Lecture 3: Nonresponse issues and imputation

- Sources of nonresponse
- Real life examples
- Effects of nonresponse –bias
- Estimation methods for reducing the effect of nonresponse
  - Weighting, for unit nonresponse only (very short)
  - Imputation methods, for item nonresponse mainly
  - Nonignorable response mechanisms

- Nonresponse occurs in almost all surveys, even “compulsory” ones
  - Labor Force survey in Norway, quarterly, 10% nonresponse
- Perceived to have increased in recent years
- Besides sampling error, the most important source of error in sampling
- Nonresponse is important to consider because of
  - (Potential) **bias** (will almost always result in bias), sample is not representative of the population
  - Increased uncertainty in the estimates

- Nonresponse is the failure to obtain complete observations on the survey sample
- Unit nonresponse: unit (person or household) in the sample does not respond
  - Can be very high proportion, can be as much as 70% in postal surveys
  - 30% is not uncommon in telephone surveys
  - 50% in the Norwegian Consumer Expenditure survey, up from about 30% 15 years ago
- Item nonresponse: observations on some items are missing for unit in sample
- Remedies: *Weighting* for unit nonresponse, *imputation* for item nonresponse

Sources of unit nonresponse

- Non-contact: failure to locate/identify sample unit or to contact sample unit
- Refusal: sample unit refuses to participate
- Inability to respond: sample unit unable to participate, e.g. due to ill health, language problems
- Other: e.g. accidental loss of data/questionnaire

Sources of item nonresponse

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

Amounts

- some variables only 1-2%
- Often highest for financial variables, e.g. total household income may have 20% missing data

- The response rate is the most widely reported quality indicator
- But need not be related to how large the bias may be
- 2 examples to illustrate how nonresponse can lead to very misleading statistical analysis, even when the response rate is high

1. Classical example, response rates 81-85%

- Political polling before the American presidential election in 1948
  - Democratic candidate: Truman
  - Republican candidate: Dewey
  - Institute: Roper
  - Surveys : July, August, September, October
  - Election: November

	July	August	Sept	Oct	Election
Truman	37.8	37.0	35.2	40.4	49
Dewey	55.5	52.4	57.0	53.4	45
Others	6.7	10.5	7.7	6.2	6
Sample size	3011	3490	3490	3500	
responses	2510	2951	2936	2841	
Nonresponse (Percentage)	501 (18.6)	539 (15.4)	554 (15.9)	659 (18.8)	

- Bias: Larger nonresponse rate among the economically poorer groups
- Compensating for nonresponse: Model the probability of response dependent on which candidate the person will vote for, within in each socio-economic group

  Nonignorable response mechanism

- Gives Truman 51%
- Method: Imputation, estimate 93-99% will vote for Truman in the nonresponse group

2. Election survey in Norway 1993

- Sample: 3000 persons
- Number of responses, after two callbacks: 1403
- Estimate the voting proportion
- Of the 1403, 1190 said they voted in Parliament election: 84.8%
- Margin of error: 2%
- True voting proportion = 75.5%
- Estimate 84.8% is biased because higher nonresponse rate among nonvoters
- The response sample is not representative of the nonresponse group (typically the case)

$$\text{Margin of error}: 2 \cdot SE = 2\sqrt{\frac{0.848 \cdot 0.152}{1403}} = 2 \cdot 0.0095 = 0.019$$

Effect of nonresponse

Fixed population model of nonresponse:

U = finite population of N units

$$r_i = \begin{cases} 1 \text{ if unit } i \text{ does/would respond} \\ 0 \text{ if not} \end{cases}$$

$$i = 1, \ldots, N$$

r_i's are fixed, not random

$$U_R = \{i \in U : r_i = 1\} = \text{responding subpopulation}$$

$$U_M = \{i \in U : r_i = 0\} = \text{nonresponding subpopulation}$$

$$N_R = \text{size of } U_R$$

$$N_M = \text{size of } U_M$$

Bias of standard estimator

Simple random sample of size n

Response sample: $s_r = s \cap U_r$, size n_r

Estimate the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Population means of U_R and U_M : $\bar{Y}_R$ and $\bar{Y}_M$

$$\bar{Y} = \frac{N_R \bar{Y}_R + N_M \bar{Y}_M}{N} = q_R \bar{Y}_R + (1 - q_R)\bar{Y}_M$$

$q_R = N_R / N =$ expected response rate

Standard estimator :

observed sample mean : $\bar{y}_r = \frac{1}{n_r} \sum_{i \in s_r} y_i$

Given n_r : the response sample s_r is a random sample from U_R

$$\Rightarrow E(\bar{y}_r) = \bar{Y}_r$$

$$\Rightarrow \text{Bias} = E(\bar{y}_r) - \bar{Y} = \bar{Y}_R - \bar{Y}$$

$$= \bar{Y}_R - q_R \bar{Y}_R - (1 - q_R)\bar{Y}_M = (1 - q_R)(\bar{Y}_R - \bar{Y}_M)$$

No bias if either $q_R = 1$ or $\bar{Y}_R = \bar{Y}_M$

Mean square error:

$$E(\bar{y}_r - \bar{Y})^2 = Var(\bar{y}_r) + [E(\bar{y}_r) - \bar{Y}]^2$$
$$= EVar(\bar{y}_r \mid n_r) + [E(\bar{y}_r) - \bar{Y}]^2$$
$$\approx \left(1 - \frac{q_R n}{N_R}\right)\frac{\sigma_R^2}{q_R n} + (1 - q_R)^2 (\bar{Y}_R - \bar{Y}_M)^2$$

We notice that even if there is no bias, the uncertainty increases because of smaller sample size

Expected sample size decreases from n to $q_R n$

For example, if we want a sample of 1000 units and we know q_R: $n = 1000/q_R$

If expected response rate is 60% : need n= 1000/0.60= 1667

$$\text{Bias} = E(\bar{y}_r) - \bar{Y} = (1 - q_R)(\bar{Y}_R - \bar{Y}_M)$$

Possible consequenses of nonresponse:

1. Bias is independent of n, can not be reduced by increasing n
2. Bias increases with increasing nonresponse rate (1-q_R)
3. Bias increases when $|\bar{Y}_R - \bar{Y}_M|$ increases
4. If $\bar{Y}_R = \bar{Y}_M$: *ignorable* nonresponse mechanism

Unrealistic to assume $\overline{Y}_R = \overline{Y}_M$,

But within smaller subpopulations it may not be unreasonable,

especially if the variable used to partition the population is highly correlated with y

Called: poststratification

Widely used tool to correct for nonresponse

Estimation methods for reducing the effect of nonresponse

- Handling nonresponse:
  - Reduce the *size* of nonresponse, especially by callbacks
  - Reduce the *effect* of nonresponse, by estimating the bias and correcting the original estimator designed for a full sample
- Estimation methods:
  - Weighting, for unit nonresponse
  - Imputation, especially for item nonresponse

Weighting

Basic idea:

- Some parts of the population are underrepresented in the response sample
- Weigh these parts up to compensate for underrepresentation
- Population-based
  - Reduces sampling error
  - Adjusts for unit nonresponse

Poststratification

1. Stratify using variables that partitions the population in homogeneous groups
2. Stratify according to varying response rates

H poststrata. For poststratum h, U_{Rh} is the responding substratum and U_{Mh} is the nonresponding substratum

$$q_h = \text{response rate in postratum } h$$

$$W_h = N_h / N, \text{ where } N_h \text{ is the size of poststratum } h$$

$$\bar{Y}_{Rh} = \text{mean in responsestratum } h$$

$$\bar{Y}_{Mh} = \text{mean in nonresponsestratum } h$$

Simple random sample and $\bar{y}_r$ estimating $\bar{Y}$

$$E(\bar{y}_r) - \bar{Y}$$

$$= \frac{1}{q_R} \sum_{h=1}^{H} \bar{Y}_{Rh} W_h (q_h - q_R) + \sum_{h=1}^{H} (1 - q_h) W_h (\bar{Y}_{Rh} - \bar{Y}_{Mh})$$

1. component: Bias because of different response rates in the poststrata, can be estimated

2. Component can not be estimated if response and nonresponse means are different

Poststratification estimates the first component

Choose poststrata such that most of the bias is in the first component

$\Rightarrow q_h$ should vary as much as possible, and $\bar{Y}_{Rh} \approx \bar{Y}_{Mh}$

First component: $\bar{Y}_R - \sum_{h=1}^{H} W_h \bar{Y}_{Rh}$

Observed mean from poststratum h: $\bar{y}_h$

$\Rightarrow$ unbiased estimator for this component: $\bar{y}_r - \sum_{h=1}^{H} W_h \bar{y}_h$

$\Rightarrow$ adjusted estimator:

$$\hat{\bar{y}}_{post} = \bar{y}_r - (\bar{y}_r - \sum_h W_h \bar{y}_h) = \sum_{h=1}^{H} W_h \bar{y}_h$$

$$= \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h, \text{ and for the total } \hat{t}_{post} = \sum_{h=1}^{H} N_h \bar{y}_h$$

The poststratified estimator

Weights for each observation in poststraum h:

N_h / n_{rh}, n_{rh} is the size of the response sample in postratum h

Estimating the number of one-person households, Norwegian consumer expenditure survey 1992

Poststrata: Fam. size $x = h$	1	2	3	4	5+
Observed rate with household size 1, $\bar{z}_h$	0.5123 (.6001)	0.0391 (0.0527)	0.0472 (0.0753)	0.0067 (0.0106)	0.0048 (0.0084)

$$z_i = 1 \text{ if household size is } 1$$

$$\hat{t}_{post} = \sum_{h=1}^{5+} N_h \bar{z}_h = 486{,}032$$

Compared to

1) *unweighted estimate = 390,501*

2) *Modelbased estimate = 595,462* (nonignorable nonresponse)

Poststratification reduces the bias about 50%

Imputation

- Used for item nonresponse
- Item nonresponse creates problem even when the nonresponse happens at random, leaves us with few complete cases
- Imputation: filling in for each missing data value by predicting the missing values
- For a given variable *y*, for estimating population total or mean, use estimator constructed for the full sample, based on the observed and imputed data:
  - Imputation based estimator
- Need proper variance estimates
- Also want to produce complete data sets that allow for standard statistical analysis
  - Important that the imputed values reflect the right variation in the data

1. Regression-based imputation methods

Regression imputation

Assume a regression model for Y given x,

where x is available also for the nonresponse group

f.ex. $E(Y \mid x) = \beta x, Var(Y \mid x) = \sigma^2 x$

Estimate β from the response sample s_r with

$\hat{\beta}_r = \sum_{i \in s_r} Y_i \sum_{i \in s_r} x_i$,

and for all $i \in$ nonresponse group, predict y_i with

$y_i^* = \hat{\beta}_r x_i$

Problem: Not enough variation to account for the variability in the nonresponse group

Residual regression imputation

Since $Var\{(Y_i - \beta x_i)/\sqrt{x_i}\} = \sigma^2$

Standardized observed residuals: $e_i = (y_i - \hat{\beta}_r x_i)/\sqrt{x_i}$

For $i \in s - s_r$, draw the value e_i^* at random from the set of standardized observed residuals in the response sample $\{e_j : j \in s_r\}$

Imputed *y*-value is given by:

$$e_i^* = (y_i^* - \hat{\beta}_r x_i)/\sqrt{x_i}$$

that is:

$$y_i^* = \hat{\beta}_r x_i + e_i^* \sqrt{x_i}$$

If the model assumption also includes a distributional assumption, say normality:

Draw imputed values from the estimated $N(\hat{\beta}_r x_i, \hat{\sigma}_r^2 x_i)$

Underlying assumption on the response mechanism:

Missing at random (MAR): Probability of response for unit i may depend on x_i, while independent of y

If basic full sample estimator is the ratio estimator,

$$\hat{T}_R = X \frac{\sum_s Y_i}{\sum_s x_i},$$

then the imputation - based estimator becomes

$$\hat{T}_{R,I} = X \frac{\sum_{s_r} Y_i + \sum_{s-s_r} Y_i^*}{\sum_s x_i}$$

Standard imputation methods, much used in National Statistical Institutes

(i) *Mean* imputation $y_i^* = \bar{y}_r$

Within poststrata: poststratification

(ii) *Hot-deck* imputation (typically within poststrata) :

y_i^* is drawn at random from the observed y values,

with replacement

(iii) *Nearest neighbour* imputation: Find a donor in the response sample based on closeness of auxiliary variables

Nonignorable nonresponse : How to proceed

The response probabilities are assumed to depend on variable of interest

$$P_\psi(R_i = 1 \mid x_i, y_i) \text{ is modeled, } f_\psi(r_i \mid x_i, y_i)$$

Population model for Y_i given x_i: $f_\theta(y_i \mid x_i)$

Joint distribution of Y_i and R_i:

$$f_{\theta,\psi}(y_i, r_i \mid x_i) = f_\theta(y_i \mid x_i) f_\psi(r_i \mid x_i, y_i)$$

Conditional distribution of Y_i given nonresponse, R_i=0

$$f_{\theta,\psi}(y_i \mid x_i, R_i = 0)$$

$$f_{\theta,\psi}(y_i \mid x_i, R_i = 0) = f_\theta(y_i \mid x_i) P_\psi(R_i = 0 \mid y_i, x_i) / P_{\theta,\psi}(R_i = 0 \mid x_i)$$

where

$$P_{\theta,\psi}(R_i = 0 \mid x_i) = \int f_\theta(y_i \mid x_i) P_\psi(R_i = 0 \mid y_i, x_i) dy_i$$

Maximum likelihood estimates: $\hat{\theta}, \hat{\psi}$

Likelihood function, independence between (Y_i, R_i):

$$l(\theta,\psi) = \prod_{i \in s_r} f\ (y_{obs,i} \mid x_i) P_\psi(R_i = 1 \mid x_i, y_{obs.i}) \prod_{i \in s - s_r} P_{\theta,\psi}(R_i = 0 \mid x_i)$$

Note: Likelihood function could be quite flat in ψ, numerical difficulties for finding maximum.

Imputed values: $y_i^* = E_{\hat{\theta},\hat{\psi}}(Y_i \mid x_i, R_i = 0)$

or drawing a value from the estimated conditional distribution

Example

Binomial case $P(Y_i = 1) = \theta$

$$P(R_i = 1 \mid y_i) = \begin{cases} \psi & \text{if } y_i = 0 \\ 2\psi & \text{if } y_i = 1 \end{cases}$$

$$P(Y_i = 1 \mid R_i = 0)$$

$$= \frac{P(Y_i = 1)P(R_i = 0 \mid Y_i = 1)}{P(Y_i = 1)P(R_i = 0 \mid Y_i = 1) + P(Y_i = 0)P(R_i = 0 \mid Y_i = 0)}$$

$$= \frac{\theta(1-2\psi)}{\theta(1-2\psi) + (1-\theta)(1-\psi)} = \frac{\theta(1-2\psi)}{1-\psi-\theta\psi}$$

$$= \theta - \theta\frac{\psi(1-\theta)}{1-\psi-\theta\psi}$$

Maximum likelihood estimates

Let $v = \sum_{s_r} y_i$, the number of "successes" in the response sample and n_r is the size of s_r

Likelihood function

$$l(\theta,\psi) = \prod_{i \in s_r} f_\theta(y_{obs,i}) \prod_{i \in s_r} P_\psi(R_i = 1 \mid y_{obs.i}) \prod_{i \in s - s_r} P_{\theta,\psi}(R_i = 0)$$

$$= \theta^v (1-\theta)^{n_r - v} (2\psi)^v \psi^{n_r - v} (1 - \psi - \theta\psi)^{n - n_r}$$

and

$$\log l(\theta,\psi) = v \log\theta + (n_r - v)\log(1-\theta) + v\log 2 + n_r \log\psi$$
$$+ (n - n_r)\log(1 - \psi - \theta\psi)$$

Likelihood equations:

$$\text{(I)}\ \ \partial l / \partial \psi = 0 \Leftrightarrow \frac{n_r}{\psi} - (1+\theta)\frac{n-n_r}{1-\psi(1+\theta)} = 0$$

$$\Leftrightarrow \psi = \frac{n_r}{n(1+\theta)}$$

$$\text{(II)}\ \partial l / \partial \theta = 0 \Leftrightarrow \frac{v}{\theta} - \frac{n_r - v}{1-\theta} - \psi\frac{n-n_r}{1-\psi(1+\theta)} = 0$$

$$\overset{\text{(I)}}{\Leftrightarrow} \frac{v}{\theta} - \frac{n_r - v}{1-\theta} - \frac{n_r}{n(1+\theta)} \cdot \frac{n-n_r}{1-(n_r/n)} = 0 \Leftrightarrow \frac{v}{\theta} - \frac{n_r - v}{1-\theta} - \frac{n_r}{1+\theta} = 0$$

$$\Leftrightarrow \theta = \frac{v}{2n_r - v}$$

$$\hat{\theta} = \frac{v}{2n_r - v} = \frac{\bar{y}_r}{2 - \bar{y}_r}$$

$$\hat{\psi} = \frac{n_r}{n(1+\hat{\theta})} = \frac{1}{2} \cdot \frac{n_r}{n}(2 - \bar{y}_r)$$

Note:

$$E(\bar{Y}_r) = E(Y_i \mid R_i = 1) = P(Y_i = 1 \mid R_i = 1)$$

$$= \frac{\theta \cdot 2\psi}{\psi + \theta\psi)} = \frac{2\theta}{\theta + 1} \quad (< \theta)$$

A reasonable estimate would satisfy

$$\bar{y}_r = \frac{2\hat{\theta}}{\hat{\theta} + 1}, \text{ that is}: \hat{\theta} = \frac{\bar{y}_r}{2 - \bar{y}_r} = MLE$$

$$y_i^* = E_{\hat{\theta},\hat{\psi}}(Y_i \mid x_i, r_i = 0) = P_{\hat{\theta},\hat{\psi}}(Y_i = 1 \mid R_i = 0)$$

$$= \hat{\theta} - \hat{\theta}\frac{\hat{\psi}(1-\hat{\theta})}{1-\hat{\psi}-\hat{\theta}\hat{\psi}} = \hat{\theta} - \frac{n_r}{n-n_r}(\bar{y}_r - \hat{\theta})$$

Estimate the total number of successes in the population, $T = \sum_{i=1}^{N} Y_i$

Basic estimator without nonresponse: $\hat{T} = N \cdot \bar{Y}_s$

Imputation-based estimate:

$$\hat{t}_I = N\frac{1}{n}(\sum_{i \in s_r} y_i + \sum_{i \in s - s_r} y_i^*)$$

$$= N\frac{1}{n}\{n_r \bar{y}_r + (n-n_r)(\hat{\theta} - \frac{n_r}{n-n_r}(\bar{y}_r - \hat{\theta}))\} = N \cdot \hat{\theta}$$

What happens if we erroneously assume ignorable response mechanism ? Then

$$\hat{\theta} = \bar{y}_r,$$

$$y_i^* = P_{\hat{\theta}}(Y_i = 1) = \hat{\theta} = \bar{y}_r$$

$$\text{Estimate}: \hat{t}_* = N \cdot \bar{y}_r$$

$$\text{Bias}: \frac{1}{\text{N}} E(\hat{T}_* - T) = \{E(\bar{Y}_r) - \theta)\}$$

$$= \frac{2\theta}{1+\theta} - \theta = \frac{\theta(1-\theta)}{1+\theta}$$

$$\frac{1}{N} E(\hat{T}_I - T) \approx \frac{\theta(1-\theta^2)}{n_r}$$

Approximately unbiased for large n_r

Lecture 4: Variance estimation in the presence of imputed values

Consider simplest possible case

- Simple random sample
- Random nonresponse
- No auxiliary information

Two possible imputation methods

(i) *mean* imputation : $y_i^* = \bar{y}_r$

(ii) *hot-deck* imputation :

y_i^* is drawn at random from the observed y values, with replacement

- Mean imputation can not be used if the completed data set shall reflect expected variation in the nonresponse group

- Look at standard design-based analysis based on the completed sample: observed and imputed data

$$\text{Problem: Estimate } \bar{Y}$$

$$\bar{y}_s = \text{sample mean if the whole sample } s \text{ is observed}$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2$$

Large n,N-n

Standard 95% confidence interval:

$$\text{CI}: \bar{y}_s \pm 1.96\hat{\sigma}\sqrt{\frac{1}{n}-\frac{1}{N}}$$

With nonresponse:

Standard CI based on the completed data set with observed and imputed values:

$$\text{CI}^*: \bar{y}_s^* \pm 1.96\hat{\sigma}_*\sqrt{\frac{1}{n}-\frac{1}{N}}$$

$\bar{y}_s^*, \hat{\sigma}_*^2 : \bar{y}_s, \hat{\sigma}^2$ based on the completed sample with observed and imputed values

Coverage with mean imputation

$$W_r = \frac{\bar{y}_r - \bar{Y}}{\hat{\sigma}_r \sqrt{\frac{1}{n_r} - \frac{1}{N}}} \sim N(0,1) \quad \text{approximately}$$

$$\hat{\sigma}_*^2 = \frac{n_r - 1}{n - 1}\hat{\sigma}_r^2 \quad \text{such that} \quad CI^* \approx \bar{y}_r \pm 1.96 \frac{n_r}{n}\hat{\sigma}_r \sqrt{\frac{1}{n_r} - \frac{1}{N}}$$

Confidence level: $C_* = P(\bar{Y} \in CI_*) = P(|W_r| \leq \frac{n_r}{n} 1.96)$

Nonresponse (%)	0	10	20	30	40	50
Conf. level	0.95	0.922	0.883	0.830	0.760	0.673

Coverage with hot-deck imputation

La $\bar{y}^*_{mis} = \frac{1}{n-n_r}\sum_{i\in s-s_r} y^*_i$

$$E(\bar{y}^*_{mis} \mid data) = \bar{y}_r$$

$$Var(\bar{y}^*_{mis} \mid data) = \frac{1}{n-n_r}\cdot\frac{n_r-1}{n_r}\hat{\sigma}^2_r$$

$\Rightarrow$

$$E(\bar{y}^*_s) = EE(\bar{y}_* \mid data) = \bar{Y}$$

$$Var(\bar{y}^*_s) = EVar(\bar{y}^*_s \mid data) + VarE(\bar{y}^*_s \mid data)$$

$$= EVar(\frac{(n-n_r)\bar{y}^*_{mis}}{n} \mid data) + Var(\bar{y}_r)$$

$$= \frac{(n-n_r)^2}{n^2}\cdot\frac{1}{n-n_r}(1-\frac{1}{n_r})\sigma^2 + \sigma^2(\frac{1}{n_R}-\frac{1}{N})$$

$$\approx \sigma^2\left[\frac{(n-n_r)}{n^2}\cdot(1-\frac{1}{n_r}) + \frac{1}{n_r} - \frac{1}{N}\right]$$

$$E(\hat{\sigma}_*^2) \approx \sigma^2$$

$$W_* = \frac{\bar{y}_s^* - \bar{Y}}{\hat{\sigma}_* \sqrt{\frac{n-n_r}{n^2}(1-\frac{1}{n_r}) + \frac{1}{n_r} - \frac{1}{N}}} \sim N(0,1) \quad \text{approximately}$$

Confidence level:

$$C_* \approx P(|W_*| \le 1.96 \sqrt{\frac{1}{n} - \frac{1}{N}} / \sqrt{\frac{n-n_r}{n^2}(1-\frac{1}{n_r}) + \frac{1}{n_r} - \frac{1}{N}})$$

$$\approx P(|W_*| \le 1.96 / \sqrt{1 + \frac{n}{n_r} - \frac{n_r}{n}})$$

Confidence levels of CI*:

Nonresponse (%)	Mean imputation	Hot-deck imputation
0	0.95	0.95
10	0.922	0.925
20	0.883	0.896
30	0.830	0.864
40	0.760	0.826
50	0.673	0.785

One possible solution: Multiple imputation

m repeated hot-deck imputations for each missing value:

m completed samples

$\bar{y}_s^*(i), \hat{\sigma}_*^2(i)$ for $i = 1,\ldots,m$

$\bar{y}_s, \hat{\sigma}^2$ based on the m completed samples

Averages: $\bar{\bar{y}}_s^* = \sum_{i=1}^m \bar{y}_s^*(i)/m$ and $\bar{\sigma}_*^2 = \sum_{i=1}^m \hat{\sigma}_*^2(i)/m$

A "direct" standard CI:

$$\mathrm{CI}^* : \bar{\bar{y}}_s^* \pm 1.96\bar{\sigma}_* \sqrt{\frac{1}{n} - \frac{1}{N}}$$

Also too short

$\bar{\sigma}_* \sqrt{\frac{1}{n} - \frac{1}{N}}$ measures the variation only *within* the samples

It is necessary to include a measure of variation *between* the *m* samples; to measure the uncertainty due to imputation

$$B_* = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{y}_s^*(i) - \bar{\bar{y}}_s^*)^2$$

Rubin's MI method combines the standard analyses as follows:

Replace $\hat{\sigma}_*^2$ with: $V_* = \bar{\sigma}_*^2 (\frac{1}{n} - \frac{1}{N}) + (1 + \frac{1}{m}) B_*$

and corresponding 95% CI: $\bar{\bar{y}}_s^* \pm 1.96 \sqrt{V_*}$

Requires that the imputations are based on a Bayesian model, drawing imputed values from the posterior distribution given nonresponse

How does it work for hot-deck imputation ?

Depends on f_{mis} = the nonresponse rate

It can be shown that the confidence level is approximately equal to, when $m > 1$:

$$C_{*m} = P\left(|N(0,1)| \leq 1.96\sqrt{1 - \frac{f_{mis}^2}{1 + m^{-1} f_{mis}(1 - f_{mis})}} \right)$$

Decreasing as m increases !!

Note that for $m = 1$:

$$C_{*1} \approx P(|N(0,1)| \leq 1.96 / \sqrt{1 + \frac{n}{n_r} - \frac{n_r}{n}})$$

$$= P(|N(0,1)| \leq 1.96 \sqrt{\frac{1 - f_{mis}}{1 + f_{mis}(1 - f_{mis})}})$$

$$== P(|N(0,1)| \leq 1.96 \sqrt{1 - \frac{f_{mis} + f_{mis}(1 - f_{mis})}{1 + f_{mis}(1 - f_{mis})}})$$

Confidence levels

Nonresponse (%)	*m =1* *Hot-deck*	*m* =2 (infinity)
0	0.95	0.95 (0.95)
10	0.925	0.949 (0.949)
20	0.896	0.946 (0.945)
30	0.864	0.940 (0.939)
40	0.826	0.930 (0.927)
50	0.785	0.916 (0.910)

Quite an approvement over single imputation, but this standard MI procedure does not quite achieve the desired cofidence level.

Reason: The hot-deck imputations do not display enough variability. In Rubin's term, the imputations must be "proper".

The between-imputation component must be given a larger weight k than $(1+1/m)$.

Non-Bayesian multiple imputation

- Formulation of problems in non-Bayesian multiple imputation for variance estimation
- An approach for developing a general theory for non-Bayesian MI –how to combine standard procedures
- Simple examples
- For a given data model and imputed values: Can the same combination of standard procedures be used for any estimation problem?

Suggested approach

- Full sample: $s = (1,\ldots,n)$
- Planned data: $\mathbf{y} = (y_1,\ldots,y_n)$
- Objective: Estimate θ
- Response sample s_r of size n_r
- Observed:

$$y_{obs} = \{(y_i : i \in s_r), s_r\}$$

Estimator based on the full sample data **y:** $\hat{\theta}$

Variance estimator: $\hat{V}(\hat{\theta})$

For $i \in s - s_r$ we impute by some method: y_i^*

Completed data:

$$\mathbf{y}^* = (y_i : i \in s_r, y_i^*, i \in s - s_r)$$

Based on $\mathbf{y}^*$: $\hat{\theta}^* = \hat{\theta}(\mathbf{y}^*)$

$$\hat{V}^* = \hat{V}(\hat{\theta}^*)$$

m repeated imputations:

m completed data-sets with *m* estimates $\hat{\theta}_i^*, i = 1,...,m.$

and related variance estimates $\hat{V}_i^*, i = 1,...,m$

The combined estimate: $\bar{\theta}^* = \sum_{i=1}^{m} \hat{\theta}_i^* / m$

The within-imputation variance: $\bar{V}^* = \sum_{i=1}^{m} \hat{V}_i^* / m$

and the between-imputation component :

$$B^* = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i^* - \bar{\theta}^*)^2$$

The total estimated variance of $\bar{\theta}^*$:

$$W = \bar{V}^* + (k + \frac{1}{m}) B^*$$

Must determine k such that

$$E(W) = Var(\bar{\theta}^*)$$

Typically:

$$E(\overline{V}^*) \approx Var(\hat{\theta})$$

$$E(B^* \mid y_{obs}) = Var(\hat{\theta}^* \mid y_{obs})$$

$$\Rightarrow \quad E(W) = Var(\hat{\theta}) + (E(k) + \frac{1}{m}) EVar(\hat{\theta}^* \mid Y_{obs})$$

approximately

$$Var(\overline{\theta}^* \mid y_{obs}) = Var(\hat{\theta}^* \mid y_{obs}) / m$$

$$E(\overline{\theta}^* \mid y_{obs}) = E(\hat{\theta}^* \mid y_{obs})$$

$$\Rightarrow Var(\overline{\theta}^*) = \frac{1}{m} E\{Var(\hat{\theta}^* \mid Y_{obs})\} + Var\{E(\hat{\theta}^* \mid Y_{obs})\}$$

$$Var(\hat{\theta}) + \{E(k) + \frac{1}{m}\} EVar(\hat{\theta}^* \mid Y_{obs})$$

$$= \frac{1}{m} EVar(\hat{\theta}^* \mid Y_{obs}) + VarE(\hat{\theta}^* \mid Y_{obs})$$

$$\Leftrightarrow E(k) EVar(\hat{\theta}^* \mid Y_{obs}) = VarE(\hat{\theta}^* \mid Y_{obs}) - Var(\hat{\theta})$$

$$E(k) = \frac{VarE(\hat{\theta}^* \mid Y_{obs}) - Var(\hat{\theta})}{EVar(\hat{\theta}^* \mid Y_{obs})}$$

Three examples

- Estimating population average with hot-deck imputation
- Estimating regression coefficient with residual imputation
  - Ratio model
  - Linear regression model
- Nonresponse mechanism: MCAR

Population average with hot-deck imputation

$$p_r = P(R_i = 1)$$

$$E(k) = \frac{1}{p_r}$$

$$k = \frac{1}{1 - f_{mis}}, f_{mis} = (n - n_r)/n$$

Regression coefficient with residual imputation

- Ratio model:

$$Y_i = \beta x_i + \varepsilon_i \qquad Var(\varepsilon_i) = \sigma^2 x_i$$

$$i = 1, \ldots, n$$

<u>All</u> x_i's are known, also in the nonresponse sample

The full data estimator of β: $\hat{\beta} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} x_i$

Estimate based on observed sample s_r: $\hat{\beta}_r$

Imputation method

The standardized residuals:

$$e_i = (y_i - \hat{\beta}_r x_i) / \sqrt{x_i}$$

For $i \in s - s_r$: Draw e_i^*

at random from the set of observed residuals

$$e_i, i \in s_r$$

The imputed y-values:

$$y_i^* = \hat{\beta}_r x_i + e_i^* \sqrt{x_i}$$

$$X_s = \sum_{i=1}^{n} x_i$$

$$X_r = \sum_{i \in s_r} x_i$$

$$X_{nr} = \sum_{i \in s - s_r} x_i = X - X_r$$

All considerations : conditional on n_r and X_r

$$f_X = X_{nr} / X_s$$

$$k \approx \frac{1}{1 - f_X}$$

Regression coefficient with residual imputation

- Linear regression model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \ , \ Var(\varepsilon_i) = \sigma^2 , \ i = 1,\ldots,n$$

$$\text{Observed residuals}: e_j = (y_j - \hat{\alpha}_r - \hat{\beta}_r x_j) \ , \ j \in s_r$$

For $i \in s - s_r$: Draw e_i^*
at random from the observed residuals,
with replacement

The imputed y-values: $y_i^* = \hat{\alpha}_r + \hat{\beta}_r x_i + e_i^*$

$$k = 1/(1 - f_{mis})$$

Future research:
1. For the same situation and imputation method: Use the same *k* for all estimands ?

- General answer: No
- Illustration by example 1, hot-deck imputation
- For estimating population mean with sample mean:
- $k = 1/(1- f_{mis})$
- This *k* is not valid for all other estimation problems
- Start by look at this problem in the simplest case

- MCAR, the response variables R_i are independent $p_r = P(R_i = 1)$
- Hot-deck imputation
- No assumption on sampling design, general $p(s)$
- Three possible cases:
  1. s is a sample from a finite population with design model. The observed stochastic variables are (s, s_r) and y_{obs} is equivalent to (s, s_r)
  2. Same as in case 1, but with a population model. Then

$$y_{obs} = \{(y_i : i \in s_r), s_r, s\}$$

  3. Observational study where **y** is modeled and

$$y_{obs} = \{(y_i : i \in s_r), s_r\}$$

One obvious requirement : $E(\hat{\theta}^* \mid \mathbf{y}, s) = \hat{\theta}$

$$\text{Case 1}: E(\hat{\theta}^* \mid s) = \hat{\theta}$$

$$\text{Case 2}: E(\hat{\theta}^* \mid \mathbf{y}, s) = \hat{\theta}$$

$$\text{Case 3}: E(\hat{\theta}^* \mid \mathbf{y}) = \hat{\theta}$$

Consider estimates linear in $(y_i : i \in s)$: $\hat{\theta} = \sum_{i \in s} a_i(s) y_i$

Theorem

$\hat{\theta}$ satisfies $E(\hat{\theta}^* \mid \mathbf{y}, s) = \hat{\theta}$ if and only if

$$a_i(s) = a(s) \quad \text{for all} \quad i \in s$$

and then $E(k) \approx 1/p_r$

and we can use $k = 1/(1 - f_{mis})$

1. $a(s) = 1/n$ sample mean, $k = 1/(1-f_{mis})$

2. Regression coefficient for regression through the origin:

$$\hat{\beta} = \sum_{i=s} y_i / \sum_{i \in s} x_i \qquad a(s) = 1/\sum_{i \in s} x_i$$

$$k = 1/(1 - f_{mis})$$

3. A case where (*) does not hold : regression coefficient in linear regression not through the origin:

$$\hat{\beta} = \frac{\sum_{i \in s} (x_i - \bar{x}_s) y_i}{\sum_{i \in s} (x_i - \bar{x}_s)^2} \qquad a_i(s) = \frac{x_i - \bar{x}_s}{\sum_{j \in s} (x_j - \bar{x}_s)^2}$$

$$E(\hat{\beta}^* \mid s) \approx p_r \hat{\beta}$$

In case 3 we know from earlier that we can use $k=1/(1-f_{mis})$ with residual hot-deck imputation

For regular regression problems hot-deck imputation cannot work. Obviously:

When y is correlated to known x in nonresponse group:

One should utilize this in the imputations regardless of the estimation problems one consider.

Need to generalize these results:

- To other response mechanism, at least MAR, where the response probablities depend on auxiliary variables, known for the whole sample
- To more general imputation methods, like nearest neighbourhood

Future research: Other issues to be studied

- Robustness of the choice $k = 1/(1-f_{mis})$, in simulation studies
- k is a measure of the proportion of missing information in the response sample as compared to the full sample. Generalize this by defining relevant measures of missing information
- Study of coverage of confidence intervals

- Missing data in explanatory variables are common

in observational studies like epidemiological research.

Some initial simulations indicate that $k = 1/(1-f_{mis})$

can be used in this case with residual hot-deck imputation

- Compare this MI method with alternative methods for variance estimation with imputed data (jackknife,bootstrap)
  - Conditions under which the approaches are valid
  - Extend to which the methods provide unified approaches for sets of estimands
  - Standard criteria: efficiency of point and variance estimation
  - Computational burden

Some references- after topic

- **Standard imputation methods**
  - Lundström and Särndal (2002). *Estimation in the presence of Nonresponse and Frame Imperfections,* ch.7. CBM, Statistics Sweden
- **Nonignorable nonresponse**
  - Baker and Laird (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association,* 83, 62-69.
  - Forster and Smith (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse (with discussion). *Jorunal of the Royal Statistical Society B,* 60, 57-70.
  - Greenlees, Reece and Zieschang (1982). Imputation of missing values when the probability of response depends on the variable being imputed.

- Little (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association,77,* 237-250.
- Belsby, Bjørnstad and Zhang (2005). Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse. Applied to the Norwegian Consumer Expenditure Survey. *Survey Methodology,*31, 197-212.

- **Variance estimation with non-Bayesian imputed data**
  - Bjørnstad (2007). Non-Bayesian Multiple Imputation (with dsicussion). *Journal of Official Statistics*, 23, 433-491
  - Shao and Sitter (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association,* 91, 1278-1288.