MANCHESTER 1824

The University of Manchester

Total Survey Error, Handling Missing Data, Statistical Data Editing and Imputation

Natalie Shlomo

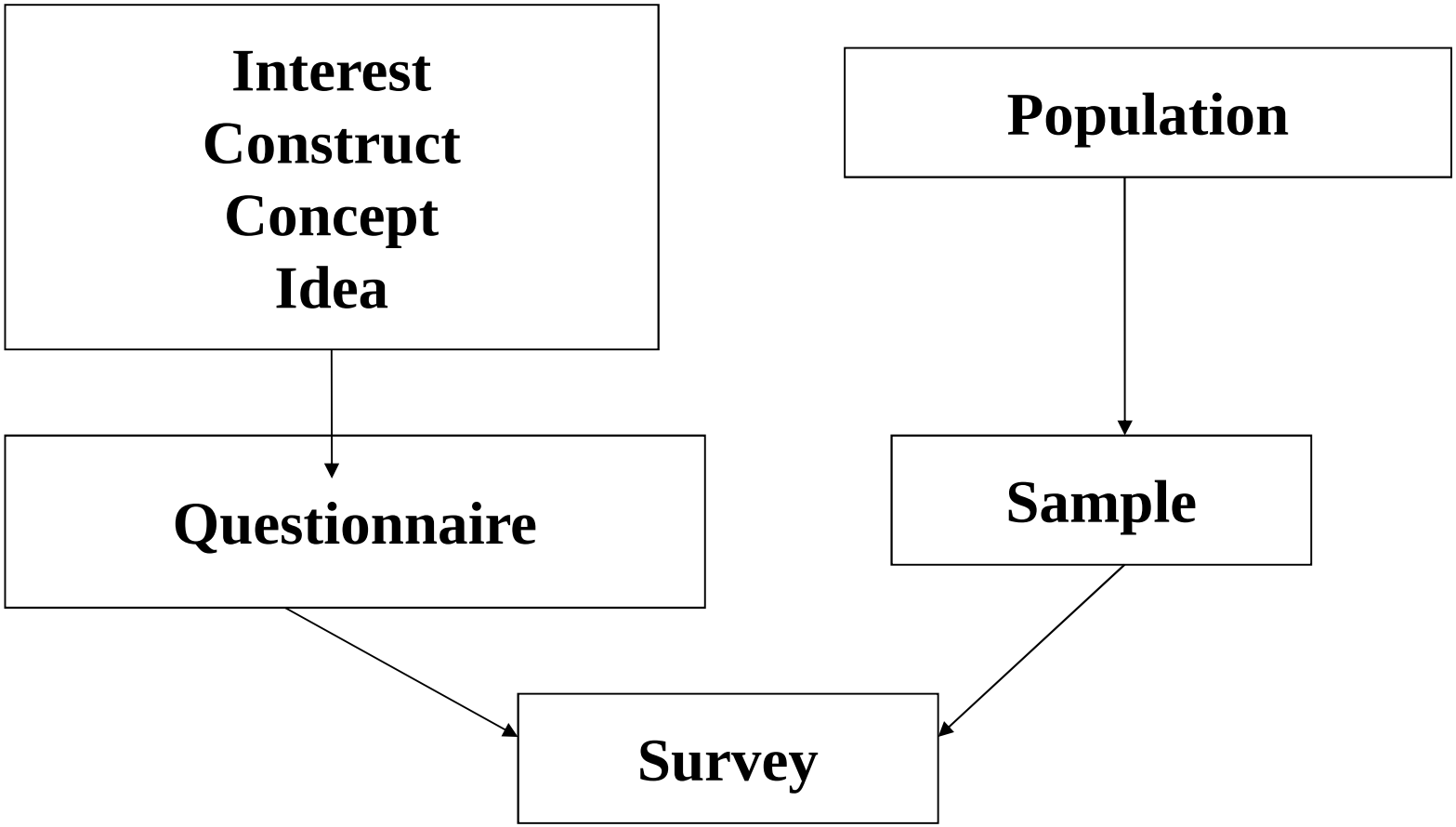
Natalie.Shlomo@manchester.ac.uk

Topics Covered

- Total Survey Error
- Unit non-response vs Item non-response
- Mechanisms of Response
- Compensating for item non-response – imputation methods
- Compensating for unit non-response – survey weights
- Statistical Data Editing

Total Survey Error

Measurement and Representation



Survey design: from abstract to concrete
Inference: from concrete to abstract

Total Survey Error

Measurement:

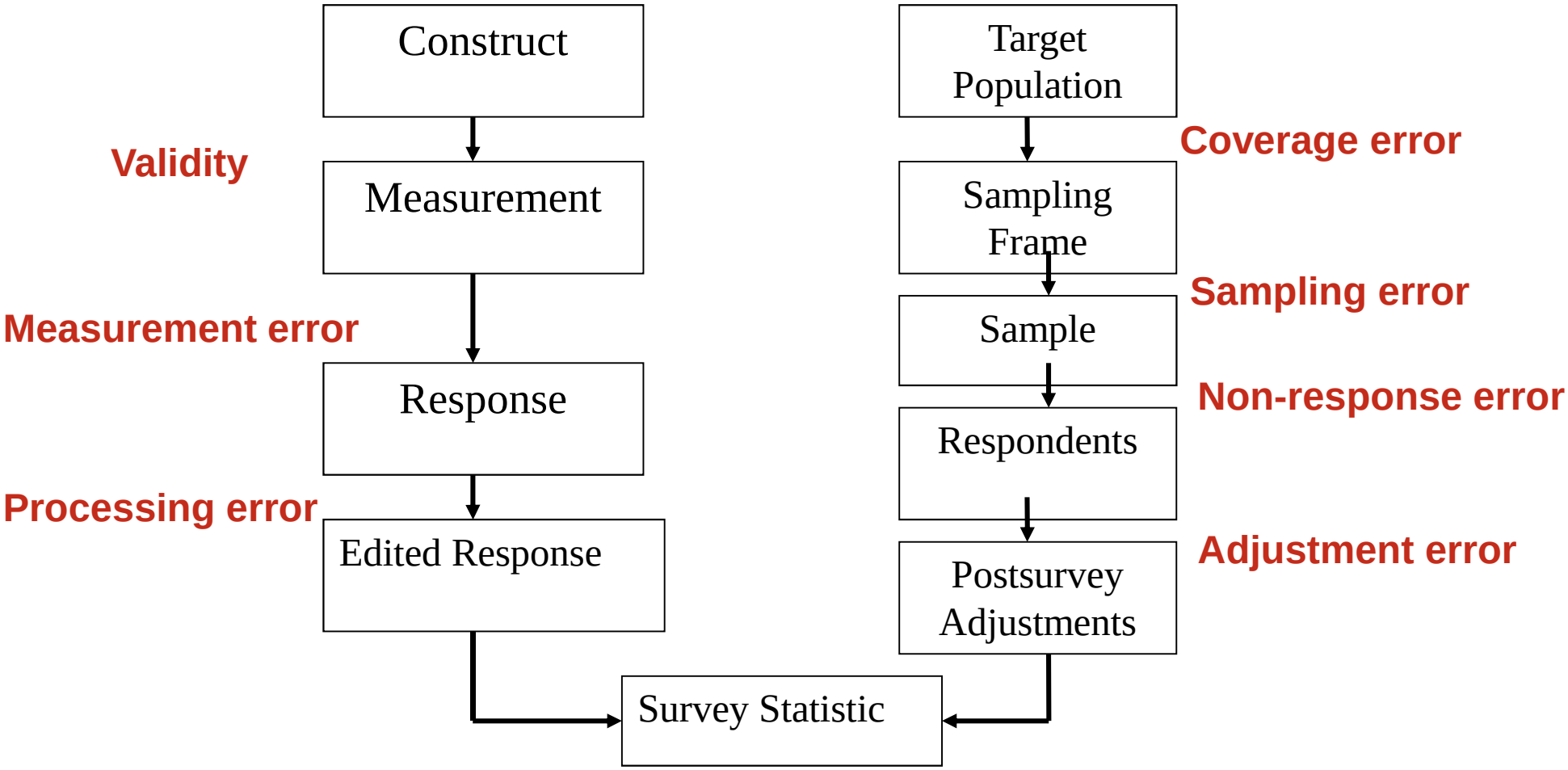
- Validity is the extent to which the measures reflect the underlying construct
- Measurement error: departure from the true value to that reported, i.e. response bias where bias is a systematic distortion of a response process; may result from questionnaire, mode of data collection, interviewers, etc.
- Processing error:
  - removal of outliers impacts on results (might be wrong to exclude an outlier, example charitable giving)
  - coding (text answers coded into categories), might be done differently by different people, i.e. poor training of coders
  - Imputation of missing data

Total Survey Error

Representation:

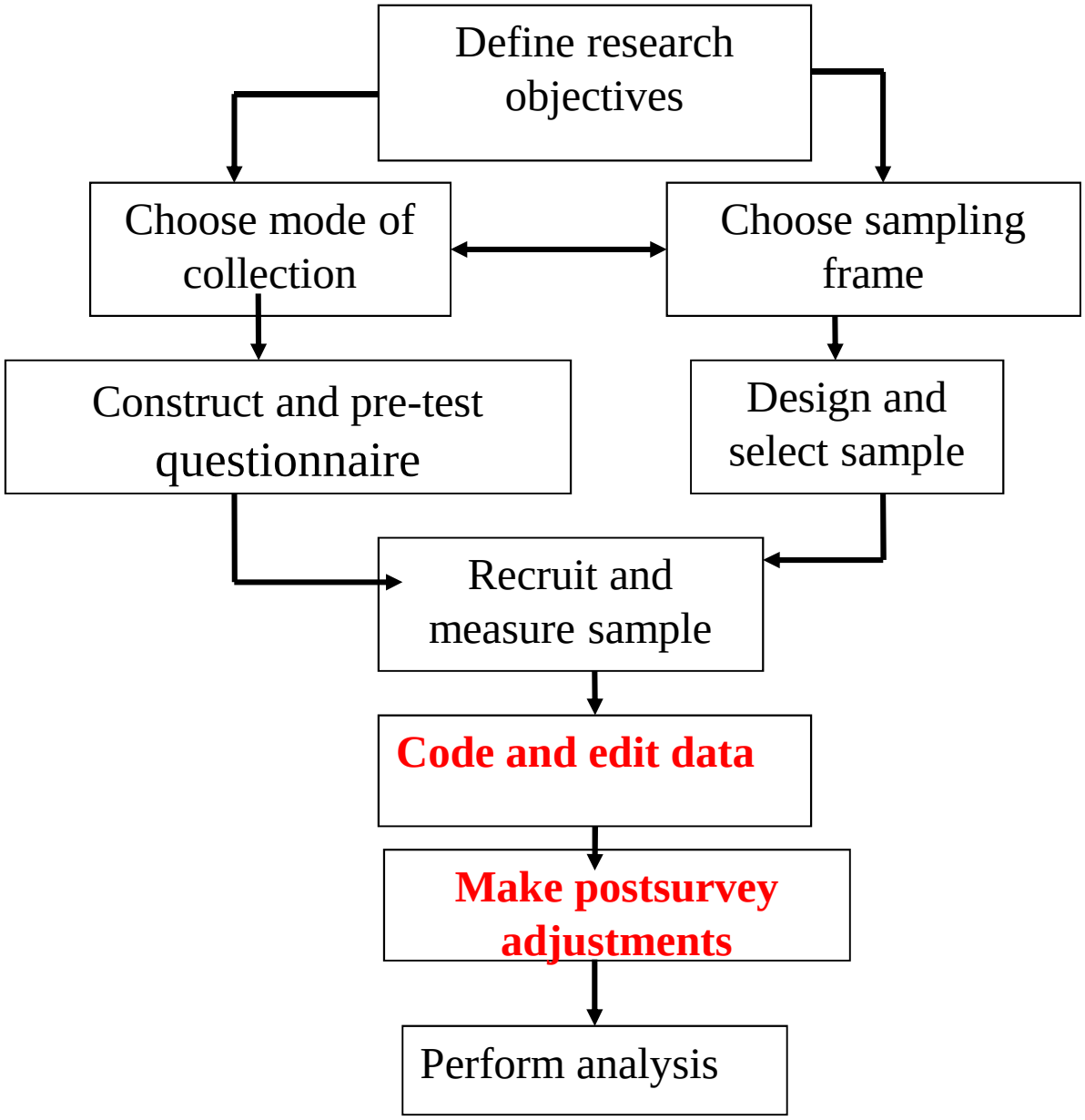
- Coverage error, e.g. telephone surveys: lower coverage of poor people and people who possess only mobile phones (undercoverage); ineligible units and duplicates (overcoverage) businesses. Two things important: how many people not covered. How different are covered and not covered people?
- Sampling error resulting from the variability in using a randomly selected fraction of the population
- Nonresponse error (bias) results from how different nonrespondents are from respondents, in particular if the missingness is related to the target variable
- Adjustment error in estimation increase variability (imprecision)

Components of the total survey error

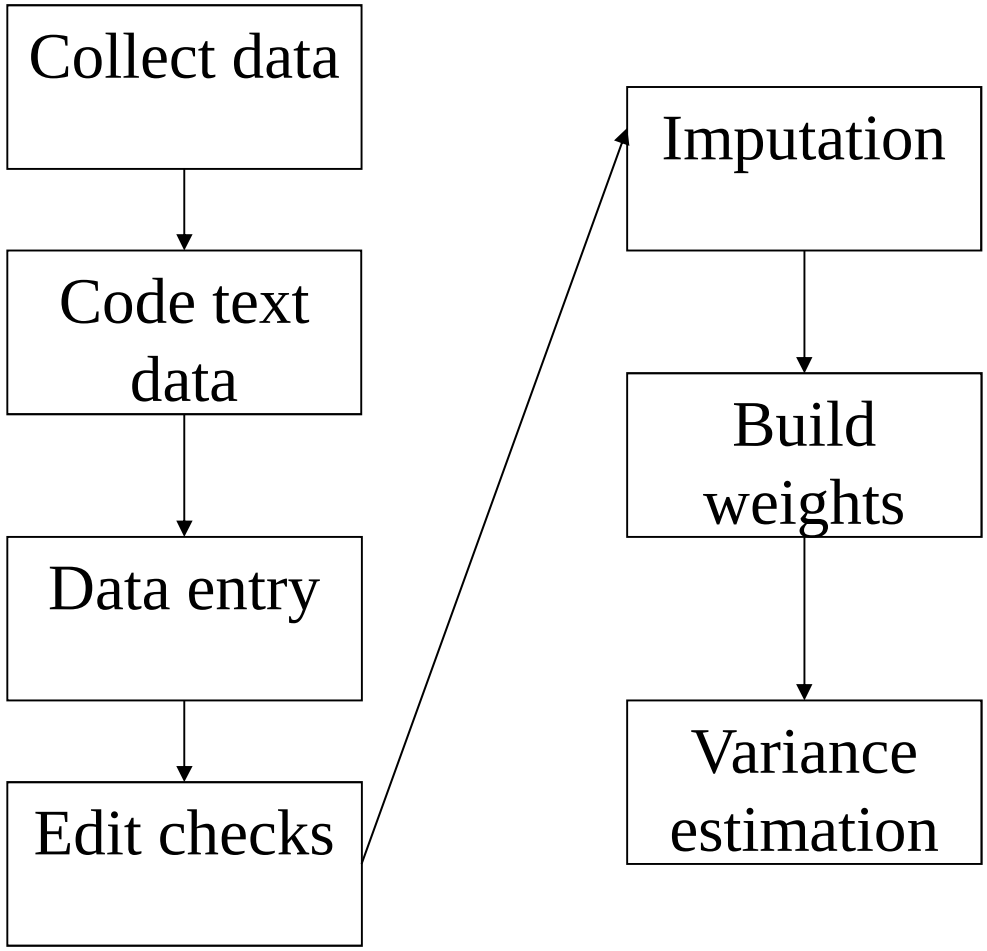


Aim: minimise total survey error by minimising errors between steps

Survey Process



Survey processing



Unit and Item Non-response

Definitions

Non-response is the failure to obtain complete measurements on the (eligible) survey sample (eligible = in-scope = target population). It occurs in almost all surveys.

Unit non-response – *no measurements on unit*

Item non-response – *measurement for some items missing for unit*

Under-coverage arises when there are units in target population but not in the (sampled) study population. Distinct but related to non-response

Non-response is important because of: potential bias; increased variance; cost implications; quality perceptions

Sources of Unit Non-response

- non-contact
- failure to locate/identify the sample unit or to contact the sample unit
- non co-operation
- refusal of sample unit to participate
- inability to respond
- inability of sample unit to participate (e.g. due to ill health, language barriers)
- other, e.g. accidental loss of data/questionnaire

Sources of Item Non-response

- **Respondent**
  - answer not known
  - refusal (sensitive or irrelevant question)
  - accidental skip
- **Interviewer**
  - does not ask question
  - does not record response
- **Processing**
  - response rejected at editing stage

Prevention of Non-response

Minimising non-contact

- timing of calls (Sunday, Monday evening best)
- number of calls (at least 7 to reduce to 4%)

Minimising refusal

- **general design strategies**
  - length of fieldwork
  - multiple modes
  - allow proxy response
- **strategies aimed at respondents**
  - advance letter/telephone call
  - incentives
  - reducing respondent burden
  - confidentiality assurance
- s**trategies aimed at interviewer**
  - interviewer characteristics/selection
  - number of interviewers
  - use of best interviewers
  - workload size
  - refusal conversion strategy

Prevention versus Compensation

Before/during data collection vs. after data collection

Prevent non-response where possible

Compensation is not an alternative to prevention

- it is difficult and must always make strong assumptions

Main aim of compensation: *to reduce bias from systematic non-response*

Methods of Compensation

Unit Non-response

Weighting

Item Non-response

Imputation

Modelling

Partial Non-response

Combination of Approaches (e.g. attrition in longitudinal surveys)

Unit vs. Item Non-response

Unit non-response

Apply a common method for all variables

Item non-response

Need to allow for different missing values on different variables

Analysis with Non response

- Treat missing values as separate category
- Complete Case Analysis – omit all cases with missing values on any variable
- Available Case Analysis - omit cases with missing values on any of the variables required for a given analysis
- Imputation – fill in missing values
- Weighting – for univariate analyses, item non-response may be combined with unit non-response
- Modelling Methods for Incomplete Data – allow for missing data in model fitting

Response Mechanisms

Response mechanisms

We can assess response mechanisms on the basis of the following information:

Y = incomplete observed outcome variable of interest (with missing values)

X_i = set of observed variables

R = response probability

Three response mechanisms

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Missing Completely At Random (MCAR)

- Response probability (R) is independent of Y and X_i
- Non-respondents form a random subsample of the complete sample

Y

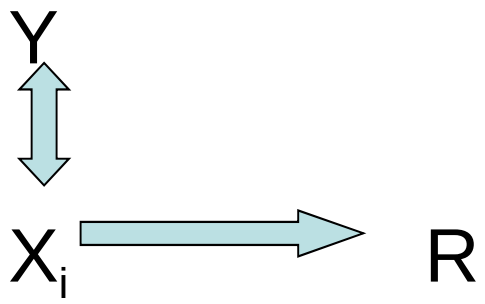
R

X_i

Missing At Random (MAR)

R dependent on X_i, but under control of X_i, no effect of Y on R

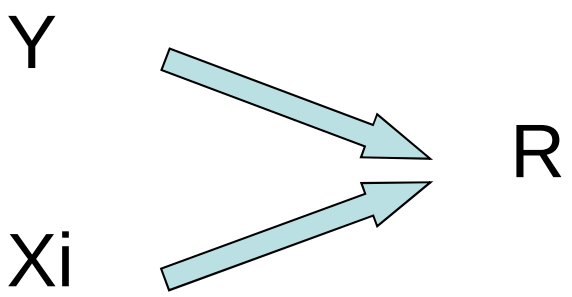
Weighting for X_i-variables or imputing on basis of X_i variables adjusts for bias in Y-variable



Not Missing At Random (NMAR)

- R dependent on X_i, as well as on Y
- No correction possible through weighting or imputation and need to move to more advanced modelling techniques
- Results tend to be biased, i.e. the value of the variable that's missing is related to the reason it's missing

  An example of this is if a certain question on a questionnaire tend to be skipped deliberately by participants with certain characteristics



Compensating for Item Non-response - Imputation

Imputation Methods

Main problem of imputation is preservation of statistical distribution of (complete, but partly unknown) data as well as possible

Methods:

- Deductive and rule-based
- Mean or mode imputation; mean or mode within class
- Hot Deck (random and nearest neighbour using distance function)
- Predictive mean matching
- Using regression models
- Maximum likelihood imputation
- Multiple imputation

Deductive and Rule-based

Examples

- Age = 9 so deduce marital status = single
- Earnings last month = E so impute annual earnings as 12E
- Impute amount of pension or welfare payment according to rules for entitlement

Example: Survey of Health and Safety Enforcement

Y1 = number of prohibition notices

Y2 = number of improvement notices

Y3 = number of formal notices

$$Y1 + Y2 = Y3$$

If one variable is missing can deduce value from other two variables.

If Y1 and Y2 are missing and Y3 = 0 can deduce Y1 = 0, Y2 = 0 since Y1, Y2, Y3 ≥ 0

Mean or Mode Imputation

Impute all missing values of y by (weighted) respondent mean $\bar{y}_r$ of y, if y continuous

Impute by respondent mode if categorical variable e.g. number of cars.

<u>Mean or Mode within Class</u>

Define homogeneous classes and impute mean or mode within class

Want y mean (or mode) same for respondents and non-respondents within classes or the probability of response for respondents and non-respondents is the same within classes

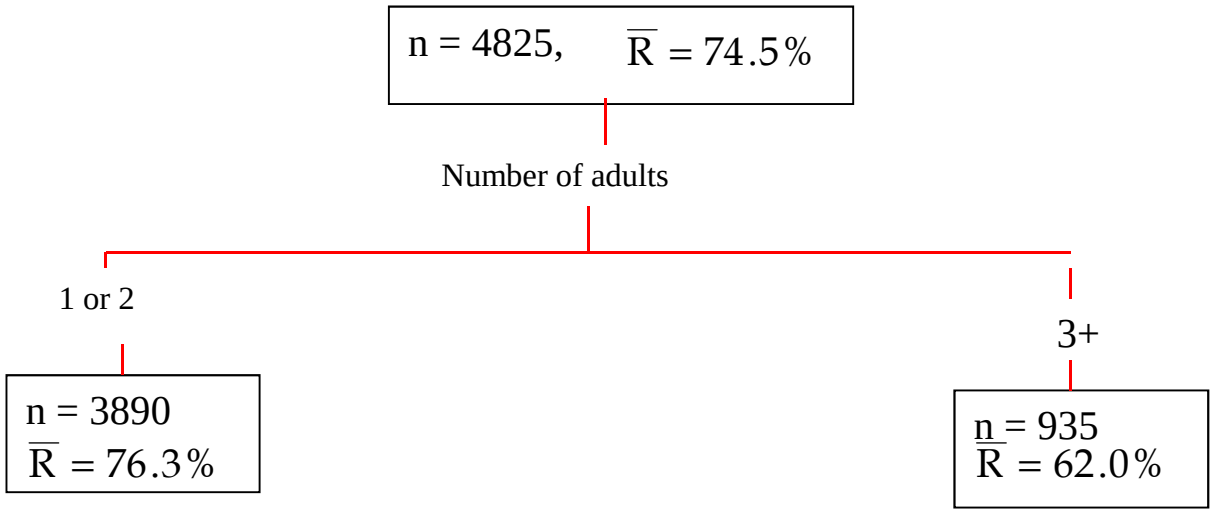
May use segmentation algorithm or models to develop homogenous classes

Segmentation Algorithm

Segmentation methods, such as CHAID, available in SPSS

The sample is split progressively into subgroups where dependent variable is response indicator and we are maximising differences in average response rate $\overline{R}_h$

Family Expenditure Survey: refusals



Random Imputation

Donors = respondents (on y) chosen at random

Missing values imputed by donor's values
Different methods of donor selection

If y categorical can impute by random number generator e.g. smoker with probability 0.3 and non-smoker with probability 0.7 (if respondent proportion of smokers = 0.3)

Hot Deck Imputation

Random Hot Deck: Form of random imputation within classes

- Records ordered on class variables within homogenous group
- Impute value from random donor in same class

Nearest Neighbour Hot Deck:

- Within classes, records are sorted according to a distance function
- Donor record is the record ‘closest’ to the recipient

Example of hot deck to impute income: produce homogeneous classes, sort records within class by hours worked and find the nearest donor to impute value

Regression Imputation

Fit regression model

$$y_i = x_i' \beta + \in_i$$

to respondents, where y is variable with missing values and x_i is vector of variables known for all cases in sample

Impute for case with missing y by predicted value

$$\hat{y}_i = x_i' \hat{\beta}$$

Mean Imputation and Regression Imputation

If there are k classes and x_i is a $k \times 1$ vector of indicator variables for these classes, then

$\hat{\beta}$ is the vector of class means of y

$x_i' \hat{\beta}$ is the mean of y within the class containing unit *i*

Mean imputation (within classes) is thus a special case of regression imputation

Both will distort distributions and lead to too little variability

Random Regression Imputation

To preserve distributions it is better to add random residual

$$\hat{y}_i = x_i^{'} \hat{\beta} + e_i$$

Different choices of e_i

- Simulate from $N\left(0, \hat{\sigma}_{\epsilon}^2\right)$ - parametric, model dependent
- Draw at random from respondent residuals – nonparametric
- Select residual for ‘similar’ respondent - nonparametric

Similarity to Within Class Hot Deck

Suppose again missing y,

x_i is the vector of indicators of k classes (there is no intercept)

$\hat{\beta}$ is a vector of class means

Then drawing randomly from residuals within the same class is equivalent to random within class hot deck imputation
so random regression imputation includes hot deck imputation as special case

If regression is non-linear than selecting residual for respondent with similar x protects against model mis-specification, e.g. from non-linearity

Regression Imputation for Constrained and Categorical Variables

If constraints, e.g. $y \geq 0$ then can transform, e.g.

$$\log y = x'\beta + \varepsilon$$

If y binary can fit logistic model

$$\log [P(y = 1) / P(y = 0)] = x'\beta$$

And impute y=1 if $\hat{P}(y = 1) > 0.5$

Predictive Mean Matching

Fit model to main y variable (e.g. income), e.g.

$$y_i = x_i'\beta + \in_i$$

and then select donor for case with missing y which is nearest on $x_i' \hat{\beta}$

Advantages:

- may be used to impute for several related variables (not all continuous);
- imputed values always obey constraints and are feasible;
- robust to model mis-specification

Disadvantage:

- may not make most efficient use of the data

Model based Multiple Imputation

- Assumes a statistical model for the data, for example a multivariate normal distribution or a non-parametric approach
- Start by replacing missing values by their means
- Fits the model and then replaces the missing values with a sample from their predictive distribution given the data
- Do this repeatedly until the pattern stabilises
- You then have a complete data set to work with
- Release multiply imputed datasets to obtain more precise variance estimates using Rubin's combining rules

For m multiply imputed datasets: $\bar{y} = \frac{1}{m}\sum_{j=1}^{m}\bar{y}_j$

and for variance:

$$Var(\bar{y}) = \frac{1}{m}\sum_{j=1}^{m} var(\bar{y}_j) + \left(1+\frac{1}{m}\right)\frac{1}{m-1}\sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2$$

Multivariate Imputation

- Preserve covariances and correlations
  - Domain means e.g. mean income of age group 20-29
  - Proportions in 2-way tables e.g. proportion who have no car and no job

Tendency for bias if each variable imputed independently (attenuation of correlations and differences between domain means)

For multivariate parameters involving both X and Y, with missing data on either X or Y, to avoid bias:

- Impute missing (X,Y) from common donor or joint predictive distribution
- Sequential regression (often used in Multiple Imputation): start with a 'complete' dataset; impute missing X dependent on Y and other regressors; impute missing Y dependent on X and other regressors, continue....

Properties of Imputation Methods

Bias Properties – Means

$\bar{y}^*$ = sample mean with imputed values

$= [\sum_r y_i + \sum_{nr} y_i^*] / n$ where y_i^* = imputed value

Under what conditions is expectation of $\bar{y}^*$ equal to $\bar{Y}$?

Assumptions:

- average imputed values same as average values for respondents, within imputation classes or given x variables used. This is reasonable for several imputation methods (eg. hot deck, regression)
- average value of y within imputation classes (or given x) is same for respondents and non-respondents. This depends on non-response mechanism

Bias Properties – Variances and Distributions

We have already seen simple regression imputation leads to downward bias in estimation of variances

Regression imputation also causes bias in estimation of quartiles, e.g. lower quartile overestimated, upper quartile underestimated

Deterministic imputation methods (e.g. simple regression, mean imputation) generally lead to bias in estimation of variances

Stochastic methods (e.g. hot deck, random regression) can be designed to reduce this bias

$$\text{var}(\bar{y}^*) = \text{var}(\bar{y}_r)[1+\bar{R}(1-\bar{R})]$$

Thus variance inflated: e.g. $\bar{R} = 0.8$ ➔ variance increase = 16%

Compensating for Unit Non-response - Weighting

Introduction

Some parts of population are underrepresented in respondent data
Weight these parts up to compensate for under representation
Example (Source: Moser and Kalton (1971) Ch.7)

Sex	Respondents		Non-respondents		Sample		Pop.
	No	%	No	%	No	%	%
Men	1360	43.6	280	58.0	1640	45.6	47.0
Women	1757	56.4	203	42.0	1960	54.4	53.0
Total	3117	100.0	483	100.0	3600	100.0	100.0

From the table:
Response rate for men is 82.9% (1360/1640)
Response rate for women is 89.6% (1757/1960)

Introduction

Example cont…

Read Daily Tabloid		
Men	1088	1088/1360 = 80%
Women	176	176/1757 = 10%
All respondents	1264	1264/3117 = 40.6%

Unweighted estimate

0.436 x 80 + 0.564 x 10 = 40.6%

Sample-based weighted estimate

0.456 x 80 + 0.544 x 10 = 41.9%

re-weight the responses based on the selected sample proportions

Population-based weighted estimate

0.47 x 80 + 0.53 x 10 = 42.9%

re-weight the responses based on the '**known**' population proportions

Population based Weighting

Uses auxiliary population information…

to reduce sampling error

to adjust for unit non-response

to adjust for non-coverage

to calibrate to external estimates

Weighted Estimates of Totals

e.g. population estimates of

number of men= 23,500,000

number of women= 26,500,000

(Population based) weighted estimate of number who read Daily Tabloid

= $23{,}500{,}000 \times 0.8 + 26{,}500{,}000 \times 0.1$

= 21,450,000

Weighting to Adjust for Unequal Inclusion Probabilities

Weighting at Stratum level

Stratum *h* has population size N_h and sample size n_h (h = 1, ..., H)

- stratum selection probability n_h/N_h

Assume disproportionate stratification, (i.e. n_h/N_h unequal)

Weighted mean $$\frac{\sum_h N_h \; y_h \; / \; n_h}{\sum_h N_h}$$

where y_h is sample total of y in stratum h

- So weight stratum sample means by N_h
- Or weight stratum sample totals by (N_h/n_h)

Example (Weighting at Stratum Level)

General Household Survey 1988

Men aged 16+

	Sample	High Alcohol Consumption	
Stratum, h	n_h	y_h	% (y_h/n_h)
England	7391	1944	26
Wales	482	135	28
Scotland	800	176	22

- Sample fraction n_h/N_h in Scotland double
- Each man in England &Wales 'represents' 2 men

$$\frac{\sum (N_h/n_h) y_h}{\sum (N_h/n_h) n_h} = \frac{2\times 1944 + 2\times 135 + 176}{2\times 7391 + 2\times 481 + 800} = 26.2\%$$

Weighting at Unit Level

π_i = probability of selection of unit i

$d_i = 1/\pi_i$ = *sampling weight (or design weight)*

Unit i in sample '*represents*' d_i units in population

e.g. If $\pi_i = 1 / 100$ then $d_i=100$

- stratum contains 500 people
- sample contains 5 people

Each sample person represents 100 people in the population…

Horvitz-Thompson Estimator

Population Total

$$Y = \sum_{U} y_i$$

(where U is the population of units)

Horvitz-Thompson Estimator (weighted estimator) of the Total

$$\hat{Y}_{HT} = \sum_{s} d_i y_i$$

(where s is the sample units)

$d_i = 1/\pi_i$ (sampling or design weight)

Example Smoking Survey

- 2 primary sampling units (PSUs) selected by simple random sampling (SRS) from stratum containing 20 PSUs.
- 5 households selected from each selected PSU by SRS.
- 1 adult selected at random from each selected household.

$$\pi_i = (2/20)\ (5/N_i)\,(1/M_i) = 1/(2N_iM_i)$$

where M_i is number of adults in the household containing adult i and N_i is number of households in PSU containing adult i.

Survey variables:
Smoking status (y_{1i} = 1 if i smokes)
Number of cigarettes smoked per week (y_{2i})

Example Smoking Survey

i	PSU	N_i	M_i	π_i	y_{1i}	y_{2i}	$d_i = \pi_i^{-1}$	$d_i y_{1i}$	$d_i y_{2i}$
1	1	200	2	1/800	0	0	800	0	0
2	1	200	3	1/1200	1	40	1200	1200	48000
3	1	200	1	1/400	1	50	400	400	20000
4	1	200	4	1/1600	0	0	1600	0	0
5	1	200	2	1/800	1	20	800	800	16000
6	2	250	2	1/1000	1	60	1000	1000	60000
7	2	250	3	1/1500	0	0	1500	0	0
8	2	250	1	1/500	0	0	500	0	0
9	2	250	4	1/2000	1	30	2000	2000	60000
10	2	250	3	1/1500	1	60	1500	1500	90000
							11300	6900	294000

Example Smoking Survey

Estimated percentage smoking = 6900/11300 = 61%

Estimated mean number of cigarettes smoked by smokers = 294000/6900= 42.6

Sample Based Weighting to Adjust for Unit Nonresponse

Two-phase approach $r \subset s \subset U$

Population (U) → Sample (s) → Respondents (r)

View response as second phase of sampling

Let $\pi_{si} = \Pr(\text{unit i selected in sample s})$

$\pi_{r|si} = \Pr(\text{unit i responds} \mid \text{i selected in sample})$

Then $\pi_i = \pi_{r|si}\pi_{si} = \Pr(\text{unit i is selected and responds})$

For well defined sampling schemes π_{si} is known

Need to estimate $\pi_{r|si}$

Then apply standard sampling theory to weight for the non-response

- this is two-phase approach to weighting

Estimation of Response Probability

Need to estimate *response probability* $\pi_{r|si}$

Usually assume $\pi_{r|si}$ depends only on i, not on s (*the sample selection*).

Write $\pi_{r|si} = \theta_i$ where θ_i = *response probability* of unit i,

- Often also assume model where θ_i is fixed within subgroups or *weighting classes (adjustment cells)*

Estimate θ_i as population proportion who respond in a subgroup.

$\hat{\theta}_i$ = response rate within subgroup

$$= \frac{\text{no. of interviews completed with eligible elements in subgroup}}{\text{no. of eligible sample elements in subgroup}}$$

Can weight estimate by sampling weights d_i

Weighted Estimation

$$\hat{Y} = \sum_r y_i / \pi_i = \sum_r y_i / \left(\hat{\theta}_i \pi_{si}\right)$$ where r=response set

$$\hat{Y} = \sum_r w_i y_i = \sum_r d_i v_i y_i$$

where $v_i = 1/\hat{\theta}_i$ *non-response weight*

$d_i = 1/\pi_{si}$ *sampling (design) weight*

$w_i = v_i d_i$ *combined weight*

Example: Smoking Survey cont..

Adults i = 2, 4 and 8 do not respond ($R_i = 0$)

Other adults respond ($R_i = 1$)

Two choices of weighting classes:

A: subgroup = whole stratum

B: subgroups = PSUs

Weight response rates by inverse household selection probabilities π_{hi}^{-1} (assuming non-response occurs at household level)

$\pi_{hi} = (2/20)(5/N_i)$, $\pi_{hi}^{-1} = 2N_i$

i	PSU	Respond(R_i)	π_{hi}^{-1}	$\pi_{hi}^{-1} R_i$
1	1	1	400	400
2	1	0	400	0
3	1	1	400	400
4	1	0	400	0
5	1	1	400	400
			2000	1200

Example: Smoking Survey cont..

i	PSU	Respond(R_i)	π_{hi}^{-1}	$\pi_{hi}^{-1} R_i$
6	2	1	500	500
7	2	1	500	500
8	2	0	500	0
9	2	1	500	500
10	2	1	500	500
			2500	2000
			4500	3200

Weighted Response Rates

$$\hat{\theta}_i^{(A)} = 3200/4500 = 71.1\% \ (i = 1,...,10) \text{ so } v_i^{(A)} = 4500/3200 = 1.406$$

$$\hat{\theta}_i^{(B)} = \begin{cases} 1200/2000 = 60\% & (i = 1,...,5) \\ 2000/2500 = 80\% & (i = 6,...,10) \end{cases} \text{ so } v_i^{(B)} = \begin{cases} 2000/1200 = 1.667 \\ 2500/2000 = 1.25 \end{cases}$$

Results of Alternative Weighting Adjustments

i	R_i	d_i	$v_i^{(A)}$	$v_i^{(B)}$	$w_i^{(A)}$	$w_i^{(B)}$	$w_i^{(A)} y_{1i}$	$w_i^{(B)} y_{1i}$	$w_i^{(A)} y_{2i}$	$w_i^{(B)} y_{2i}$
1	1	800	1.406	1.667	1125	1333	0	0	0	0
2	0	Non-response								
3	1	400	1.406	1.667	563	667	563	667	28150	33350
4	0	Non-response								
5	1	800	1.406	1.667	1125	1333	1125	1333	22500	26660
6	1	1000	1.406	1.25	1406	1250	1406	1250	84360	75000
7	1	1500	1.406	1.25	2109	1875	0	0	0	0
8	0	Non-response								
9	1	2000	1.406	1.25	2813	2500	2813	2500	84390	75000
10	1	1500	1.406	1.25	2109	1875	2109	1875	126540	112500
					11250	10833	8016	7625	345940	322510

Results of Alternative Weighting Adjustments

Estimated percentage smoking

A: 8016/11250 = 71.3%

B: 7625/10833 = 70.4%

without non-response the estimate was 61%

Estimated mean number of cigarettes smoked by smokers

A: 345940/8016 = 43.2

B: 322510/7625 = 42.3

without non-response the estimate was 42.6

Properties of Estimators

Fixed Population Model of Non-Response:

U = finite population of N units

R_i = 1 if unit i does/would respond (where $i \in U$)

= 0 if not

R_i are fixed not random

$U_1 = \{i \in U : R_i = 1\}$ responding subpopulation

$U_0 = \{i \in U : R_i = 0\}$ non-responding subpopulation

N_1 = size of U_1 , N_0 = size of U_0

$N_0 + N_1 = N$

Bias of Unweighted Estimator

Assume an SRS (s from U)

$r = s \cap U_1$ respondents, n_r = size of r

$\bar{y}_r$ = unweighted estimator of population mean $\bar{Y}$

$= \sum_r y_i / n_r$ = domain mean (domain = U_1)

Let us assume that $E(\bar{y}_r) = \sum_{U_1} y_i / N_1 = \bar{Y}_{R=1}$

$$Bias(\bar{y}_r) = \bar{Y}_{R=1} - \bar{Y}$$

$$= \bar{Y}_{R=1} - \left(N_1 \bar{Y}_{R=1} + N_0 \bar{Y}_{R=0} \right) / N$$

$$= \left(1 - \bar{R}\right)\left(\bar{Y}_{R=1} - \bar{Y}_{R=0} \right) \qquad (\bar{R} = N_1/N)$$

No bias if either $\bar{R} = 1$ (*no non-response*) or

$\bar{Y}_{R=1} = \bar{Y}_{R=0}$ (*non-response unrelated to* y_i)

Choice of Weighting Classes

Can only use information available on both respondents and non-respondents.

- Strata, PSUs
- Other information on sampling frame
- Simple question posed to non-respondents e.g. household composition
- Previous phase of survey (or wave in longitudinal survey)
- Data collected by interviewer e.g. type of dwelling

Want to choose classes that:

- $\overline{Y}_{h,R=1} \doteq \overline{Y}_{h,R=0}$ need external information to check
- $\overline{R}_h$ vary – otherwise weighting has no effect (modelling methods can be used – see later)
- sample sizes in classes not 'too small' (say <25, 30 or 50)

Post-stratification

Recall sample weighted: $\bar{y}_{sw} = \sum_h \hat{N}_h \hat{\bar{Y}}_{h,R=1} / \sum_h \hat{N}_h$

If N_h, the population number in weighting class h, is known then we may use the population weighted estimator

$$\bar{\mathbf{y}}_{\mathbf{pw}} = \sum_{\mathbf{h}} \mathbf{N_h} \hat{\bar{\mathbf{Y}}}_{\mathbf{h,R=1}} / \sum_{\mathbf{h}} \mathbf{N_h} = \sum_{\mathbf{h}} \mathbf{W_h} \hat{\bar{\mathbf{Y}}}_{\mathbf{h,R=1}}$$

where $W_h = N_h/N$

Analogous to estimator for stratified sampling, except that weighting class is not usually a stratum used for sampling.

Thus $\bar{y}_{pw}$

is also called a post-stratified estimator

Example: Smoking Survey

National estimates of proportions W_h in four age groups used to estimate proportion who smoke.

Age Group h	Population Proportion W_h	Respondents Proportion	% smoking estimated from respondents
16 – 34	0.3	0.2	40
35 – 49	0.2	0.25	30
50 – 64	0.2	0.25	30
65+	0.3	0.3	25
	1.0	1.0	

Unadjusted estimate

$= 0.2 \times 40 + 0.25 \times 30 + 0.25 \times 30 + 0.3 \times 25 = 30.5$

Post-stratified estimate

$= 0.3 \times 40 + 0.2 \times 30 + 0.2 \times 30 + 0.3 \times 25 = 31.5$

Example of Weighting Classes

Demographic estimates of numbers of persons in age × sex × marital status groups × region

Number of households of certain composition
Use register / administrative / sampling frame information:

- IDBR auxiliary information on turnover and number of employees
- information from census and/or population estimates

Can also use *high quality estimates* from other surveys:

- LFS estimates on economic status as totals for other surveys

Using Models to Construct Weights

Want to control for factors which are most strongly related to R_i. Other factors may not lead to much bias

Fit regression model, Dependent variable = R_i
Explanatory variables = variables which can be used for sample (or possibly population) weighting

Logistic Weighting

Fit model $\log\left[\theta_i/(1-\theta_i)\right] = x_i'\beta$

Then $\hat{\theta}_i = \exp\left[x_i'\hat{\beta}\right]/\left[1+\exp\left(x_i'\hat{\beta}\right)\right]$

Could use $\hat{\theta}_i^{-1}$ as response weights or can define weighting classes with similar values of $\hat{\theta}_i$

Segmentation Algorithm (see slide 25)

Calibration Methods

Horvitz-Thompson estimator

$\hat{Y}_{HT} = \sum_{i \in s} d_i y_i$ where $d_i = \pi_i^{-1}$ is design weight

Generalized Regression (GREG) Estimation

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + \left(\sum_U x_i - \sum_s d_i x_i \right)' B$$

where x_i is $J \times 1$ vector of auxiliary variables for which $\sum_U x_i$ is known

$$B = \left(\sum_s d_i x_i x_i' \right)^{-1} \left(\sum_s d_i x_i y_i \right)$$

Can be written as: $\hat{Y}_{GREG} = \sum_s w_i y_i$ where $w_i = d_i g_i$ and

$$g_i = 1 + \left(\sum_U x_i - \sum_s d_i x_i \right)' \left(\sum_s d_i x_i x_i' \right)^{-1} x_i$$

Note that: $\sum_s w_i x_i = \sum_U x_i$

Post Stratification Revisited

Auxiliary information:

N_h = size of post-stratum h ($=1,\ldots,H$)

Write $x_i = (x_{1i}, x_{2i}, ..., x_{Hi})'$ where

$x_{hi} = 1$ if $i \in$ post-stratum h, and 0 otherwise

Let n_h be number of respondents in post-stratum h

The post-stratified weights are:

$$w_i = N_h / n_h \quad \text{if } i \in \text{ post-stratum } h$$

Post-stratified weights obey:

$$\sum_r w_i x_i = \sum_U x_i$$

Can write these H calibration constraints as:

$$\sum_r w_i x_{hi} = N_h \quad \text{for } h=1,\ldots, H$$

Ratio Estimation

y_i survey variable known for respondents

x_i auxiliary variable, x_i for respondents and $X = \sum_U x_i$ are known

e.g. business survey, y = production,
x = employment, the ratio estimator of total Y is: $Y_{rat} = \frac{\bar{y}_r}{\bar{x}_r} X$

Good estimator if $y_i \; \alpha \; x_i$ approximately, i.e. when non-response is differential by x

Can write: $Y_{rat} = \sum_r w_i y_i$ where $w_i = \frac{X}{n_r \bar{x}_r}$ and n_r number of respondents

Hence $\sum_r w_i x_i = \frac{X}{n_r \bar{x}_r} n_r \bar{x}_r = X$
calibration to X

x_i	y_i	
50	20	
60	30	respondents
40	25	
70	40	non-respondents
80	50	
X = 300	Y = 165	

$\bar{y}$ = 5 x 25 = 125 biased downwards

$\bar{Y}_{rat} = \frac{20+30+25}{50+60+40} 300 = \frac{25}{50} 300 = 150$

Effect of Weighting on Variance

1. Unequal (fixed) weights increase variance.
2. Post-stratification (and calibration methods) reduce variance

$$\text{var}\left(\hat{Y}_{pw}\right) = \sum_h N_h{}^2 S^2_{hR=1} \,/\, r_h$$

depending on the extent to which within-stratum variance

$S^2_{hR=1}$ is smaller than overall variance

Increase of variance due to weighting is: $= 1 + c_w^2$

where $c_w = s_w \,/\, \overline{w}$ = coefficient of variation of w_i.

Effect of Weighting on Variance

So, to avoid inflating variance, choose weighting classes so that

- c_w^2 not too large

- samples sizes in classes not too small, e.g. if sample size <25 or non-response weight > 2 then collapse weighting classes

Statistical Data Editing

Statistical Data Editing

- Statistical data editing is the process of checking collected data and correcting them if necessary
- If a violated edit restriction involves several variables, it is not immediately clear which of the variables (if any) are in error
- Statistical data editing can be subdivided into three steps:
  - Finding erroneous records and erroneous fields in those records (Error Localization)
  - Replacing erroneous and missing fields by best possible way (Imputation)
  - Adjust imputed values such that all edits become satisfied (Consistency)

SDE and the survey process

- We focus on identifying and correcting errors
- Other goals of SDE are
  - identify error sources in order to provide feedback on the entire survey process;
  - provide information about the quality of the incoming and outgoing data;
- Role of SDE is slowly shifting towards these goals
  - feedback on other survey phases can be used to improve these phases and reduce amount of errors arising in these phases

Edit Restrictions

Edit restrictions

- Edit restrictions, or edits for short, often used to determine whether a record is consistent or not
- Edit restrictions capture subject-matter knowledge of admissible (or plausible) values and combinations of values in each record
- Inconsistency of data values with edit restrictions means that there is an error or in any case that the values are implausible
- Consistent records that are also not suspicious otherwise, e.g. are not outlying with respect to the bulk of the data, are considered error-free

Edit Restrictions

- Examples
  - A male cannot be pregnant
  - A female cannot have given birth to more than 20 children
  - Balance Edit: $a_1 x_1 + ... + a_n x_n = b$ eg., Profit of enterprise should be equal to its total turnover minus its total costs

    $T - C - P = 0$
  - Inequality Edit: $a_1 x_1 + ... + a_n x_n \geq b$, eg. Profit of enterprise should be less than 50% of its total turnover

    $0.5 \times T - P \geq 0$

Fellegi & Holt (1976) Algorithm

Principles:

1. The record has to pass all edit restrictions with minimum changes to fields
2. Automatic localization of errors
3. Marginal and joint distributions should not change as a result of the correction

Fellegi & Holt (1976) Algorithm

Error Localization: Consider explicit edits defined for age, marital status and relationship to head of hh:

E1={age<15, married, . }

E2={. , married, spouse}

Generate implicit edit that can be logically derived from first 2 edits: E3={age<15, . , spouse}, i.e. if E3 fails then necessarily edit E1 or E2 fails

Consider record r={age<15,married, spouse} and the record fails E1 and E3

If we change marital status then r fails E2

Implicit edits contain information about edits that do not fail but may fail after fields are changed

Must change at least one field in common with E1 and implicit edit E3

Edit Restrictions

For continuous variables, assume the following edits:

$$100 \geq y$$

$$y \geq x$$

$$x \geq 50$$

Suppose we have a record in which x and y are missing (or were deleted in the error localization)

If we first impute a variable and neglect edit restrictions:

- Assume impute x=150 or y=40
- Can’t satisfy all edit restrictions

Edit Restrictions

- We take variables with missing values into account by eliminating them (Fourier-Motzkin elimination)
- Elimination leads to implicit edit restrictions not involving these variables
- Implicit edit restrictions contain relevant information on remaining variables

By taking these implicit edit restrictions into account, it is **guaranteed** that we can find allowed values for all variables to be imputed

Edit Restrictions

Recall edits: $100 \geq y$

$y \geq x$

$x \geq 50$

If we eliminate y we obtain the implicit edit $100 \geq x$

And the new edit restrictions are: $100 \geq x$

$x \geq 50$

So to impute x, we take a value between 50 and 100

This will guarantee a value for y that will satisfy all edit restrictions

Recapitulating

- We impute variables one by one
- For each variable we impute records one by one
  - For each record we fill in observed and already imputed values in edit restrictions
  - For each record we eliminate variables still to be imputed
  - This leads to an admissible interval for the current variable to be imputed

Statistical Data Editing Methods

Statistical Data Editing

- Several techniques are available:
  - Interactive editing
  - Selective editing
  - Automatic editing
  - Macro-editing

Interactive editing

- When data are edited interactively in a modern survey processing system (eg., Blaise), effects of adjusting data in terms of failed edit restrictions or distributional aspects can be seen immediately
- This immediate feedback, in combination with data themselves, direct subject-matter specialist to potential errors
- Interactive editing
  - Requires subject-matter knowledge
  - Edit restrictions to guide interactive editing process
  - In basic form all records have to be edited which is costly but generally considered high quality

Selective editing

- Selective editing aims to identify records with potentially influential errors
- Most common form of selective editing up to now is based on score functions that are used to split data into two streams
  - critical stream: records that are the ones most likely to contain influential errors
  - noncritical stream: records that are unlikely to contain influential errors
- Records in critical stream are edited interactively
- Records in non-critical stream are either not edited or are edited automatically

Selective editing

- Score for a record is often a weighted sum of scores for each of a number of important target parameters (local scores)
- Local scores are often defined as product of two components
  - likelihood of potential error ("risk"): measured by comparing raw value with "anticipated" value
  - contribution on estimated target parameter ("influence"): measured as (relative) contribution of anticipated value to estimated total

  Example: $\sum d_i \, |Y_i - Y^*|$ with d_i design weight, Y_i observed value, Y_i* anticipated value
- Records with scores above certain threshold are directed to interactive editing
- Edit restrictions are generally not used in the construction of score functions

Selective editing

- Advantage:
  - Selective editing improves efficiency in terms of budget and time
- Disadvantage:
  - No ‘best’ technique for combining local scores into a global score have been identified if there are many variables
- Selective editing has gradually become a popular method to edit business (numerical) data
- Anticipated value can be based on a response from the business at a previous wave according to a prediction model

Automatic editing

- Two kinds of errors: systematic and random
- Systematic error: error reported consistently among (some) responding units
  - gross values reported instead of net values
  - values reported in units instead of requested thousands of units (so-called “thousand-errors”)
  - Incorrect minus sign, interchanged digits
  - Rounding errors
- Random error: error caused ‘by accident’
  - observed value where respondent mistakenly typed in the wrong value

Correcting errors

- Systematic errors often easy to correct once detected
  - Detection and correction based on same method or rule
- Random errors often hard to correct after detection
  - Detection and correction based on separate methods
  - Detection used to set suspicious values to missing
  - Correction based on imputation methods for missing values
- Edit restrictions play a fundamental role in these techniques

Automatic editing of random errors

- Random errors occur by accident, not by systematic reason
- Methods can be subdivided into three classes:
  - methods based on statistical models and outlier detection
  - methods based on deterministic checking rules: “if components do not sum up to total, total is erroneous”
  - methods based on solving a mathematical optimization problem
    - Use paradigm of Fellegi and Holt: data in each record should be made to satisfy all edit restrictions by changing fewest possible number of fields
    - Edit restrictions play a fundamental role in latter approach

Error localization as mathematical optimization problem

- Guiding principle is needed
  - Freund and Hartley (1967): minimize sum of distance between observed data and “corrected” data and a measure for the violation of edits
  - Casado Valera et al. (1996): minimize quadratic function measuring distance between the observed data and “corrected” data such that “corrected” data satisfy all edits
  - Bankier (1995): impute missing data and potentially erroneous values by means of donor imputation, and select imputed record that satisfies all edits and that is “closest” to original record

Fellegi-Holt paradigm: (dis)advantages

- Advantages:
  - Drastically improves efficiency in terms of budget and time
  - In comparison to deterministic checking rules less detailed, rules have to be specified
- Disadvantages:
  - Class of errors that can safely be treated is limited to random errors
  - Class of edits that can be handled is restricted to so-called hard (or logical) edits which hold true for all correctly observed records
  - Risky to treat influential errors by means of automatic editing

Macro-editing

- Macro-editing checks whether data set as a whole is plausible and examine potential impact on survey estimates to identify suspicious data in individual records
- We distinguish between two forms of macro-editing:
  - Aggregation method
  - Distribution method

- Edit restrictions hardly play a role in macro-editing

Macro-editing: aggregation method

- Verification whether figures to be published seem plausible
- Compare quantities in publication tables with
  - same quantities in previous publications
  - quantities based on register data
  - related quantities from other sources

Macro-editing: distribution method

- Available data used to characterize distribution of variables
- Individual values compared with this distribution
- Records containing values that are considered uncommon given the distribution are candidates for further inspection and possibly for editing

Macro-editing: graphical techniques

- Exploratory Data Analysis techniques can be applied
  - box plots
  - scatter plots
  - (outlier robust) fitting
- Other often used techniques in software applications
  - anomaly plots
  - time series analysis
  - outlier detection methods
- Anomaly plots are graphical overviews of important estimates, where unusual estimates are highlighted
- Once suspicious data have been detected on a macro-level one can drill-down to sub-populations and individual units

Macro-editing: (dis)advantages

- Advantages:
  - Directly related to publication figures or distribution
  - Efficient in term of budget and time
- Disadvantages:
  - Records that are considered non-suspicious may still contain influential errors
  - Publication of unexpected (but true) changes in trend may be prevented
  - For data sets with many important variables graphical macro-editing is not most suitable SDE method
    - Most persons cannot interpret 10 scatter plots at the same time

Integrating SDE Techniques

- Use an SDE approach that consists of following phases:
  - Correction of “evident” errors
  - Application of selective editing to split records in critical stream and non-critical stream
  - Editing of data:
    - records in critical stream edited interactively
    - records in non-critical stream edited automatically
  - Validation of the publication figures by means of (graphical) macro-editing

Integrating SDE techniques

- All editing and imputation methods have their own (dis)advantages
- Integrated use of editing techniques (selective editing, interactive editing, automatic editing, and macro-editing) as well as various imputation techniques can improve efficiency of SDE and imputation process while at same time maintaining or even enhancing statistical quality of produced data

Thank you for your attention

Questions?