r(k)$

Comparison of the imputation methods



- Auxiliary information required
 - ⇒ All except
 - Logic
 - Mean
 - Hot-deck
 - Cold-deck (other file)
 - Matching
 - Previous value
 - Unit trend
 - Group trend

Comparison of the imputation methods



Nonresponse

- Uniform (MCAR)

⇒ All methods

- (MAR)

⇒ Methods which use auxiliary information

- (NMAR)

⇒ Response model

Comparison of the imputation methods



■ Computing speed (relative)

Slow: Nearest neighbour

- Medium: Stochastic methods

- Fast: Other

■ Complexity

Complex: Nearest neighbour

Some model methods

- Simple: Most others



Other imputation methods

- Pro-rating
- Historical revisions
- Manual adjustments



Pre-dissemination methods

- Pro-rating
 - Description:
The values of the components of a total are adjusted to the total
 - Example: $X + Y = Z$
 $X = 2$, $Y = 3$ and $Z = 6$
 X is imputed by $6 \cdot 2/5$ and Y by $6 \cdot 3/5$.
 - =>Corresponds to a ratio calculated on only one unit with responses acting as the auxiliary variables
 - Can be used to adjust all or some parts of a total after other imputation methods



Pre-dissemination methods

- Historical revision

Description:

When a published series is adjusted, the old values are adjusted to the level of the new publication

$$\hat{y}_k = \frac{\text{New total}}{\text{Old total}} y_k$$

=>Corresponds to ratio imputation with the auxiliary total calculated using the new method



Statistics
Canada

Statistique
Canada

Canada



Pre-dissemination methods

- Manual adjustment

Description:

Any adjustment performed by anyone involved in processing or analysis of the data

$$\hat{y}_k = \text{Adjustment} * y_k$$

$$\hat{y}_k = y_k + \text{Adjustment}$$

$$\hat{y}_k = \text{Adjustment} = z_k$$



Statistics
Canada

Statistique
Canada

Canada

Characteristics of nonresponse



Types of nonresponse

- Do respondents and nonrespondents have the same level for auxiliary variables?

(Is nonresponse uniform or does it depend on one of the auxiliary variables?)

- Is there past information on the respondents?

(Attempt at verifying whether nonresponse may depend on the Y variable)

Adjustment for nonresponse which depends on Y

Characteristics of nonresponse



Nonresponse depends on	Type	Action
1) No variable (it is assumed not to be in 3)	Uniform (MCAR)	Any method
2) One or more auxiliary variable	Non-confounded (MAR)	These variables must be used for a) Imputation and/or b) Creation of imputation classes
3) No variable (or Y according to previous data)	Confounded (NMAR)	a) Pretend uniform b) Use past data as auxiliary variables (like in 2) c) Adjust like (Rancourt et al, 94) d) Use an assumed response model

C. Issues



Required results

- Is a complete data set required?
 - Yes => Imputation
 - No => Re-weighting or imputation
- Who are the data analysts?
- External clients:
 - => Simple methods, identifiers
- Internal to the agency
 - => More complex methods

Issues



Nature of the data

- Categorical or continuous?
- What are the parameters of interest?
- What is the frequency of the data?

Sources of data

- Are there any other variables available?
- Are there previous occasions of the survey?
- Are there other sources which could be used through record linkage?

Issues



Editing

- Link between editing and imputation
- Variables involved in edits
- Edit classes

Auxiliary variables

- Evaluate quality / model
- Same occasion / previous occasion
- Use to create classes (not all combinations!)
- Verify relationships (assess model)
- Hierarchy of variables

Issues



Donor imputation

- Determine classes (donor pool)
- Imputed used as donors?
- Choose distance function
- Limit number of times a donor is used
- Keep track of donors
- For mass imputation / Data fusion

Issues



Protection of the distributions

- Donors or methods with residuals
- Univariate: higher moments
- Multivariate: correlations, etc.
- General model

5. Nonresponse treatment: Principles and approaches



Outline



- A. Context
- B. Re-weighting
- C. Imputation
- D. Issues
- E. Other related topics

A. Context



Full response

Population → Sample

Context



Nonresponse

Population → Sample → Respondents

Context



=> 2 approaches

1. Re-weighting
2. Imputation

B. Re-weighting



Description

Only complete respondents are kept (as soon as a value is missing, the record is completely discarded);

and

the weights of the records are kept and adjusted to take the nonrespondents into account.

Re-weighting



Characteristics

- Often simple
- Complete file
- Does not invent data
- Many software available
- Weights adjusted to eliminate or reduce bias
- Efficiency varying with the information used to compute weight adjustments

Re-weighting



Examples

1.	No	X1	X2	W
	1	3	-	6
	2	6	4	6
	3	-	-	6
	4	2	0	6
	5	3	1	6

We only keep 2, 4 and 5 and $W^* = 10$

2. Monthly survey :
Questionnaires returned after the end of the collection period
3. Two phases :
Respondents are the 2nd phase

Issues



- Re-weighting classes
- Number of classes
- Number of units per class
- Building classes or using the independent variable in a model.
- Number of sets of weights
- Using boundaries for the weights
- Software: Standard vs other (Sudaan, Carp, Wesvar, Vplex, Poulpe)
- Normalising the weight

Approaches



1. Using observed counts
 - ➔ Adjustment by the response rate
2. Using a response model
3. Using auxiliary data
 - ➔ Calibration

Adjustment by the response rate



Description

Within each weighting class, the response rate is calculated and its inverse is used to adjust the weights of respondents.

A complete data set (respondents) is then obtained and each record has an adjusted weight.

The total sum of weights is therefore the total number of units in the population.

Adjustment by the response rate



Method

From a sample with n units and m respondents, the response rate is calculated within each re-weighting class c

2 Cases :

1. Unweighted response rate : $Rate_c = \frac{m_c}{n_c}$

2. Weighted response rate : $Rate_c = \frac{\sum w_k}{\sum_{s_c} w_k}$

Then $\hat{Y}_r = \sum_{c=1}^C \frac{1}{Rate_c} \sum_{r_c} w_k y_k$

Adjustment by the response rate



Examples

- Auxiliary information not available
- Minor adjustments (high response rates)
- Post-stratification (=case 2)
- Census adjustment

Response probability models



Description

Within each re-weighting class, the response probability is modelled and its inverse is used to adjust the weights of the respondents.

A complete data set (respondents) is then obtained and each record has an adjusted weight based on the response model.

The total sum of weights is not necessarily the total number of units in the population.

Response probability model



Method

The quantity to estimate is $Y_U = \sum_U y_k$

If the Horvitz-Thompson estimator is used, we have

$$\hat{Y}_s = \sum_s \frac{y_k}{\pi_k} = \sum_s w_k y_k$$

If respondents only are available, one can use

$$\hat{Y}_r = \sum_r \frac{y_k}{\pi_k \hat{p}_k} = \sum_r w_k^* y_k$$

where \hat{p}_k is the estimated response probability of unit k

Some models



Uniform model

$$P(k \in r) = p_k, \hat{p}_k = \hat{p} = \frac{m}{n} \text{ for all } k$$

$$\hat{Y}_r = \sum_r \frac{y_k}{\pi_k \hat{p}} = \frac{n}{m} \sum_r w_k y_k$$

Class uniform model (or homogeneous response groups
– HRG)

$$P(k \in r) = p_k, \hat{p}_c = \frac{m_c}{n_c}$$

$$\hat{Y}_r = \sum_r \frac{y_k}{\pi_k \hat{p}_c} = \frac{n_c}{m_c} \sum_r w_k y_k$$

→ Corresponds to response rate adjustment

Using auxiliary information



Description

The estimates obtained from the respondents are adjusted to auxiliary known totals.

The auxiliary information may come from external sources.

This approach is also called calibration.

Using auxiliary information



Ratio adjustment

Known total for groups g are used :

$$\hat{Y}_{r-rat} = \sum_{g=1}^G \left(\frac{\sum_{U_g} x_k}{\sum_{r_g} x_k / \pi_k \hat{p}_k} \sum_{r_g} y_k / \pi_k \hat{p}_k \right)$$

Note 1 : If $x_k = 1 \quad \forall k$, then the method corresponds to post-stratification.

Note 2 : If groups g are equal or included within the re-weighting classes c , then the \hat{p}_k cancel each other when obtained using a model.

C. Imputation



Description

For each missing value (or identified as such), a **replacement** value is found.

The process is carried out to the best of the knowledge (according to the availability of the auxiliary information).

Imputation may be carried out manually or using a computer.

Imputation



Characteristics

- Complete sample
(allows for the use of complete data software)
- Utilise all data
- Consistent for different analysts
- « Invents » data
- Data after imputation may be misleading

Disadvantages of imputation



- The basic assumptions must be satisfied
- Can reduce the relationships between variables
- May lead users into believing in too high a data quality
- Simple concept
 - ➔ often performed without enough care
- Invents data

Imputation and Modelling



They are the same!

Modelling



- Which independent variables to use? (including interactions, higher orders, categorical variables and groups)
- Is the function linear?
- Is the variance of residuals constant?
- Are the errors independent?
- Are there outliers?
- Are the residuals normally distributed?


Modelling



- Test the significance of coefficients
- Create groups
- Tests on different data sets
- Robust methods
- Perform transformations

What to model?



- 1) Response probability $P(R = 1 | s)$
since S  R is unknown.

And/or

- 2) Variable of interest (y)

Example of the modelling process



Available

Y, X1, X2, X3, X4, X5, R

- 1) Help from a subject matter specialist
- 2) Modelling the variable of interest

Example of the modelling process



We find

$$Y = a + bX_1 + cX_2 + dX_1X_2 + \varepsilon_k$$

If

$$y_k^* = \hat{a} + \hat{b}X_1 + \hat{c}X_2 + \hat{d}X_1X_2$$

→ Imputation method: Multiple regression

Possibilities / methods



- X2 is judged not to be always important

$$Y = a + bX_1 + \varepsilon_k$$

$$y_k^* = \hat{a} + \hat{b}X_1$$

→ Imputation method: simple regression

Possibilities / methods



- If X1 and X2 are not available

$$Y = a + \varepsilon_k$$

$$y_k^* = \hat{a} = \bar{y}_R$$

→ Mean imputation

Possibilities / methods



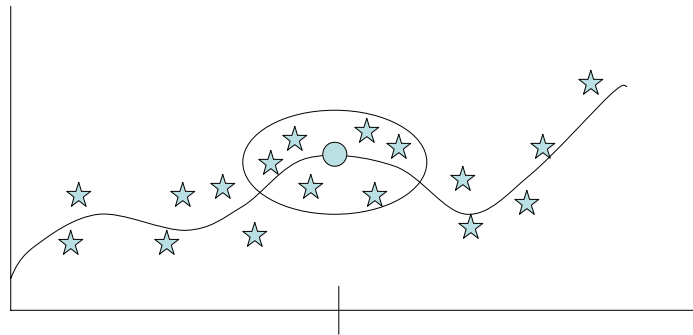
- If need for robustness
 - Many data sets
 - Outliers
 - Non-linear relationships

→ Imputation method: non-parametric regression

Non-parametric regression



- Use of points in the « neighbourhood »



Non-parametric regression



➡ Creation of floating groups

Limit case ! : Groups of size 1

➡ Nearest neighbour imputation

Question:



Which imputation method
is the best?

Question:



Which imputation method
is the best?
WRONG QUESTION!

Right question:



What is the best
resulting model?

Other question:



A nonresponse model or a data model?

Answer:



The strongest one to be favoured, given implementation context.

D. Issues



- Simplicity
- Participation of subject matter specialists
- Reduction of the use of manual imputation
- Importance of manual imputation
- Partial vs. total nonresponse



Imputation approach

- Flags MUST be produced:
 - Respondents - Nonrespondents
 - Imputation method
 - Imputation class
 - Auxiliary variable(s) used
 - Donor



Hierarchy of imputation methods

- First, methods using auxiliary information
- Use of information from the respondent (ex.: historical imputation)
- Previous value imputation for large units
- Methods with a stochastic component



Hierarchy of imputation levels

- Start at the domain level
- Obtain subject matter's opinion
- Pre-establish the levels
- LIMIT the number of levels
- Evaluate the model fit at EACH level



E. Imputation-Related Topics

- Longitudinal Surveys
- Mass Imputation
- Data Fusion

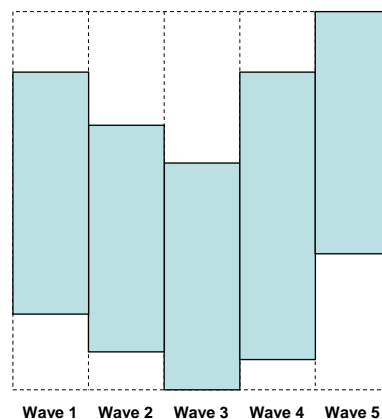
Imputation in longitudinal surveys



Issues

- Unit vs. item nonresponse
- Nonresponse waves (patterns)
- Attrition
- Tracing
- Cross-sectional imputation may introduce artificial transitions (change)
- Backward imputation

Imputation in longitudinal surveys





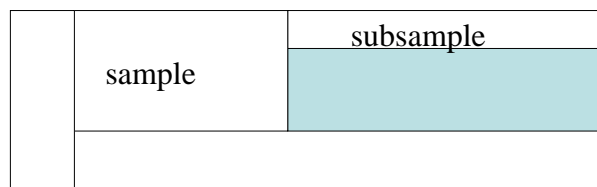
Longitudinal imputation methods

- Direct longitudinal substitution (historical imputation)
- Deterministic imputation of change (trends)
- Longitudinal regression imputation
- Longitudinal hot-deck
- Longitudinal nearest-neighbour



Mass imputation

- Sampling vs. subsampling



- Weighting vs. imputation
- e.g. Canadian Census of Construction

Disadvantages



- Can introduce biases when appropriate variables not controlled for
- Imputing large volumes of data (behaviour not well known)
- Theory not easily tractable
- Variance / covariance estimation a problem
- May need two files

Advantages



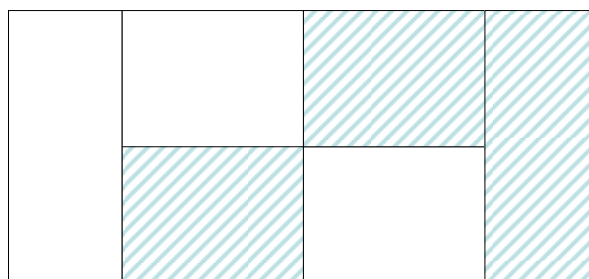
- Data “missing” at random
- Complete data set
- Quick ad hoc estimates
- Good if sample informative of subsample (especially if subsample not random, but ignorable)
- Good if weights difficult to calculate
- Can make better use of aux. information (preserve inter variable relationships)

Cautions



- Choice of imputation method important
- Imputation values must be flagged
- Critical and periodic evaluation needed
 - Simulation studies can be useful
- Dutch experiences are negative

Data Fusion



Common variables File 1 File 2 Imputed variables

- Completing files with missing data
- Adding variables from external sources

Data Fusion



Advantages

- Responds to greater data needs
- Makes use of auxiliary data

Disadvantages

- Heavy modelling involved or weak model

Data Replacement



- Administrative data instead of collected data when

$$y_k = x_k$$

- Can be viewed as imputation when using a more general model (e.g. $y_k = \beta x_k + \varepsilon_k$)

6. Variance due to nonresponse and imputation



Outline



- A. Context
- B. Methods
- C. Comparisons
- D. SEVANI



A. Context

- Simple case

$$\hat{Y}_s = \sum_S w_k y_k$$

- With auxiliary information

$$\hat{Y}_s = \sum_S a_k g_k y_k$$



What variance?

- Not the population variance!

$$s_{yU}^2 = \sum_U (y_k - \bar{y})^2$$

- Variances of the estimates

$$V(\hat{Y}_s)$$

over all possible sampling and response sets

Why estimate the variance?

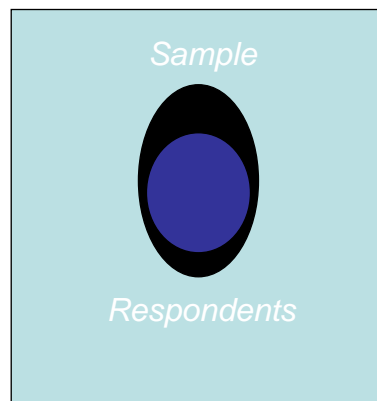


- Measure of the quality of the estimates
- Helps drawing the right conclusions
- Contributes in correctly informing users

Nonresponse



Population





Imputation and estimation

- Data after imputation:

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k & \text{if } k \in o \end{cases}$$

r: respondents

o: nonrespondents

hence

$$\hat{Y}_{\bullet s} = \sum_S a_k g_k y_{\bullet k}$$

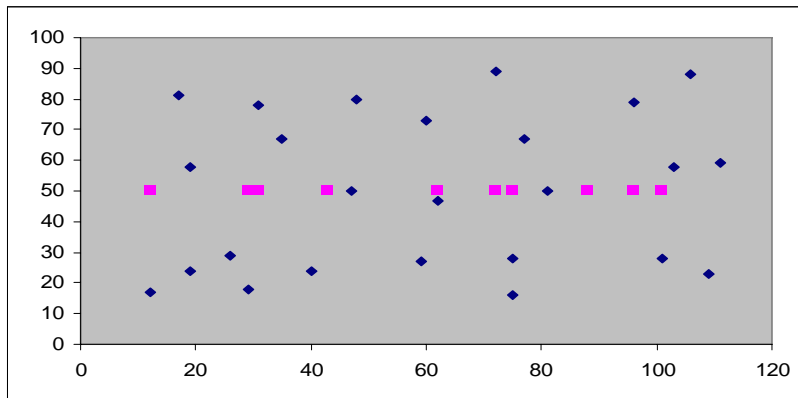


Why estimate the imputation variance?

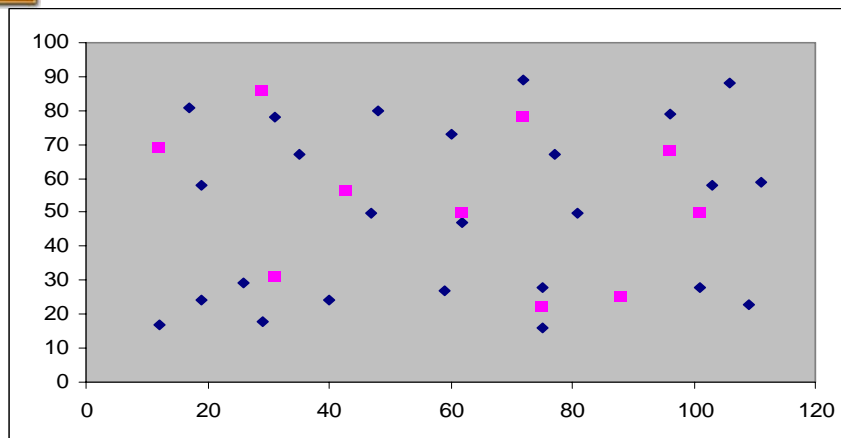
- Examples
- Total error
- Reasons



Example: Mean imputation



Example: Hot-deck imputation



Definition of the problem



Total error :

$$\hat{Y}_{\bullet s} - Y_U = (\hat{Y}_s - Y_U) + (\hat{Y}_{\bullet s} - \hat{Y}_s)$$

$$E_p E_q [\hat{Y}_{\bullet s} - Y_U]^2 = E_p E_q [(\hat{Y}_s - Y_U) + (\hat{Y}_{\bullet s} - \hat{Y}_s)]^2$$

If the bias is null, we have

$$V_{TOT} = V_{SAM} + V_{IMP} + 2V_{MIX}$$

Definition of the problem



We want to estimate

$$V_{TOT} = V_{SAM} + V_{IMP} + 2V_{MIX}$$

but only

$$\hat{V}_{ORD} = N^2 \frac{1-f}{n} \sum \frac{(y_{\bullet k} - \bar{y}_{\bullet s})^2}{n-1}$$

is available (in the case of simple random sampling without replacement).

- This estimator assumes that the data after imputation have the same variability as if the complete sample were available;
- This estimator under-estimates V_{SAM} and completely misses V_{IMP} .

Why estimate the imputation variance



- To give the right picture and know the impact of imputation

- Results from simulation studies

- Artificial population
- 50% nonresponse
- Uniform nonresponse
- Nearest neighbour imputation
- Ratio estimation

Results from simulations



$$V_{TOT} = V_{SAM} + V_{IMP}$$

$$28.03 = 9.33 + -$$

$$\hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{IMP} + \hat{V}_{MIX}$$

$$27.79 = 9.27 + 19.00 + (-0.48)$$

Why estimate the imputation variance ?



- To better allocate resources between the sample and the imputation/follow-up procedures
- Example (Percentage of total variance):

	V_{SAM}	V_{IMP}
a)	90%	10%
b)	30%	70%

B. Methods



- Two-phase approaches
- Reverse approaches
- Re-sampling approaches
- Multiple imputation

Two-phase approaches



- Nonresponse model
(two-phase)
- Data model
(model-assisted)

Two-phase approach



- Rao (1990), Rao & Sitter (1995)
- Assumption:
→ Response set = 2nd phase sample

$$\hat{V}_{\text{TWO-PHASE}} = \hat{V}_{\text{PHASE 1}} + \hat{V}_{\text{PHASE 2}}$$

Two-phase approach



Principle :

The respondents are assumed to form the second phase of a two-phase sample design.

Variance :

$$\hat{V}_{2P1} = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \sum_r (y_k - \bar{y}_r)^2 / (m-1) + N^2 \left(\frac{1}{m} - \frac{1}{n} \right) S_{er}^2$$

$$\hat{V}_{2P2} = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \hat{B}^2 S_{zs}^2 + 2N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \hat{B} S_{zer} + N^2 \left(\frac{1}{m} - \frac{1}{N} \right) S_{er}^2$$

where

$$\hat{B} = \sum_r y_k / \sum_r z_k, \quad S_{zer} = \sum_r e_k z_k / (m-1)$$

and $e_k = y_k - \hat{B}z_k$.

Model assisted approach



- Särndal (1990, 1992), Deville & Särndal (1991, 1994)
- Using an imputation model

$$\hat{V}_{\text{MODEL}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + \hat{V}_{\text{MIX}}$$



Model assisted approach

Principle :

Use a model of the form

$$\xi : y_k = \beta z_k + \varepsilon_k; \quad E_\xi(\varepsilon_k) = 0;$$

$$E_\xi(\varepsilon_k^2) = \sigma^2 z_k; \quad \text{and} \quad E_\xi(\varepsilon_k \varepsilon_{k'}) = 0 \quad \text{for} \quad k \neq k'$$

to construct an estimator for each of the terms in V_{TOT} .

$$\text{Then we have : } \hat{V}_{TOT} = \hat{V}_{SAM} + \hat{V}_{IMP} + 2\hat{V}_{MIX}$$

(Särndal 90, 92)

(Deville and Särndal 91, 94)



Model assisted approach

Two terms for \hat{V}_{SAM}

1) $\hat{V}_{ORD} = N^2 \frac{1-f}{n} S_{y \bullet s}^2$, calculated on the data after imputation, with

$$S_{y \bullet s}^2 = \frac{1}{n-1} \sum_s \{y_{\bullet k} - (\sum_s y_{\bullet k} / n)\}^2$$

Usually, \hat{V}_{ORD} under-estimates V_{SAM}

2) \hat{V}_{DIF} , constructed to satisfy

$$E_\xi\{\hat{V}_{DIF}\} = \frac{N^2(1-f)}{n} E_\xi\{S_{ys}^2 - S_{y \bullet s}^2\}$$

We obtain $\hat{V}_{SAM} = \hat{V}_{ORD} + \hat{V}_{DIF}$



Model assisted approach

Notes

- General method which allows for the derivation of each estimator;
- Ex: Nearest neighbour imputation:

$$\hat{V}_{\text{IMP}} = \left\{ \sum_r \left(\sum_{o_d} w_k \right)^2 z_l + \sum_{o_d} w_k^2 z_k \right\} \hat{\sigma}^2$$

with

$$\hat{\sigma}^2 = \sum_r (y_k - \hat{B}z_k)^2 / (m - 1)$$



Reverse approach

- Shao and Steel (1999)
- Inverse approach

$$U \Rightarrow S \Rightarrow R$$

Becomes

$$U_r \Rightarrow S_r \Rightarrow R$$

Re-sampling approaches



- Jackknife
- Bootstrap
- BRR

Jackknife



- Rao & Shao (1992)
- At each iteration, adjust imputed values when deleting a respondent

$$\hat{V}_{\text{JKNF}} = \frac{n-1}{n} \sum_{j \in S} \left(\hat{Y}_{j, \text{adjusted}} - \hat{\bar{Y}}_{\text{adjusted}} \right)^2$$

Jackknife technique



Principle :

Iterative method where a unit is removed at each iteration and the estimator is re-calculated. Then, the imputed values are adjusted when the removed unit is a respondent.

Ordinary jackknife :

$$\hat{V} = \frac{n-1}{n} \sum_{j \in s} (\hat{Y}_{\bullet s}^{(j)} - \hat{Y}_{\bullet s})^2$$

Corrections for imputation :

$$y_{\bullet k}^{(aj)} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k + a_k^{(j)} & \text{if } k \in o \text{ and } j \in r \\ \hat{y}_k & \text{if } k \in o \text{ and } j \in o \end{cases}$$

Jackknife technique



Variance :

$$\hat{V}_{JK} = \frac{n-1}{n} \sum_{j \in s} (\hat{Y}_{\bullet s}^{(aj)} - \hat{Y}_{\bullet s}^{(a)})^2$$

where

$$\hat{Y}_{\bullet s}^{(aj)} = \frac{N}{n-1} \sum_{k \neq j \in s} y_{\bullet k}^{(aj)}$$

and

$$\hat{Y}_{\bullet s}^{(a)} = \frac{1}{n} \sum_{j \in s} \hat{Y}_{\bullet s}^{(aj)}$$

(Rao and Shao 92)

Bootstrap



- Shao & Sitter (1996)
- For each sample, the imputation process is reproduced

$$\hat{V}_{\text{BOOT}} = \frac{1}{B} \sum_{b=1}^B \left(\hat{Y}_{b,boot} - \hat{Y}_{boot} \right)^2$$

BRR (Balanced repeated replication technique)



Principle :

The sample is divided into sub-samples and within each, the imputed data are adjusted.

Variance :

$$\hat{V}_{\text{BRR}} = \frac{1}{R} \sum_{r=1}^R \left(\hat{Y}_r - \hat{Y} \right)^2$$

(Shao, Chen and Chen 98)

Description



Description :

M sets of data are completed from the missing data predictive distribution.

M analyses are performed and the results are then combined.

Inference is achieved using the multiply-imputed data sets.

(Rubin 87)

Purpose



- Produce a consistent analysis;
- Incorporate knowledge of the person doing the imputation;
- Produce complete data sets;
- Reflect the uncertainty present in the data after imputation;
- Create data bases which can be released to users.

Estimation



The aim is to estimate a parameter such as

$$T = \sum_U y_k$$

Let

$\hat{t}_{\cdot j}$ Estimator on the j th completed data set.

Point estimator :

$$\hat{T}_{IM} = \frac{\sum_{j=1}^M \hat{t}_{\cdot j}}{M}$$

Variance



Variance estimator :

Let

$\hat{v}_{\cdot j} = \hat{v}(\hat{t}_{\cdot j})$ (on the j th data set)

$$\bar{V}_{WITHIN} = \sum_{j=1}^M \frac{\hat{v}_{\cdot j}}{M} \text{ (internal variance)}$$

$$\bar{V}_{BETWEEN} = \sum_{j=1}^M \frac{(\hat{t}_{\cdot j} - \hat{t}_{IM})^2}{M-1}$$

$$\hat{V}_{MI} = \bar{V}_{WITHIN} + \frac{M+1}{M} \bar{V}_{BETWEEN}$$

C. Comparisons



- Imputation with residuals
- “Moving” slope
- Proper imputation

Comparison of the approaches



- Breakdown of the approaches
- Number of imputations
- Other aspects

Imputation variance



- Provide sampling-imputation breakdown:
 - Multiple imputation
 - - Model assisted
 - - Two-phase
 - - Hot-deck
- Others do not

Number of imputations



- Single imputation:
 - Model assisted
 - Two-phase
 - Jackknife
- Others require multiple imputations

Other comparisons



- Identifiers (all but multiple imputation)
- Uniform nonresponse (two-phase)
- Models
- Users (multiple imputation)

Conclusion



- Problem common to all surveys
- Important issue
- Several methods
- Importance of distinguishing
 V_{SAM} and V_{IMP}

D. SEVANI



System for Estimation of the Variance due to Nonresponse and Imputation

Characteristics



- SAS – based
- Production system
- Variance due to nonresponse and imputation

Framework



- Quasi-multi-phase approach
- Nonresponse
- Imputation

Imputation



- Linear regression
- Auxiliary value
- Nearest neighbour

Outputs



Provides

- V_{nrp}
- V_{imp}
- V_{total}

7. Software and Quality assessment



Outline



- A. Imputation software
- B. Simulation studies
- C. Genesis
- D. Measuring quality

A. Imputation Software



- Types of Systems
- Historical Perspective
- International Perspective
- Quick Overviews

Types of systems



- Manual or automated process?
- Specialized or general purpose system?
- Tailor-made or generalized?
- Concentrate on imputation (not editing for follow-up, selective editing, etc.)

Historical perspective



- Processing systems (all encompassing)
- Specialized, tailor-made systems (include editing and imputation)
- Fellegi-Holt problem
- Fellegi-Holt solution
 - Edits drive imputation actions
 - Preserve distributions
 - Minimum change
- Many systems written as a result in various degrees of generality

Generalized systems



- High costs of data processing
- Need for fast processing
- Importance of comparisons
- A survey can be broken down into smaller steps
- Need for reproducibility
- Advancement of computer sciences (hardware as well as software)



Advantages and disadvantages

Advantages:

- State-of-the-art methods available
- Constant technical support available
- Reproducibility
- Greater consistency
- Flexibility

Disadvantages:

- High costs of creating
- Users must still evaluate results (no guarantees)
- External control



International examples

- Processing Systems
 - Blaise (Netherlands)
- Specialized Systems
 - IMPS (USBC)
- Fellegi-Holt Systems
 - CanEdit (Canada)
 - AERO (Hungary)
 - DIA (Spain)
 - SCIA (Italy)

International examples



- Modified F-H for quantitative data
 - NEIS/GEIS (Canada)
 - SPEER (USBC)
 - Aggies (NASS)
 - Cherry Pie/LEO (Netherlands)
- Minimum change systems
 - NIM/CANCEIS (Canada)
 - SEDDIM/DIESIS (Italy)
- Euredit
 - looking into neural networks

Quick overviews



- Fellegi-Holt systems
- GEIS type systems
- NIM type systems

Fellegi-Holt systems



- Mostly for qualitative data
- Satisfy three F-H principles
- Users specify conflict rules
- Full set of implied edits is generated (weak link)
- Minimum number of fields to be imputed is identified
- Imputation by hot-deck is customary but not necessary
- Imputed record must satisfy all edits

GEIS type systems



- Mostly for quantitative data
- Based on F-H principles
- Users specify edits as linear inequalities (disadvantage)
- Minimal set of edits is identified
- Minimum number of fields to be imputed is identified by means of linear programming (weak link)
- Any imputation method possible, nearest neighbour often the choice
- Imputed record must satisfy the edits

NIM type systems



- For both qualitative and quantitative data (but needs to be tested)
- Reverses search-for-donors and identification-of-fields-to-impute operations
- Users specify edits (virtually any form)
- “Nearest neighbours” are found (minimum change options)
- Must use donors (weak link)
- Minimum change while passing all edits

Future outlook



- Modularizing generalized systems
- Combining qualitative and quantitative data (Fellegi-Holt type system)
- Neural nets
- Evaluation software (SEVANI / GENESIS)

B. Simulation Studies

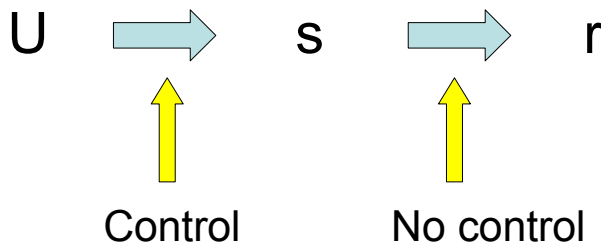


- Context and definition
- Goals
- Characteristics
- Implementation
- Notes

Context and definition



- Assessing quality



Context and definition



- Computer program
- Controlled conditions
- Large number of iterations

→ Monte Carlo experiment

Goals



- Learn / confirm properties
- Determine potential impacts
- Better understand methods
- Compare methods under some conditions

Characteristics



- Provides quantitative description
- Helps discover unforeseen situations
- Can be tedious
- Never the truth!

Structure



0. Create / choose population

1. Population characteristics

Many iterations



1.1 Sample selection

1.1.1 Generation of response

1.1.2 Basic calculations

2. Summary Summary statistics

3. Comparisons

Implementation



- Population
- Samples
- Response sets
- Basic calculations
- Summary measures
- Comparisons

Population



- True population
 - Actual population
 - Sample (with imputed values)
 - Response set
- Generated
 - From parameters modeled on the realized sample
 - From a known distribution

Sample selection



- Sampling scheme
- Sample size
- Number of samples (iterations)

Generation of a response set



- Response mechanism
- Response model
- Number of response sets
- Expected number of respondents

Basic calculations



- Estimator ($\hat{\theta}$)
- Variance estimator $\hat{V}(\hat{\theta})$
- Confidence interval

Summary statistics



- Average
- Variance
- Number of times interval covers true value
- Distributions

Comparisons



- $BIAS(\hat{\theta}) = AV_{MC}(\hat{\theta}) - \theta$
- $BIAS[\hat{V}(\hat{\theta})] = AV_{MC}[\hat{V}(\hat{\theta})] - VAR_{MC}(\hat{\theta})$
- Coverage of confidence interval:
- AV_{MC} (# times interval covers true value θ)
- Relative bias
- Mean squared error
- ⇒ Monte Carlo error

Notes



- Not Bootstrap or Multiple imputation
- Not the truth
- Very useful

Impact of imputation



- Bias
- Variance

Bias



Always unknown, but we can evaluate

- % of the total imputed
- Record linkage with external files
- Evolution in time
- Evaluation by subject matter specialists
- Simulations



C. GENESIS



GENeralised

Simulation

System

Characteristics



- Simulation system
- SAS – based
- Modules
 - Full response
 - Imputation
 - Classes

Full Response Module



- Sampling schemes (SRS, PPS)
- Estimators (H-T, ratio, regression)
- Relative Bias
- MSE
- Graphics

Imputation Module

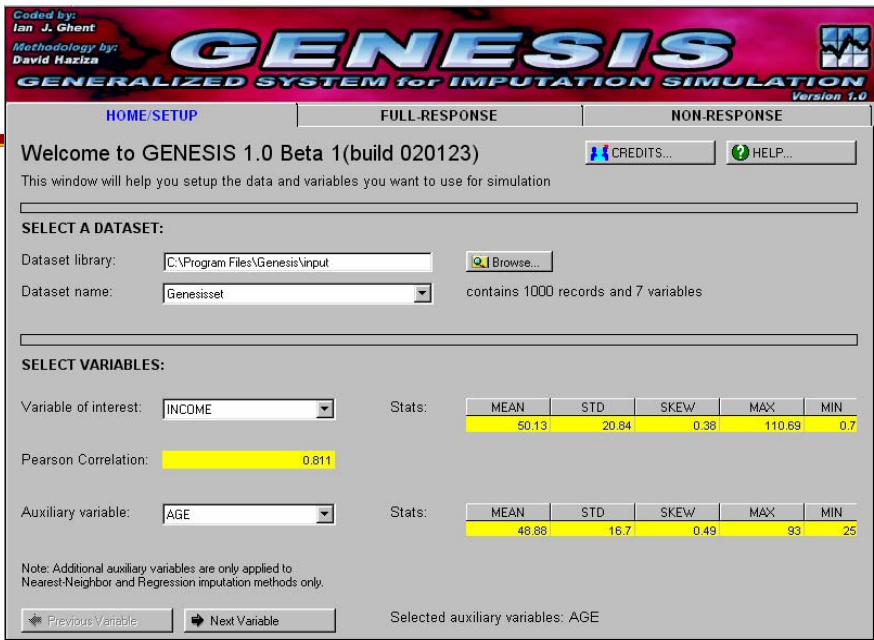


- Nonresponse (MCAR, MAR, NMAR)
- Imputation methods
- Variance due to imputation
- Monte-Carlo measures

Classes Module



- Cross classification by variables
- Scores approach using \hat{y} and \hat{p}
- Monte-Carlo measures



GENESIS
GENERALIZED SYSTEM for IMPUTATION SIMULATION
Version 1.0

HOME/SETUP FULL-RESPONSE NON-RESPONSE

Welcome to GENESIS 1.0 Beta 1(build 020123) CREDITS... HELP...

This window will help you setup the data and variables you want to use for simulation

SELECT A DATASET:

Dataset library: C:\Program Files\Genesis\input Browse...

Dataset name: Genesisset contains 1000 records and 7 variables

SELECT VARIABLES:

Variable of interest: INCOME Stats: MEAN STD SKEW MAX MIN

50.13	20.84	0.38	110.69	0.7
-------	-------	------	--------	-----


Pearson Correlation: 0.811

Auxiliary variable: AGE Stats: MEAN STD SKEW MAX MIN

48.88	16.7	0.49	93	25
-------	------	------	----	----

Note: Additional auxiliary variables are only applied to Nearest-Neighbor and Regression imputation methods only.

Previous Variable Next Variable Selected auxiliary variables: AGE

 Statistics Canada / Statistique Canada

Canada

D. Measuring Quality



Examples :

- 1) Nonresponse rate
- 2) Imputation rate
 - Before imputation
 - After imputation
 - By method
- 3) Number of failed edit rules;
- 4) Number of times donors are used;
- 5) Number of attempts for finding donors;
- 6) Number of units by cause of nonresponse.

Using auxiliary information in comparisons



- 1) Macro-editing
 - Comparison of rates against other surveys;
- 2) Micro-editing
 - Comparison with previous occasions;
 - Comparison with other sources; (micro-matching)

Studies



- 1) Size of imputation classes;
- 2) Variance due to imputation;
 - Magnitude;
 - Importance relative to variance due to sampling;
- 3) Variation of nonresponse through time.
- 4) Importance of the response burden

Informing users



- Prevention measures
- Identifiers
- Importance of imputation
- Precision

Identifiers



1. Respondents – nonrespondents;
2. Imputation / re-weighting methods;
3. Classes;
4. Donor;
5. Hierarchy of methods;
6. Hierarchy of levels

Precision



1. Size of imputation classes;
2. Variance due to imputation;
3. Total variance;
4. Percentage of variance due to imputation;
5. Percentage of variance due to sampling.

8. Examples and other topics



Outline



- Key references
- Statistics Canada software
- Statistics Canada examples
- Imputation-related activities at Statistics Canada
- Questions

Key References



Some internet sites

Statistics Canada Quality Guidelines

<http://www.statcan.ca/english/freepub/12-539-XIE/12-539-XIE.pdf>

Statistical Data Editing Workshop

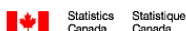
<http://www.unece.org/stats/documents/2003.10.sde.htm>

Euredit Project

<http://www.cs.york.ac.uk/euredit>

Multiple Imputation Online

www.multiple-imputation.com



Canada

Key References



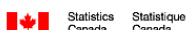
Papers & Books

BANKIER, M., LACHANCE, M., POIRIER, P., "A Generic Implementation of the New Imputation Methodology", *Proceedings of the Survey Research Methods Section*, 548-553, 1999.

BOUCHER, L., SIMARD, J.-P., GOSSELIN, J.-F. "Macro-editing, a case study: Selective editing for the Annual Survey of Manufactures conducted by Statistics Canada", *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 362-367, 1993.

FELLEGI, I.P., HOLT, D., "A systematic approach to automatic edit and imputation", *Journal of the American Statistical Association*, 71, 17-35, 1976.

FELX, P. RANCOURT, E., "Applications Of Variance Due To Imputation In The Survey Of Employment, Payrolls And Hours" *Methodology paper*, BSMD 2001-009E



Canada



GRANQUIST, L. and KOVAR, J.G., "Editing of survey data: How much is enough?" in *Survey Measurement and Process Quality*, Lyberg, L. et al eds., J. Wiley and Sons, New York, 1997.

HIDIROGLOU, M.A., BERTHELOT, J.-M. "Statistical editing and imputation for periodic business surveys", *Survey Methodology*, 12, 73-83, 1986.

KALTON, G., KASPRZYK, D., "The treatment of missing survey data", *Survey Methodology*, 12, 1-16, 1986.

LIU, T., RANCOURT, E., "Constrained Categorical Imputation for Nonresponse in Surveys", ICSN 1999, Portland.

LEE, H., RANCOURT, E., SÄRNDAL, C.-E., "Experiment with variance estimation from survey data with imputed values", *Journal of Official Statistics*, 10, 231-243, 1994.



LEE, H., RANCOURT, E., SÄRNDAL, C.-E., "Variance estimation in the presence of imputed data for the Generalized Estimation System", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389, August 1995.

LUNDSTRÖM, L., and SÄRNDAL, C.-E. *Estimation in surveys with nonresponse*, J. Wiley and Sons, 2005.

OH, H.L., SCHEUREN, F.J., "Weighting adjustment for unit nonresponse", in *Incomplete Data in Sample Surveys*, vol. 2, ed. : W.G. Madow, I. Olkin and D.B. Rubin, New York : Academic Press, 143-184, 1983.

OUTRATA, E., CHINNAPPA, B.N., "General survey functions design at Statistics Canada", *Bulletin of the International Statistical Institute*, 53 : 2, 219-238, 1989.

PANEL ON INCOMPLETE DATA, *Incomplete Data in Sample Surveys*, 3 volumes, Academic Press, 1983.



PIERZCHALA, M., "A review of three editing systems", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 111-120, 1990.

RANCOURT, E., BEAUMONT, J.-F., HAZIZA, D., MITCHELL, C., "Statistics Canada's New Software To Better Understand And Measure The Impact Of Nonresponse And Imputation", Conference Of European Statisticians, October 2003.

RANCOURT, E. (2001). Edit and Imputation: From suspicious to scientific techniques. Proceeding Actes, International Association of Survey Statisticians, 634-655.

RANCOURT, E., LEE, H., SÄRNDAL, C.-E., "Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse", *Survey Methodology*, 20, 137-147, 1994.

RAO, J.N.K., SHAO, J., "Jackknife variance estimation with survey data under hot-deck imputation". *Biometrika*, 79, 811-822, 1992.

RAO, J.N.K., SITTE, R.R., "Variance estimation under two-phase sampling with application to imputation for missing data", *Biometrika*, 82, 453-460, 1995.



RUBIN, D.B., "Basic ideas of multiple imputation for nonresponse", *Survey Methodology*, 12, 37-47, 1986.

SÄRNDAL, C.-E., "Methods for estimating the precision of survey estimates when imputation has been used", *Survey Methodology*, 18, 241-252, 1992.

SÄRNDAL, C.-E., "Estimation in Surveys with Nonresponse", *Sixteen Lundström*, J. Wiley & Sons, 2005.

SHAO, J., SITTE, R.R., "Bootstrap for imputed survey data", *Journal of the American Statistical Association*, 91, 1278-1288, 1996.

SHAO, J., STEEL, P., "Variance estimation for survey data with composite imputation and nonnegligible sampling fractions", *Journal of the American Statistical Association*, 94, 254-265, 1999.

Statistics Canada Software



- CANCEIS (NIM)
- BANFF (GEIS)
- IMPUDON
- GENESIS
- SEVANI

Statistics Canada Examples



- Census
- Tax replacement
- Survey of Employment, Payrolls and Hours
- Survey of Household Spending

Imputation-Related activities at Statistics Canada



- Quality Guidelines
- Committee on Imputation Practices (COPI)
- Imputation Research
- Imputation Bulletin



Questions, Discussion?



Thank you!

Varianza

SEVANI: software estimación varianza debida a imputación. Métodos que utiliza?

¿GENESIS no hace estimaciones de varianza?

¿Por qué dos software diferentes?

MIR(I)

- Aplicación que integra aplicaciones de homogeneización, validación, depuración e imputación. Variables cualitativas.
- **Imputación determinística transversal (hot-deck):** emplea el método deductivo mediante el cual el dato faltante se deduce de otra/s variable/s de la estadística.
- **Imputación determinística longitudinal (cold-deck):** Se asigna determinísticamente un valor a partir de información auxiliar de períodos de tiempo anteriores

MIR(II)

- Imputación aleatoria transversal: Se seleccionan una serie de variables de agrupación y se asigna aleatoriamente según la distribución de la variable en la subpoblación correspondiente.(hot-deck)
- Imputación aleatoria longitudinal: El análisis es similar al anterior empleando información de estudios previos (cold-deck)

MIR(III)

- Nuevo método hot-deck (aleatorio):
 - Selección registro donante completo (ó al menos más de 1 variable)
 - Criterio proximidad para definir grupos donantes y receptores
 - Selección aleatoria del donante dentro cada grupo