Survey Sampling Methods

Moritz Baten, Burak Demiral, Sergej Kaiser

June 21, 2016

1 Critical reading

In the following we are going to first summarize "The problem of nonrespone in sample survey" by Hansen and Hurwitz [2004]in combination with "Multiphase Sampling in Nonresponse Problems" by Srinath [1971]. Hurwitz and Hansen describe sampling procedures to obtain a cost optimal allocation of survey mail and follow up face to face interviews. Substaintially the authors argue that combining both one may obtain the advantages from both methods, thus low cost of mail with higher repsonse and infromation conerning the representativity of the sample to the population. - first point sample size is only one factor of sampling error - shows based on a variance formula that same reliability may arrise from various sample sizes - shows optimal number of mails and in person interviews for a desired average error can be calculated if an approximate response rate is available. - if the response rate is unkown the authors showed that the optimal number of mails and in person interviews can be obtained for a given range of reponse rates. - further highlighted two alternatives plans to achieve minimum cost for a desired avg. level of precision (for a response rate of 100% assuming similarity between respondents and non respondents). - Alternative 1 initially send out 1000 mails and follow up on a fraction of the non respondents depending on the square root of the expected unit cost divided by the unit cost of sending a questoinare plus the expected unit cost of processing returned questionnaires. Result: Higher cost than optimal combination, for response rate ¿30, cost increase from 10 to 24 percent - Alternative 2 preferable if approximate reponse is unkown, specify maximum number of responss sent out. Second step after sample return and response rate is known, set in person interviews such that desired level of precision is achieved - Striking result for alternative 2 even though response rate not known, for a given level of precision may be reached for low level of cost (not much higher than optimal level of cost). - Critique: Cost depend on approx. rate of response. This may be a problem in practice that either the targeted avg. sampling error is missed or cost budget is out of control. - Further mailing may increase a missing - Further short comming is that the authors only focus on non response. The reliability of an sample depends as well on the quality of the respondents answers. K.P. Srinath extends the sampling plan of Hansen and Hurwitz. The authors contribution is that he proposes to use subsampling fractions, which vary in response to sample non response. The authors variance of the estimator which do not depend on the unkown sampling variance and the subsampling fractions are adjusted to ensure a fixed precision.

2 Own Analysis of Survey Data

The joint effects of age (in 7 categories), income level (in 3 categories) and sex on having a stable general practitioner or not were studied by running a logistic regression with dummy coding. There are two analyses one taking the survey design into account and the other not (i.e. analyzing the survey as if simple random sampling was used). There were 436 units with variable values not available, they were deleted pairwise, so in total 8128 units of observations were used.

Here we will compare a regression equation where none of the design features were taken into account and a one where all of them; finite population correction, stratification, clustering and weighting which adjusts for the differential selection probability are taken into account. Tables below show the estimates for log odds and standard errors for all the variables.

	Level	Estimate	S.E.	t value	$Pr > \|t\|$
Intercept		-3.13	0.22	-14.16	0.0
SEX	1	0.44	0.08	5.78	0.0
AGE7	1	1.05	0.22	4.69	0.0
AGE7	2	1.33	0.21	6.21	0.0
AGE7	3	1.14	0.22	5.31	0.0
AGE7	4	0.72	0.23	3.19	0.0
AGE7	5	0.67	0.23	2.88	0.0
AGE7	6	0.26	0.24	1.08	0.3
FA3	1	-0.19	0.10	-1.84	0.07
FA3	2	-0.45	0.11	-3.96	0.0

Table 1: No consideration for survey design

	Level	Estimate	S.E.	t value	$Pr > \|t\|$
Intercept		-3.45	0.49	-7.08	0.0
SEX	1	0.32	0.10	3.11	0.00
AGE7	1	1.03	0.48	2.14	0.03
AGE7	2	1.16	0.47	2.45	0.01
AGE7	3	1.11	0.48	2.32	0.02
AGE7	4	0.60	0.48	1.25	0.21
AGE7	5	0.20	0.49	0.41	0.68
AGE7	6	0.24	0.53	0.46	0.64
FA3	1	-0.12	0.19	-0.63	0.53
FA3	2	-0.41	0.21	-1.98	0.05

Table 2: Consideration for survey design

Parameter estimates vary slighly between the two designs. The tables show that all of the standard errors went up, doubled for most of the parameters when the survey design was taken into account. As dummy coding was used estimate for the

intercept shows the log odds for the reference category (female, older than 75 and highest income group) and the parameter estimates show the results compared to the reference of their category, in this case odds of having a stable GP is 4% of not having one ($exp(-3.45)$). Men have higher odds of having a GP than women. Odds of having a GP gets lower as persons get older and poorer.

Below are Wald tests for parameters. In both cases age and sex effects are significantly different than zero, whereas when design is taken into account income effect is not significantly different than zero at % 5 level, but it is significantly different than zero when design is not taken into account.

Effect	DF	Wald Chisq	Pr > ChiSq
SEX	1	33.36	0.0
AGE7	6	89.80	0.0
FA3	2	17.08	0.0

Table 3: Type 3 Analysis of Effects w/out design

Effect	DF	Wald Chisq	Pr > ChiSq
SEX	1	9.65	0.0
AGE7	6	38.87	0.0
FA3	2	5.25	0.07

Table 4: Type 3 Analysis of Effects w/ design

3 Design of Health Interview Survey

3.A Description of the Belgian Health Survey

The sample for the Health Interview Survey (HIS) 2013 consists of 10.000 units each of which selected from the National Registry throughout the year. Belgium consists of 3 regions and in order to be able to talk about these regions, the sample is stratified into 3; 3.500 of which from Flanders, 3.500 Wallonia and 3.000 Brussels.[1] The regions are further stratified into provinces but this time proportional to their size so that a province won't be over or underrepresented.[2] This stratification ensures a geographical spread and if the provinces are homogenous within themselves it should result in higher precision compared to simple random sampling. In the end this results in total 12 strata.

This 2-step stratification is followed by a multi-stage sampling with three levels in total. First the municipalities within provinces, then households within municipalities and lastly individiuals within households.

In order to select the municipalities systematic sampling method was used in each province. The way this is done in a province is as follows: First municipalities were sorted in a descending order according to their population size and listed with a sum of population of the municipality itself and the cumulative sum of previous municipalities upto that municipality (C). If we want to pick N groups we set an interval I (total population in the province divided by N) according to that number N. And we iterate through the list until we go through N steps and pick N groups, at each step we calculate the threshold $((step_no-1)*I+R)$ and pick a group of 50 from the municipality with the closest C that is greater than this threshold. It should be noted that it is possible to select more than one groups from one municipality.

In the 2013 survey 225 groups of 50 were selected from 158 municipalities out of 589. This selection procedure ensures that large municipalities are selected since they are at the beginning of the list and also at least one of the smaller municipalities since they are lumped together at the end of the list. If it is assumed that smaller provinces share similar charecteristics it should be enough to pick at least one of them. Selecting groups of 50 is creates what is called a clustering effect decreasing the precision, similar individuals compared to a simple random sample, but this is still done in order to decrease the amount of effort in the fieldwork.

After selecting the municipalities that the groups of 50 will come from now another sampling is needed to be done in order to determine the households. This is also achieved by systematic sampling, but this selection is done throughout the year in four quarters, because it is likely that households move or people die within a year but also to take into account the temporal representativity, so average number of interviews per group per quarter is then 12.5.

This time households are ordered hierarchically according to 3 categories, first by (1) statistical sector, this is only relevant for municipalities that will give more than 1 group of 50 to the sample[3], then by (2) household size and lastly by (3) age of the reference person.

[1] On top of this there is also 600 additional sampling units for the province of Luxembourg as they payed extra money for this in order to acquire the desired bound of error for their province.

[2] The province Liège was further stratified into two in order to study the German Community.

[3] what is done is that in these municipalities are divided into statistical sectors so that one group comes out of each sector

As before a step size is calculated by dividing the total number of households (N) by the required number of households (n, this is simply calculated by dividing the adjusted mean household size[4] size to 12.5) but this time instead of just picking one household from the list three other consecutive households are also taken from the list. This group of 4 households is called a cluster, this is done in order to have replacement households in case are non-responses from the selected household. The step size further divided by 2 so that there are double the amount clusters that is needed. And again a random number is picked and each at iteration the step size is added to this random number and the household with the ordering closest to this number and the consecutive 3 is picked. In the end the clusters are listed and the first available household within a cluster is selected for the interview.

This sampling taken from an ordered list ensures that municipalities are well represented with respect to statistical sectors, household sizes and ages of the reference person.

Within a household maximum number of people to interview was capped at 4 in order to avoid familial correlation. This is avoided, because individuals in the same household tend to be similar compared to the individuals in other households, so interviewing one more individual within a household doesn't give much information about the sample in general. Then if a household contains more than 4 individuals there has to be a selection rule to pick 4 members out of the total household. Ideally this selection procedure should be done in total randomness in order to avoid bias, but there are some practical problems with this as it may be difficult to explain the reference person that they will not be interviewed and also the information for the general household questionnaire should come from the reference person, thus for the households with more than 4 members, reference (and the partner if there is one) is always interviewed and the following 3 (or 2 if there is a partner) members are selected by the birthday rule, which is whoever has the closest upcoming birthday to the day of interview is selected and since not everyone has the same selection probability in households with more than 4 member, weights are attached in proportion to inverse of the selection probability in order to mitigate the bias.

3.D Comparison of European Social Survey with General Social Survey (US)

Since the ESS survey is a rather young survey, it seems fitting to compare it to an incumbent with a lot of experience, the GSS from the United States.

The GSS uses a multi-stage area probabiltity sampling. The Primary Sampling Units are the Standard Metropolitan Statistical Areas (SMSAs) or counties in rural areas. Both are stratified by region, age and race before selection. The Secondary Sampling Units (SSU) are block groups (BG) or enumeration districts (ED). The latter describe an area which one interviewer can cover. These BGs and EDs are also stratified to race and income before selection. The Tertiary Sampling Units (TSU) are blocks who are selected proportional to size. In places without block statistics field counting is used.

[4]Calculated as the mean household size but households with more than 4 members are counted as 4 member households as only 4 people are interviewed per household

In addition quotas on sex, age and employment status are used to account for the not-at-homes: District figures from census data are used to keep the sex proportional. Similarly proper portion of employed and unemployed women and proper portion of men over and under 35 were enfored. The argumentation here is that especially young women and men under 35 are difficult to sample otherwise.

Furthermore non-respondent cases are sub-sampled at the end of sampling (about ten weeks before the first release) period to spent more ressources on these difficult cases. This sub-sample is then weighted and added to the final sample.

	GSS	ESS
PSU	counties or metropolitan areas	census sections
stratification	yes: region, race and age	yes: region and town size
SSU	block groups or districs	list of individuals†
stratification	yes: race and income	no
TSU	blocks	–
stratification	no	no
Quota Sampling	yes on sex, age and employment	no, forbidden
oversampling	yes	not recommended, requires authorisation
non-respondent actions	oversampling, non-resp. sub-sample	urge to use more contact attempts

† applies to ESS countries who have a registry, for others fieldwork agencies and random route techniques are used

Table 5: Comparison of ESS and GSS

The ESS uses a two stage sampling procedure. The PSU are selected usually sections of the last census. The PSUs are stratified by region and town-size before selection. Afterwards they are ordered per target population and from this lists then the individuals are sampled (by SRS). There is no stratification on this level.

This procedure is applied if registries are available. In countries where no reliable lists of addresses or households is available like Portugal or Bulgaria PSUs are used but then use fieldwork agencies or random route techniques. Fieldwork agencies create lists of people within an area, while random route techniques use algorithms to randomly select sampling units. However it is not clear how random these random route techniques are. Thus they should be avoided if possible (Lyberg, 2000).

For smaller countries the ESS uses a minimum size of 800 for countries with a population of less than two million. In addition ESS urge the national coordinators to not over-sample anticipated low response rates.

4 Methodology

4.1 What are missing data in surveys?

Survey data may contain different kind of non response. First unit non response describes the situation if there is no data of the target unit (mostly a person) or item non response describes that a sampling unit did not respond to a particular item.

Schafer and Graham [2002] describe that traditionally for the first kind survey statistician used reweighing whereas for the latter they used single imputation. Further the author note that this methods may provide in special cases similar performance as modern methods like maximum likelihood based methods or multiple imputation methods, which are more general.

In the following we will first outline a missing data classification based on Rubin [1976]. Next, based on the classification we explain why we need missing data techniques. The discussion of missing data techniques follows then. In particular, we will discuss traditional methods and compare them to more advanced methods. Finally, we discuss modern missing data methods.

4.2 Which kind of missing data exist?

Since Rubin [1976] missing data is analyzed with probability models. To illustrate we will discuss the case of an arbitrary pattern of nonresponse. In a survey of Y questions, we may or may not observe a nonresponse to a question from each participant. We will record for each question the non response with a binary variable R_Y. The variable value is equal to zero, if the survey participant answered a particular question and is one if he did not answer. For an arbitrary non response pattern our survey data set, where each of the Y questions corresponds to a separate column and each of the n survey participants corresponds to a row, we thus obtain a set of binary indicators R. The innovation of Rubin [1976] is to study the missingness mechanism Little and Rubin [1986].The missingness mechanism describes the relationship between the missingness and the variables in the data set of variables. Rubin [1976] showed how to study the mechansim by treating R as a random variable with a distribution.

Based on the missing data mechanism we can classify non response into three categories missing completely at random (MCAR), missing at random(MAR), not missing at random (MNAR), we discuss each category in turn. Missingness completely at random describes that the probability that a value is missing is independent of our missing or observed data and the missing data mechanism. Missing at random describes that the probability of a missing value does only depend on our observed data but not on the missing values. Last the assumption not missing at random describes that missingness depends on observed and unobserved parts of the data.

4.3 Why are missing data techniques necessary?

The concerns about non response are (1) efficiency losses (2) complications in data handling and data analysis (3) bias due to differences between data values for those between respondents and non respondents (cf. Schafer [1999])

Point (1) and (3) can be ignored if the missing data is MCAR. Therefore the missingness in our survey would be independent to all other collected survey variables. MAR is in practice unlikely to hold.

Usually the missingness is related to other variables in the data set and hence either MAR or MNAR. If we assume that the missingness in our survey is MAR we assume that we can correct for the missingness in our data set with the rest of our data Van den Beuren (?). Most of the modern missing data are based on this assumption. Further, if the missingness is MNAR the missingness mechanism has to be taken into account in the analysis. An example would be if respondents would self select into the survey based on unobserved characteristics.

4.4 Traditional Missing Data techniques

- Listewise Complete Observations This method keeps only the responses from the survey participants which completely answered the questions which are analysed. E.g. if we build an regression model for sentiment towards immigrants this method keeps only the observations, which have completed all questions included as independent or dependent variables. In general this method may be used if the missingness is MCAR. .

  However in the case of multivariate analysis for a large number of variables, even small fractions of missing values for each variable can lead to a large reduction of the sample size with this method. Therefore the authors note that even under MCAR listewise complete observations may be not efficient.

  Under certain patterns of MAR the method yields valid und efficient estimates of the regression coefficients. This does not extend to correlation. In general if the fraction of missing values is not to large, this method will not lead to much bias Schafer and Graham [2002]

- Pairwise Complete Observations This method keeps different sets of sample units for different parameters. To highlight this we look at the same example as before. Further, we refine the example by specifying that regression model studies the relation between $Y_{1,i}$ and $X_{1,i}$ and $X_{2,i}$ for all sampling units i, which where obtained by simple random sampling. Under pairwise complete observations the parameter β_1, which describes the relation between $Y_{1,i}$ and $X_{1,i}$, would be obtained using all available observations. The same holds for the parameter β_2. For each pair $Y_{1,i}, X_{1,i}$ and $Y_{1,i}, X_{2,i}$ the pairwise complete observations would be used, therefore the sets of observations would be different for each parameter. According to Schafer and and Graham (2002) the computation of the standard errors thus becomes difficult. Further, they showed a problem of this method with the example of a correlation. Eg. to calculate the correlation between $Y_{1,i}, X_{1,i}$, one may use all the available values of $X_{1,i}$ to calculate the standard deviation of $X_{1,i}$ and all available pairs of $Y_{1,i}, X_{1,i}$ to obtain the covariance. However, a correlation obtained like this may be not bound in the interval $[-1, 1]$. Therefore the authors conclude that although the principle to use as much data as available may be good, the particular implementation is not good.

- Weighting The method is used in combination with listewise complete observations. Schafer and Graham [2002] note that under specific conditions weighting can reduce the bias of the method in the case of MAR and NMAR. As in listewise complete observation the sample is reduced to the set of complete observations for a particular analysis. The sample is in the next step adjusted to resemble more closely the population or the full sample with regards to covariates. The weights are based on estimates of the probability to respond, which may be obtained from the data. The weighting eliminates the bias due to non response for the included variables in the probability model. However, the bias for any variables not included in the model will not be reduced.

- Single imputation

  - Mean substitution Imputing a missing value due to item non response with the mean value of the corresponding question yields unbiased estimates. However, we do not account for the uncertainty introduced by the missing value. The method downward biases the sample variance and overstates the number of observations. Further Schafer and Graham [2002] report that this method biases the covariance between variables and the interclasscorrelation.
  - Regression imputation This statistical technique imputes the missing values with the predicted values from a regression. The method produces unbiased estimates of the mean under MCAR and MAR however the imputed data will show less variation than the complete data Baraldi and Enders [2010]. The problem can be mitigated by adding a random component to each predicted score, which is drawn from a normal distribution with mean zero and a variance equal to the residual variance. This method yields unbiased estimates (both under MCAR and MAR) , however the standard errors do not account for the uncertainty of the estimates. The result is that the standard error is downward biased and statistical test will have a higher type 1 error.

4.5 Modern Techniques

In the following we describe **multiple imputation** based on the outline in Schafer [1999].

Multiple imputation is a simulation based technique. Each missing value is imputed with m simulated values, where m is usually a small number (3-10). The method leads to valid imputation under a frequentist perspective under the assumption of MAR and an additional technical assumption about the parameters of the missigness mechanism and the analysis model.

Little and Rubin [1986] recommend to implement MI with bayesian techniques. Therefore the simulated values are draws from the predictive posterior distribution, which is obtained under bayesian estimation.

With this technique a specific missing data model can be specified, which can be different from the analysis model. The choice of the imputation model is not completely free and must be done with attention to the analysis model. A guidance

is that the posterior distribution of the imputation model must reflect uncertainty about the analysis model.

After the imputation the data user to analyze each of the m imputed data sets using complete data techniques. Finally, the m results can be combined using Rubin's Rules Rubin [1987]. Rubin showed that an estimate can be pooled using an average of the m results. Further the estimated total variance is a combination of the average within imputation variance and the between imputation variance. Statistical tests about the estimate may be conducted using a student t-distribution with a MI specific degrees of freedom (cf. Schafer [1999]).

The flexibility of this approach is due to the seperation of the imputation and the analysis step. In general as Schafer [1999] notes that the imputation need to be reasonable and reflect uncertainty. As the analysis and imputation are separated it may be the case that the assumptions of the imputation model and the analysis model are incompatible. Schafer [1999] notes that if the imputation model makes less assumptions than the analysis model than the MI estimate is valid however a loss of power may occur. Further, in case the imputation model makes more assumptions and these assumptions are valid the MI estimate may be more efficient than the complete data analysis.

In the following we describe the use of **maximum likelihood methods for estimation with missing data** based on Enders [2006]. Under the MAR assumption maximum likelihood methods may be used to obtain valid estimates.

The estimation of missing data maximum likelihood requires the maximization of the log likelihood function, which is the log of the likelihood function. The likelihood function describes how likely a particular parameter estimate is given the data. The ML estimation yields the most likely parameter estimate minimizing the squared distance between the parameter and the data.

With missing data present the maximum likelihood estimation changes only little. The estimation uses for each individual all available data to estimate the most likely parameter given the data.

Bibliography

Amanda N Baraldi and Craig K Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37, 2010.

Craig K Enders. A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic medicine*, 68:427–436, 2006.

Morris H. Hansen and William N. Hurwitz. The problem of nonresponse in sample surveys. *The American Statistician*, 58(4):292–294, 2004. doi: 10.1198/000313004X6328.

Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-471-80254-9.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc., 1987.

J L Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15, 1999.

Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

K. P. Srinath. Multiphase sampling in nonresponse problems. *Journal of the American Statistical Association*, 66(335):583–586, 1971. doi: 10.2307/2283533.