The Prevention of Nonresponse in Statistical Surveys

MILAN TEREK
Vysoká škola manažmentu v Trenčíne / City University of Seattle programs
Bratislava, Slovakia

Abstract: The paper deals with the possibilities of prevention of nonresponse in statistical surveys. Effects of nonresponse on the estimators are described. Then the causes of nonresponse are analyzed. Finally, the methods of increasing of response rate, based on careful choice of a mode of administration, are presented. The method of randomized response for prevention of nonresponse to sensitive questions is analysed in details and its use is illustrated on the example.

Keywords: statistical survey; nonresponse; method of randomized response

JEL Classification: C83

1 Introduction

Knowledge management can be defined as the process of creating, sharing, using and managing the knowledge and information of an organisation [4]. One method of obtaining useful information and creating the knowledge for an organisation is obtaining the information by statistical surveys and their analysis. Obtaining of data by statistical surveys is related to the problem of nonresponse.

In the past, the problem of nonresponse in statistical surveys was not so significant. Due to changes in society, there is currently a different social climate, which very often causes less willingness to provide data. It is necessary to cope with the analysis of data obtained with a higher nonresponse rate, which is common in current surveys. A high level of nonresponse can significantly impair the quality and reporting capacity of the survey results.

In general, two types of nonresponse may be considered: unit nonresponse, which lacks the values of all variables in the questionnaire. Item nonresponse means that the value of at least one but not all variables in the questionnaire is missing [8]. Both types of nonresponse reduce the accuracy of estimates, but are generally difficult to avoid. In many surveys, obtaining at least 50% of the response rate requires a lot of effort and financial resources.

Imputation means the substitution of a missing variable values by near values. It is used in the case of item nonresponse. Most often, surveys are performed by first imputing to units that have not responded partially, and then only thinking of not responding of units is being considered and weighting is being done for adjusting the nonresponse. Such an approach is called combined [8].

The influence of nonresponse to estimators and the causes of nonresponse are described. Then some methods for prevention of nonresponse are studied, and the method of randomized response to prevent the nonresponse to sensitive questions is analyzed in details.

2 The Influence of Nonresponse on Precision of Estimators

The aim of most sample surveys is to estimate, with the utmost precision, the parameters of the population, such as mean, total, or proportion. Unbiased point estimators of these parameters for different sampling designs are known (for more details see, for example, [9], [11], [12]). The main problem caused by nonresponse is the potential bias of estimators. The higher the nonresponse rate, the greater the potential bias of estimators.

We will define three different sorts of populations.

The target population[1] is the population to be studied in the survey and for which the basic inferences from the survey will be made. The target population is regarded as the ideal population to be studied.

The subset of the target population that is represented by the sampling frame is referred to as frame population. Ideally, the target and frame populations are equal. In practice, this ideal is seldom achieved. In order to select a sample from the population, one must compile a list (or frame) of all units in the target population so that an appropriate sampling can be implemented.

Finally, the respondent population is a purely hypothetical concept since it is impossible to identify all the members of this population. It is defined as that subset of the frame population that is represented by units who would respond to the survey if selected. In [2] is supposed that the frame population was divided into two strata (subsets) - the respondent stratum and the non-respondent stratum. Persons selected for the survey who respond are assumed to be randomly selected from the respondent stratum, and those that do not respond may be regarded as representing the non-respondent stratum.

Suppose that the population mean μ of variable *y* under study is estimated in the population of the size *N*. Let target and frame populations be equal.

Let:

N_R – number of units in respondent stratum,

N_{NR} – number of units in non-respondent stratum ($N_{NR} = N - N_R$),

μ_R – the mean of respondent population,

μ_{NR} – the mean of non-respondent population.

The mean of *y* in the target population is

$$\mu = \frac{N_R\mu_R + N_{NR}\mu_{NR}}{N} .$$

Suppose, the simple random sample of size *n* was realized. When the sample contains n_R of responding units and $\overline{X}$ is their sample mean, then

[1] Sometimes referred to as the inferential population.

$$E(\overline{X}) = \mu_R$$

and $B(\overline{X})$ - the bias of $\overline{X}$ is

$$B(\overline{X}) = \mu_R - \mu = \mu_R - \frac{N_R \mu_R + N_{NR} \mu_{NR}}{N} = \frac{N_{NR}}{N}(\mu_R - \mu_{NR}).$$

In general, the effect of non-responding depends on proportion of units who would non-respond to the survey if selected and on the difference between the means of units who would respond and who would non-respond. Unfortunately, the values of N_{NR}, μ_R and μ_{NR} are not usually known.

The last relation shows that the bias given by non-responding is independent of sample size *n* and cannot be reduced by its increasing. But it can be reduced by decreasing of proportion of units who would non-respond if selected ($\frac{N_{NR}}{N}$). This indicates the great importance of preventive measures to reduce the proportion of units that would not respond.

3 The Causes of Nonresponse

The quality of survey data is largely determined at the design stage. Often, when preparing a sampling plan, little time is devoted to analyzing the problem of possible nonresponse. Many less experienced, but sometimes also more experienced people, simply start collecting data without carefully considering the risks of nonresponse. They mail questionnaires to everyone in the target population and analyze those that are returned. Such surveys have frequently poor response rates. Some surveys reported in academic journals on purchasing, for example, have response rates between 10 and 15%. It is difficult to see how anything can be concluded about the population in such a survey ([7], p. 332). An analyst with good knowledge of the population should be able to anticipate the causes of nonresponse and implement effective prevention. Most analysts, however, do not know as much about the reasons for nonresponse as they think they do. Design of experiments and applying quality improvement methods to data collection and processing can be used to identify the causes of nonresponse.

The causes of nonresponse can be categorized as follows ([7], p. 333):

- Survey Content
- Methods of data collection
- Respondents characteristics

Among the three mentioned categories, mainly the methods of data collection can be effectively influenced. We will take a closer look at some of the options for increasing the response rate associated with data collection methods. Then we will study in details how to increase the response rate to sensitive questions by the method of randomized response.

4 The Prevention of Nonresponse

The following are some factors that may influence the response rate and data accuracy.

- ***Survey content***. Many surveys involve questions that persons might view as sensitive. Some persons may protect their personal information by refusing to respond to the survey or to some questions, while others may give inaccurate answers. The general advice in questionnaire design is to avoid sensitive questions if possible.

- ***Time of survey***. Some calling periods or seasons of the year may yield higher response rates than others. The vacation month of July and August, for example, would be a bad time to take a one-time household survey in Slovakia.

- ***Data-collection method***. Generally, telephone and mail surveys have a lower response rate than in-person surveys (they also have lower costs, however). E-mail and Internet surveys often have also low response rates.

- ***Questionnaire design***. The questionnaire design has a large effect on the response rate; it can also affect whether a person responds to an item on the questionnaire (for more details see, for example, [1], pp. 43 – 63, [7], pp. 11 – 16, [13], pp. 9 – 18).

- ***Respondent burden.*** Persons who respond to a survey are doing you an immense favor, and the survey should be as nonintrusive as possible. A shorter questionnaire, requiring fewer details, may reduce the burden for the respondent.

- ***Survey introduction***. The survey introduction provides the first contact between the interviewer and a potential respondent; a good introduction, giving the recipient motivation to respond, can increase response rates dramatically.

- ***Follow-up.*** The initial contact of the sample is usually less costly per unit than follow ups of the non-respondents. If the initial survey is by mail, a reminder may increase the response rate. Not everyone responds to follow-up calls, though; some persons will refuse to respond to the survey no matter how often they are contacted.

4.1 The Prevention of Nonresponse to Sensitive Questions

Sometimes inclusion of sensitive questions such as “Did you understate your income on your tax return?” or “Do you use cocaine?” is needed. These are questions that “yes” respondents could be expected to lie about. If the survey contains some sensitive questions or it is directly focused on a sensitive topic, such as financial matters or drug use, sometimes the response rate can be increased by careful choice of the mode of administration.

Many studies report that higher percentages of people say they have used illegal drugs when they fill out the questionnaire themselves than when the questionnaire is administered by an interviewer. Some surveys on sensitive topics use computer-assisted self-administration, where the respondent types answers directly into the computer. The questions are displayed on-screen and are also played through recording. An interviewer may be in the room to answer questions, but the interviewer does not see the responses typed into the computer. ([7], pp. 540 - 541).

Another interesting possibility to prevent the nonresponse to sensitive questions is applying the method of randomized response. In [7], p. 541 is described the method suggesting inclusion of a pair of questions: the sensitive one and an innocuous one. A randomizing device (such as a coin flip) determines which question the respondent should answer. If a coin flip is used as the randomizing device, the respondent might be instructed to answer the question "Did you use cocaine in the past week?" if the coin is heads, and "Is the second hand on your watch between 0 and 30?" if the coin is tails. The interviewer does not know whether the coin was heads or tails, and hence does not know which question is being answered. It is hoped that the knowledge that the interviewer does not know which question is being answered will encourage respondents to tell the truth if they have used cocaine in the past week. The randomizing device can be anything, but it must have known probability p that the person is asked the sensitive question and probability $(1 - p)$ that the person is asked the innocuous question. The key to randomized response is that the probability that the person responds yes to the innocuous question, p_I is known. We want to estimate p_s, the proportion responding yes to the sensitive question. If everyone answers the questions truthfully, then the proportion of "yes" respondents in the population is

$$\pi = P(respondent\ replies\ yes) =$$
$$= P(yes|asked\ sensitive\ question)P(asked\ sensitive\ question) +$$
$$+ P(yes|asked\ innocuous\ question)P(asked\ innocuous\ question) =$$
$$= p_s p + p_I(1-p)$$

Let $\hat{\pi}$ be the estimated proportion of "yesses" from the sample. Then p_s can be estimated by

$$\hat{p}_s = \frac{\hat{\pi} - (1-p)p_I}{p} \tag{1}$$

and the estimated variance of $\hat{p}_s$ is

$$\widehat{D}(\hat{p}_s) = \frac{\widehat{D}(\hat{\pi})}{P^2} \tag{2}$$

The larger p is, the smaller the variance of $\hat{p}_s$. But if p is too large, respondents may think that the interviewer will know which question is being answered.

Example. The n = 300 students of a university with N = 4000 students was selected by random sampling without replacement. Each selected student should fill a questionnaire containing also the following pair of questions:

Question 1: Have you ever cheated on an exam?

Question 2: Were you born in February?

Each respondent is instructed to flip a coin. If the coin is heads, he should answer question 1; if not, he should answer question 2.

We know from birth records that p_I = 0.083, and P = 0.5. Of the 300 respondents, 61 say yes to whichever question the coin indicated they should answer. Then $\hat{\pi}$ = 61/300 = 0.2033. Because

$n/N = 0.075$, the finite population correction factor cannot be neglected and the variance of $\hat{\pi}$ can be estimated by (see [13], p. 136).

$$\hat{D}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{n-1} . \frac{N-n}{N} = 0.0005.$$

We can estimate p_s by (1)

$$\hat{p}_s = \frac{\hat{\pi} - (1-p)p_l}{p} = \frac{0.2033 - (1-0.5)0.083}{0.5} = 0.3236$$

and variance of $\hat{p}_s$ by (2)

$$\hat{D}(\hat{p}_s) = \frac{0.0005}{0.5^2} = 0.002.$$

The proportion of students who have ever cheated on an exam can be estimated as 32.36 %. The variance of the used estimator is estimated to be 0.002.

Before using randomized response methods in your survey, you should test the method to see if it does indeed increase compliance and reduce bias. There is good experience with applying this method in some surveys. For example, Danermark and Swensson in [3] found that randomized response methods worked well for estimating drug use in schools and appeared to reduce response bias. Duffy and Waterton in [5], however, concluded that randomized response methods were not helpful in their survey to estimate incidence of various alcohol-related problems in Edinburgh, Scotland. Randomized response, however, increased the complexity of the interviews and some interviewers reported that many persons were confused by the method ([7], p. 542).

5 Conclusions

In general, every effort should be made to get answers from all respondents. Of the three mentioned causes of nonresponse, data collection methods can be particularly influenced to increase the response rate. Some of such methods were described in more detail.

A very interesting method is the method of randomized response, which can decrease the nonresponse to sensitive questions. However, use of this method is not necessarily effective in all surveys containing sensitive questions. Testing of the method is recommended before its use in a specific survey.

It is useful to try to obtain at least some information about non-respondents that can be used later to adjust for the nonresponse and include surrogate items that can be used for item nonresponse. There is no complete compensation for not having the data, but partial information may be better than none. Information about the sex or age of a non-respondent may be used later to adjust for nonresponse. Questions about income may well lead to refusals, but questions about cars, employment, or education may be answered and can be used to predict income (for more details see [7], p. 336).

If the nonresponse rate is not negligible, inference based only upon the respondents may be seriously flawed. In the case of item nonresponse, the methods of imputation can be used. A replacement value, often from another person in the survey who is similar to the item non-respondent on other variables, is imputed for the missing value (for more details see [6], pp. 408 – 418 and [7], pp. 346 – 350). In the case of unit nonresponse, the weighting methods for nonresponse are of interest (for more details see, for example [7], p. 340 – 345, [6], pp. 489 – 513, and [10]). Weights can be used to adjust for nonresponse.

Acknowledgments

The paper was supported by the grant from Grant Agency of VEGA no. 1/0393/16 entitled „European Union in Post Crisis Period – Macro and Microeconomic Aspects“.

References

1. BeTHLEHEM, J., 2009. *Applied Survey Methods. A Statistical Perspective.* Hoboken: Wiley and Sons.
2. COCHRAN, W. G., 1977. *Sampling Techniques. Third Edition.* New York: J. Wiley and Sons.
3. DANERMARK, B., and SWENSSON, B., 1987. Measuring drug use among Swedish adolescents: Randomized response versus anonymous questionnaires. *Journal of Official Statistics*, 3, pp. 439–448.
4. GIRARD, J. P., and GIRARD, J. L., 2015. "Defining knowledge management: Toward an applied compendium" (PDF). *Online Journal of Applied Knowledge Management*. 3 (1): 14.
5. DUFFY, J. C., and WATERTON, J. J., 1988. Randomized response vs. direct questioning: Estimating the prevalence of alcohol related problems in a field survey. *The Australian Journal of Statistics*, *30*, pp. 1–14.
6. LEVY, P. S., and LEMESHOW, S., 2008. *Sampling of Populations. Methods and Applications. Fourth Edition*. Hoboken: J. Wiley and Sons.
7. LOHR, S. L., 2010. *Sampling: Design and Analysis. Second Edition*. Boston: Brooks/Cole.
8. SÄRNDAL, C.-E., and LUNDSTRÖM, S., 2005. *Estimation in Surveys with Nonresponse.* Hoboken: J. Wiley and Sons.
9. TEREK, M., and HRNČIAROVÁ, Ľ., 2008. *Výberové skúmanie.* Bratislava: Ekonóm.
10. TEREK, M., 2014. Možnosti riešenia problému neodpovedania v štatistických prieskumoch. *Ekonomické rozhľady* 2/2014.
11. TEREK, M., 2017. *Interpretácia štatistiky a dát. 5. doplnené vydanie.* Košice:Equilibria.
12. TEREK, M., 2017. *Interpretácia štatistiky a dát. Podporný učebný materiál. 5. doplnené vydanie.* Košice: Equilibria.

13. TEREK, M., 2019. *Dotazníkové prieskumy a analýzy získaných dát. 1. vydanie.* Košice: Equilibria.

Contact data:

Prof. Ing. Milan Terek, PhD.
Vysoká škola manažmentu v Trenčíne / City University of Seattle programs
Panónska cesta 17, 851 04 Bratislava, Slovakia
mterek@vsm.sk