Mixed Mode data collection

A Formal Approach

by

Willem E.Saris

Robert Voogt

Notation

- The vector **f** denotes the frequencies in the different classes of a variable in the population
- We distinguish two processes in the data collection:
  - selection processes for method i is denoted by a diagonal matrix $\mathbf{S}_i$
  - transformation processes for method i is denoted by a matrix $\mathbf{P}_i$

Illustration

A potential frequency distribution of the variable "Benefit from EU membership" in the population.

opinions	absolute frequency **f**	relative frequency (1/N) **f**
Benefited	10.0 million	.625
Not benefited	5.0 million	.312
DK-No answer	1.0 million	.063
Total population	16.0 million	1.0

Selection process 1: sampling

A potential frequency distribution of the variable "Benefit from EU membership" in the sample.

opinions	absolute frequency in the population **f**	expected frequency in the sample $\mathbf{f_s}=s_s\mathbf{f}$	relative frequency in both $(1/n)\mathbf{f_s}$
Benefited	10.0 million	10 thousand	.625
Not benefited	5.0 million	5 thousand	.312
DK-No answer	1.0 million	1 thousand	.063
Total population	16.0 million	16 thousand	1.0

Selection process 1: sampling

If the probabilities are placed in a diagonal matrix, the outcome of this sampling procedure with equal probabilities can also be presented in matrix notation:

$$\begin{vmatrix} f_s(1) \\ f_s(2) \\ f_s(3) \end{vmatrix} = \begin{vmatrix} s_s & 0 & 0 \\ 0 & s_s & 0 \\ 0 & 0 & s_s \end{vmatrix} \begin{vmatrix} f(1) \\ f(2) \\ f(3) \end{vmatrix}$$

or

$$\mathbf{f}_s = \mathbf{S}_s.\mathbf{f}$$

Selection process 2: Response

Selection by nonresponse:

$$\begin{vmatrix} f_n(1) \\ f_n(2) \\ f_n(3) \end{vmatrix} = \begin{vmatrix} s_{n1} & 0 & 0 \\ 0 & s_{n2} & 0 \\ 0 & 0 & s_{n3} \end{vmatrix} \begin{vmatrix} f(1) \\ f(2) \\ f(3) \end{vmatrix}$$

or $\mathbf{f}_n = \mathbf{S}_n.\mathbf{f}$

For the response group the same equation can be specified using the matrix S_r which is defined by

$$\mathbf{S}_r = \mathbf{I} - \mathbf{S}_n$$

Transformation Process

Giving the response to a question people can move from one category to a different category.
Therefore the answering process itself is characterised by a transformation process which requires a full matrix

$$\begin{vmatrix} f_{oi}(1) \\ f_{oi}(2) \\ f_{oi}(3) \end{vmatrix} = \begin{vmatrix} p_{i11} & p_{i12} & p_{i13} \\ p_{i21} & p_{i22} & p_{i11} \\ p_{i31} & p_{i32} & p_{i33} \end{vmatrix} \begin{vmatrix} f_r(1) \\ f_r(2) \\ f_r(3) \end{vmatrix}$$

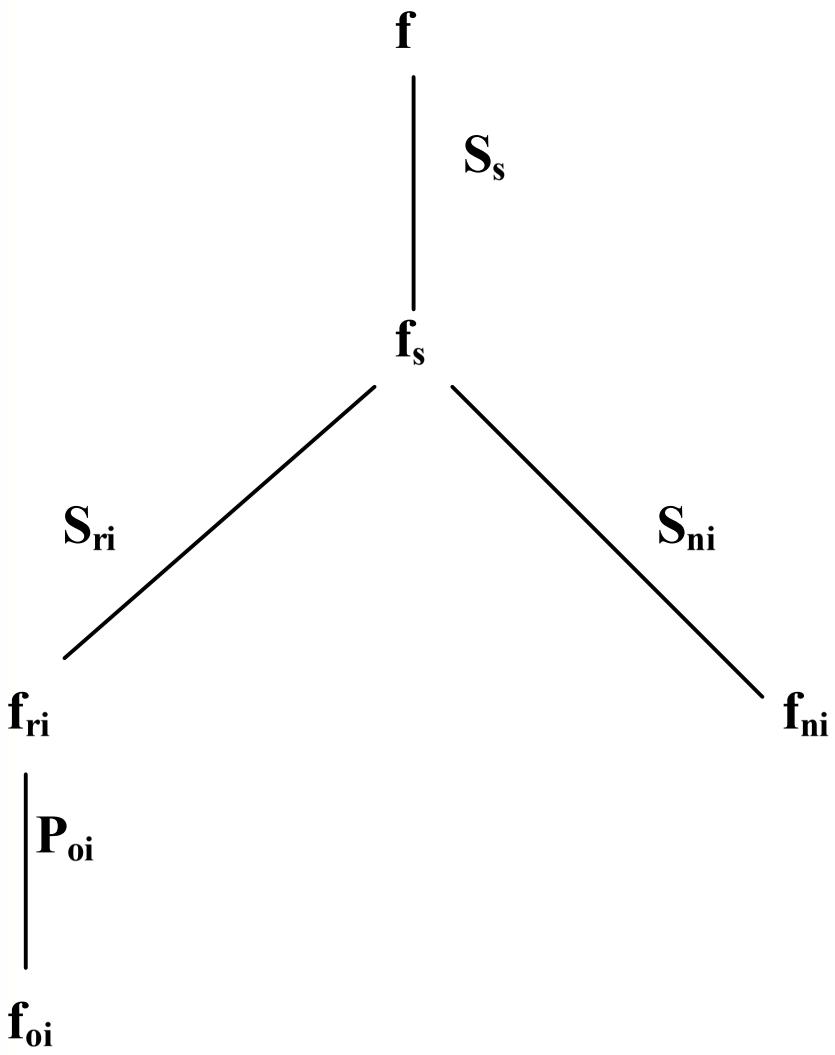
or

$$\mathbf{f}_{oi} = \mathbf{P}_i \mathbf{f}_r$$

Causes of Bias

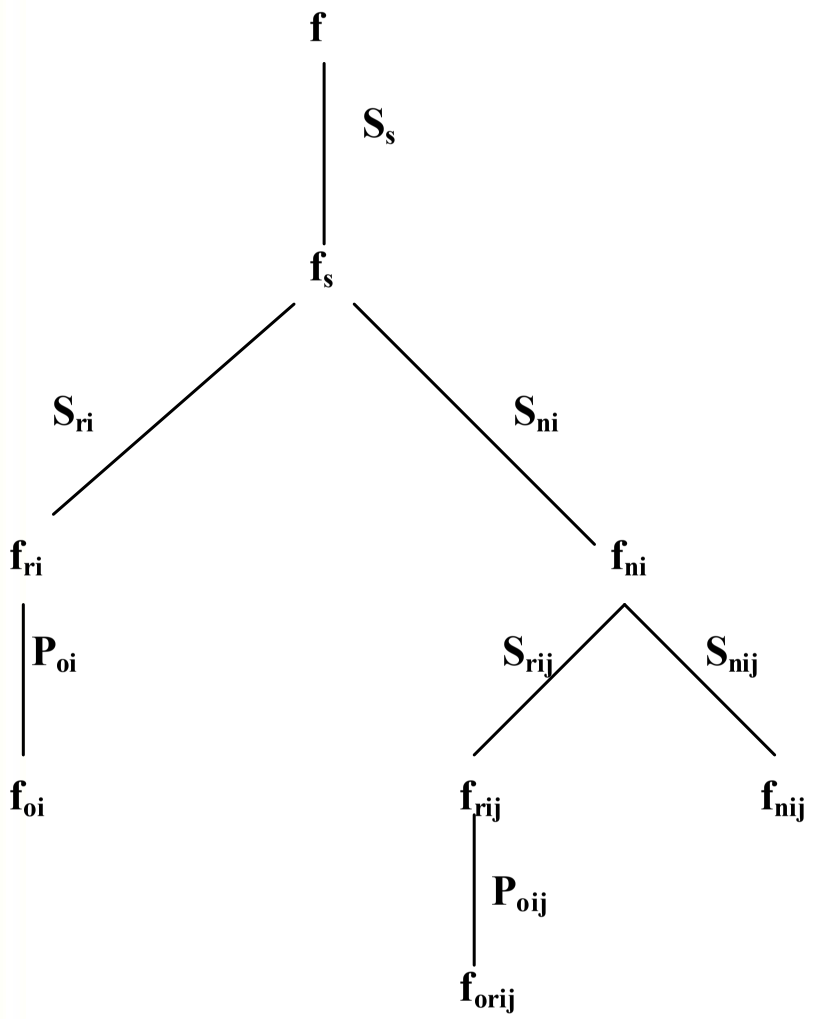
- Bias is defined as a difference between the population distribution and the expected observed distribution
- If in the selection matrices S the diagonal elements are not equal, the observed distribution will not be equal to the distribution in the population
- If the transformation matrix P_i is not an identity matrix the observed distribution will differ from the distribution in the population

A different presentation



- The response distribution is:
- $\mathbf{f}_{oi} = \mathbf{P}_{oi}\mathbf{S}_{ri}\mathbf{S}_s\,\mathbf{f}$
- The observed distribution is equal to the distribution in the population if the matrix $\mathbf{R}_i = \mathbf{P}_{oi}\mathbf{S}_{ri}\mathbf{S}_s$
- is diagonal and
- has equal probabilities on the diagonal

Extension to Mixed mode



- The response distribution is:
- $\mathbf{f}_{oi+j} = (\mathbf{P}_{oi}\mathbf{S}_{ri}\mathbf{S}_s + \mathbf{P}_{oij}\mathbf{S}_{nij}\mathbf{S}_s)\,\mathbf{f}$
- The observed distribution is equal to the distribution in the population if the matrix
- $\mathbf{R}_{i+j} = \mathbf{P}_{oi}\mathbf{S}_{ri}\mathbf{S}_s + \mathbf{P}_{oij}\mathbf{S}_{nij}\mathbf{S}_s$
- is diagonal and
- has equal probabilities on the diagonal

Two estimates of **f**

- To estimate the distribution in the population one can use:
- $\mathbf{f}_{oi+j}$ obtained using mixed mode data collection
- or
- $\mathbf{f}_{oi+j} = \mathbf{R}_{i+j}\,\mathbf{f}$ obtainable using latent variables models or $\mathbf{f} = \mathbf{R}_{i+j}^{-1}\,\mathbf{f}_{oi+j}$

- The latter approach requires estimation of all selection and transformation matrices

An Illustration of the first method

- In a election study in Zaanstad a two stage probability sample was drawn from the voting register: first voting districts and then voters
- The fact whether the people had voted or not is recorded in the register.
- So the selection and transformation processes can be followed exactly.
- We ignore the sampling

Results by telephone

- Distribution in sample: .729 voted .271 did not
- selection by telephone presence
- selection by response
- selection long interview
- transformation by observation

$$\begin{vmatrix} 618 \\ 192 \end{vmatrix} = \begin{vmatrix} .852 & .000 \\ .000 & .711 \end{vmatrix} \begin{vmatrix} 725 \\ 270 \end{vmatrix}$$

$$\begin{vmatrix} 478 \\ 98 \end{vmatrix} = \begin{vmatrix} .774 & .000 \\ .000 & .510 \end{vmatrix} \begin{vmatrix} 618 \\ 192 \end{vmatrix}$$

$$\begin{vmatrix} 420 \\ 71 \end{vmatrix} = \begin{vmatrix} .879 & .000 \\ .000 & .724 \end{vmatrix} \begin{vmatrix} 478 \\ 98 \end{vmatrix}$$

$$\begin{vmatrix} 434 \\ 57 \end{vmatrix} = \begin{vmatrix} .998 & .211 \\ .002 & .789 \end{vmatrix} \begin{vmatrix} 420 \\ 71 \end{vmatrix}$$

Bias by telephone use

- In the sample as a whole 72.9% voted and 27.1% did not
- Each selection process increases the number of voters relative to the non-voters
- Also the transformation process increases the number of voters relative to the non-voters

- The observed result was 88.4% voters versus 11.6% non-voters. The bias is 15.5%

Results for second approach by mail

- selection by non-response to telephone approach
- selection by response to mail approach
- selection by long interview
- transformation by answering

$$\begin{vmatrix} 140 \\ 94 \end{vmatrix} = \begin{vmatrix} .226 & .000 \\ .000 & .490 \end{vmatrix} \begin{vmatrix} 618 \\ 192 \end{vmatrix}$$

$$\begin{vmatrix} 34 \\ 15 \end{vmatrix} = \begin{vmatrix} .243 & .000 \\ .000 & .160 \end{vmatrix} \begin{vmatrix} 140 \\ 94 \end{vmatrix}$$

$$\begin{vmatrix} 27 \\ 9 \end{vmatrix} = \begin{vmatrix} .794 & .000 \\ .000 & .600 \end{vmatrix} \begin{vmatrix} 34 \\ 15 \end{vmatrix}$$

$$\begin{vmatrix} 28 \\ 8 \end{vmatrix} = \begin{vmatrix} .963 & .222 \\ .037 & .778 \end{vmatrix} \begin{vmatrix} 27 \\ 9 \end{vmatrix}$$

Bias after mail survey

- The non-response group of the telephone approach contains only 59.8% voters
- Selection and transformation increases this percentage again relative to the non-voters
- In the end the response group of the mail questionnaire contains 77.7% voters
- Combining the two groups hardly improves the estimate of the percentage of voters (87.6%)
- The difference is still 14.7%

Does it make sense to continue ?

- The non-response group after these two approaches contains 185 respondent of which 57% voters.
- These have been contacted by a face to face interviewer.
- In the end 99 answered the questions of which 63.8% said to have voted.
- With a response of 77% of the 810 telephone owners we get 83.8% voters.
- The bias is still 10.9% but with respect to the sample with a telephone only 7.5%.

A mixed mode design is necessary

- The telephone owners are already more often going to vote (76.3%) than the total sample (72.9%)
- In order to get a better estimate one has also to contact the respondents who do not own a telephone (185 persons).
- This is done by mail and next by face to face interviews.
- The result was that 90 out of 142 people who responded said to have voted which is 63.3%
- This group is really different.

Finally

- Adding these respondents to the others the resulting 768 people who replied contained 80% voters.
- This is still 7.1% too high
- In this study the people who did not answer in any phase of the process were also offered the possibility to answer a shorter questionnaire
- In this group 60% of the respondents said that they had voted
- Adding these people to the sample the percentage voters becomes 76.9% and the bias is only 4%.

Conclusions of method 1

- By use of mixed mode data collection in different stages the estimate of the response distribution became better and better.
- There were considerable selection effects and transformation effects
- The latter prevents to get a perfect result what ever we do because the respondents did not give the correct answer, especially the non-voters

Estimation using a latent variable model

- If one can estimate the different selection and transformation matrices one can estimate the distribution in the population using a latent variable model
- This requires special research designs.
- Selection by sampling can be controlled by the choice of the sampling design
- Selection by coverage error can be estimated in different ways depending on the sample frame

Estimation using a latent variable model

- Recently it has been shown that good estimates of the selection by response can be obtained by use of interviews with nonrespondents (Stoop 2005 and Voogt 2004)

- The transition matrices can be estimated using repeated observation (Saris 1996)

An illustration

$$S = \begin{vmatrix} .58 & .00 \\ .00 & .263 \end{vmatrix}$$

$$P = \begin{vmatrix} .998 & .211 \\ .002 & .789 \end{vmatrix}$$

$$R = \begin{vmatrix} .560 & .055 \\ .001 & .208 \end{vmatrix}$$

$$R^{-1} = \begin{vmatrix} 1.724 & -.456 \\ -.0083 & 4.809 \end{vmatrix}$$

- For the previous example we have found:
- S is selection matrix
- P is the transformation matrix
- R = P.S
- Using R^{-1} the population distribution can be estimated

Conclusion of method 2

- If the selection and transformation matrices have been estimated the response distribution can be estimated by a single mode procedure as :
- $\mathbf{f} = (\mathbf{P}_{oi}\mathbf{S}_{ri}\mathbf{S}_{s})^{-1}\,\mathbf{f}_{oi}$
- or by a mixed mode procedure
- $\mathbf{f} = (\mathbf{P}_{oi}\mathbf{S}_{ri}\mathbf{S}_{s} + \mathbf{P}_{oij}\,\mathbf{S}_{nij}\mathbf{S}_{s})^{-1}\,\mathbf{f}_{oi+j}$

- The second requires far more information

Discussion

- We showed that the estimate $\mathbf{f}_{oi+j}$ for **f** obtained by a elaborate mixed mode data collection approach is most likely biased
- The latent variable approach seems to suggest that an unbiased estimate is possible

- This is a nice topic for further statistical research
- It would also be interesting to study why one need mixed mode data collection if one can get a good estimate with one mode