

# Chapter 1

## Missing Observations and Data Quality Improvement

**Abstract** Missing data is a well-recognized problem which arises in statistical inferences and data analysis. We address different possible ways to handle missing data, to ameliorate its effect on the reliability and accuracy of survey-based inferences. Subsampling the non-respondents and imputation of missing values, are considered as methods for dealing with non-responses. This book presents the work developed on Ranked Set Sampling (RSS) in dealing with missing data. RSS is a relatively new sampling design. This chapter may be considered as an introduction to the rest of the oeuvre.

**Keywords** Non respondent • Imputation • Randomized responses • Simple random sampling • Ranked set sampling

*What the caterpillar calls the end of the world, the rest of the world calls a butterfly.*

Lao Tse

### 1.1 Missing Observations and Data Quality

Consider a finite population  $U$  of size  $N$  from which a simple random sample  $s$ , of size  $n$ , is drawn with replacement. Full response surveys are rare situations. In sample surveys it is common that some units are missing at the first measurement attempt. Let the characteristic under study determine a variable  $Y$ . For each  $i \in U$  we can determine the value of  $Y_i$ . When some units do not provide information we have that the sample is divided into two subsets.

$$\begin{aligned}s_r &= \{i \in U \mid \text{the response } Y_i \text{ is obtained}\}, s_m \\ &= \{i \in U \mid \text{the response } Y_i \text{ is not obtained}\}\end{aligned}$$

An estimate obtained from  $s_r$  only is biased and may be misleading.

Sampling survey practice compromises fixing a series of considerations. Before a survey can be developed many factors must be taken into account. Concepts, definitions, methods of collecting and processing data must be determined beforehand. They determine a working system, which is shaped by the aims of the survey and some key decisions, determined by the statisticians, are involved in the design of the inquiry.

It is common that data are not collected for all the units in the sample. Data can be missing for a part of the population and different problems arise when conclusions are to be taken using statistical methods. For different reasons the units may be unavailable, when they are going to be measured, or refuse giving information. Missing data is the common name for all cases in which the value of the variable of interest is not obtained.

The existence of missing values is one of the most pervasive problems in data analysis because they are present in many research activities. The seriousness of the problem depends on the pattern of the missing data, the distribution of missingness, how much is missing, and why it is missing. Missing data are widespread in social science surveying, as the interviewees are unable or unwilling to answer some questions. But it is a recurrent issue not only in sampling human populations. It is also a common problem in psychological, medical research and, recently, informatics is also dealing with it. The decision about how to handle missing data is very important as it affects the reliability and accuracy of the inferences about the population of interest. Missing data rates are a measure of the level of unit response. Frequently, surveyors use them as an indirect indicator of the quality of the data.

Missing data in survey research are present because:

1. An element in the target population  $U$  is not included on the survey's sampling frame (non-coverage);
2. A sampled element does not participate in the survey (total nonresponse);
3. A unit in the sample fails to provide acceptable responses (item or unit nonresponse).

Weighting adjustments are often used to compensate for non-coverage and total nonresponse (NR). Subsampling among the nonrespondents or imputation methods are used for dealing with unit nonresponse. A variety of methods have been developed trying to compensate for missing data. The magnitude of nonresponse (NR) bias may be partially assessed, Särndal and Lundström (2005) for a detailed discussion on this issue. Data quality often needs to subsample nonrespondents for following-up.

The existence of nonresponse in surveys induces a non-observational error reflecting an unsuccessful attempt to obtain the desired (needed) information from a selected unit. Unit nonresponse is a failure to obtain any data from a sample unit. Item nonresponse is defined when we deal with the measurement of  $k$  variables and some of them are not measured. Usually the values of  $Y$  in the nonrespondents are in general not similar to the values of it in the respondents. Hence, ignoring them is not a good decision. Many studies have attempted to determine if there is a

difference between respondents and nonrespondents. Some researchers have reported that people who respond to surveys answer questions differently than those who do not. Others have found that late responders answer differently than early responders, and that the differences may be due to the different levels of interest in the subject matter or to avoid being identified as belonging to a stigmatized group.

Generally surveyors decide to subsample the nonrespondents when the response rate is lower than expected and to interview them all is too costly. Another reason is that nonresponse constitutes an important potential source of bias. Subsampling the nonrespondents allows also studying the reasons for avoiding responding. Commonly a representative subsample of nonrespondents is taken (those units generating missing data) and it is used for inferring about them. The work of Hansen and Hurwitz (1946), pioneering the treatment of nonresponse, suggested a double sampling scheme for estimating population mean. Different authors have discussed approaches for subsampling the nonrespondents; see Srinath (1971) and Bouza (1981).

There is an extensive literature concerning missing data, much of which has focused on missing outcomes. The best way to deal with nonresponses (NR) is to prevent its happening. It determines that the surveyor must spend the needed time in designing surveys and building previsions, for dealing with nonresponse. To design experiments to reduce nonresponse is advisable. When NR are present it is advisable to use existent information to predict the missing data. Then a model is to be used to predict values for the nonresponse and imputation can be used for adjusting for item nonresponse.

Imputation means to substitute missing data with plausible values. Some practitioners consider that it solves the missing-data problem. But it must use some model. A naïve method, subjective evaluations or unsound modeling imputation methods may generate serious additional problems. The processes of imputation and analysis should be guided by common sense. If not, we will be dealing with bad estimates, false standard errors, and unreliable hypothesis tests. See Little and Rubin (2002) for a documented discussion. In some cases, good estimates can be obtained by substituting the missing observation by some supposedly close value of  $Y$  or by using some weighting estimation procedures. The usual approach in survey sampling is based in imposing a probability model on the complete data (observed and missing values). Surveyors are aware that real data are seldom described by convenient models. The theory of imputation for missing data requires that imputations be made conditional on the sampling design. It is advisable that an imputation model should produce imputation values, which are at least approximately compatible with the analyses to be performed on the imputed datasets. For example it must preserve the associations or relationships among items. For modeling we should consider that exists the Bernoulli variable.

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \dots, N$$

Then, at least, an additional source of randomness is present in imputation procedures. Another approach is to consider the set of simulation methods that have appeared in the statistical literature for imputing. These methods, known as Markov Chain Monte Carlo (MCMC), are being increasingly considered but rely on a knowledge of the phenomena under study, which is uncommon in survey sampling applications.

## 1.2 Ranked Set Sample in the Presence of Missing Data

[Chapter 2](#) is intended to provide the reader with an introduction to Ranked Set Sampling (RSS). It was introduced by McIntyre (1952) to estimate the pasture yields. Recently attention is being paid to the basic theory of RSS. The literature in the subject presents new techniques and approaches. RSS is a method of collecting data that improves estimation by utilizing the sampler's judgment or auxiliary information about the relative sizes of the sampling units. The procedure involves randomly drawing independently  $m$  sets of  $m$  units each from the population. Hence, the selection of the units evaluated takes into account the order of them in the combined  $m$  samples. The units in each set are cheaply ranked. From the first set of  $m$  units, the unit ranked lowest is measured; from the second set of  $m$  units, the unit ranked second lowest is measured and the process is continued until from the  $m$ -th set of  $m$  units the  $m$ -th ranked unit is measured. A sample of size  $n$  is obtained by repeating the procedure  $r$  ( $r \geq 1$ ) times independently for obtaining  $n = mr$ . RSS is an alternative to simple random sampling which has been shown to outperform simple random sampling (SRS) in many situations. RSS outperforms SRSWR in terms of efficiency, as it has a smaller variance in estimating, and increases the power in testing hypothesis, especially for nonparametric ones. As a result it provides the same accuracy using smaller sample sizes than the SRS alternatives. Auxiliary variables are commonly used in survey sampling. They may be derived from various sources as registers, administrative sources subjective evaluation of the interest variable etc. In RSS the sampled units are ranked using some non-costly auxiliary variable. The auxiliary variable  $X$  must be related with  $Y$ . We may also rank using judgments. We will deal with the estimation of the population mean.

The literature addressing how to deal with missing data can be divided by the need of obtaining information from the nonresponses and to diminish the amount of missing data.

In [Chap. 3](#) we will consider subsampling among the nonrespondents for dealing with missing data. The usual theory is presented in text books for simple random sampling, see (Cochran 1997), Hedayat and Sinha (1992). The use of ranked set sample is considered and models are discussed. Two problems are posed and studied at large:

1. Dealing with nonresponses in RSS.
2. Using RSS for subsampling among the nonrespondents using the information at hand in the population or provided in the first attempt for measuring  $Y$ .

The existence of nonresponse motivates to select a subsample among the nonrespondents or imputing the values of the interest variables on the nonrespondents. The use of imputation techniques for dealing with missing information is a theme of actuality. See for example Bouza and Al-Omari (Bouza and Al-Omari 2011a, b), Chang and Huang (2001), Fitzenberger et al. (2005), Rueda and González (2004), Young-Jae (2005), Singh and Deo (2003), Singh and Horn (2000), Toutenburg et al. (2008) and Zou and Feng (1998).

Chapter 4 is concerned with the use of imputation in RSS. The missing values can be identified with no-responses on a certain order statistic. Hence we have some missing observations but in general there are replicas of them if the RSS procedure is repeated  $r$  times (cycles) to have  $n = rm$  observations, for example. Different imputation procedures used in survey sampling are visited and developed for RSS. To study the properties of imputation-based estimators, are often considered through the consideration of a superpopulation model, the sampling mechanism generating the sample, the variable response mechanism and the imputation mechanism. In survey sampling practice it is advisable to use simple relations between the variable of interest  $Y$  and the auxiliary one  $X$ . In RSS as we may use  $X$  for ranking which seems to increase the accuracy. Some ratio relations are the simpler models. A study of the existent models is developed at large. Other models are also developed and discussed.

Chapter 5 is devoted to analyzing different numerical experiments planned for evaluating the efficiency of RSS-based estimators. They permit comparing SRSWR and the RSS alternatives. Some experiments are simulations using certain friendly probability distribution functions. The rest use real-life data and artificial populations are constituted. Monte Carlo experiments evaluate the behavior of the efficiency of the estimators.

## References

- Bouza, C. N. & Al-Omari, A. (2011a). Ratio imputation of missing data in ranked set sampling. accepted by Statistics, GSTA-2011-0026.r2. Nasr: Hindawi Publishing Corporation.
- Bouza, C. N., & Al-Omari, A. (2011b). Imputation methods of missing data for estimating the population mean using simple random sampling with known correlation coefficient. *Quality and Quantity*,. doi:[10.1007/S11135-011-9522-1](https://doi.org/10.1007/S11135-011-9522-1).
- Bouza, C. N. (1981). On the problem of subsample fraction in case of non response. (Spanish). *Trabalhos Estadística Investigation Operation.*, 32, 30–36.
- Chang, H. J., & Huang, K. (2001). Ratio estimation in survey sampling when some observations are missing. *International Journal of Information and Management Sciences*, 12, 1.
- Cochran, W.G. (1997). Sampling techniques. N. york: Wiley.
- Fitzenberger, B., Osikominu, A., & Völter, R. (2005). Imputation rules to improve the education variable in the IAB employment subsample. *ZEW Discussion Paper* 05–10, Mannheim.

- Hansen, M. H., & Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517–529.
- Hedayat, A. S., & Sinha, B. K. (1992). *Design and inference in finite population sampling*. New York: Wiley.
- McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385–390.
- Särndal, C. E., & Lundstrom, S. (2005). *Estimation in surveys with non-response*. New York: Wiley.
- Singh, S. & Horn, S. (2000): Compromised imputation in survey sampling. *METRIKA*, 51, 267–276.
- Singh, S., & Deo, B. (2003). Imputation by power transformation. *Statistical Papers*, 44, 555–579.
- Srinath, K. P. (1971). Multi-phase sampling in non-response problems. *Journal of the American Statistical Association*, 66, 583–589.
- Toutenburg, V., Srivastava, K., & Shalab, H. (2008). Amputation versus imputation of missing values through ratio method in sample surveys. *Statistical Papers*, 49, 237–247.
- Young-Jae, M. (2005). Monotonicity conditions and inequality imputation for sample and non-response problems. *Economic Review*, 24, 175–194.
- Zou, G. & Feng, S. (1998). Sample rotation method with missing data. Paper presented at the 4th ICSA Statistical Conference, Kunming, China.



<http://www.springer.com/978-3-642-39898-8>

Handling Missing Data in Ranked Set Sampling

Bouza-Herrera, C.N.

2013, X, 116 p., Softcover

ISBN: 978-3-642-39898-8