CBN Journal of Applied Statistics (JAS)

Volume 1 | Number 1 Article 7

12-2010

Addressing the Problem of Non-response and Response Bias

Fabian C. Okafor
University of Nigeria, Nsukka, Enugu

Addressing the Problem of Non-response and Response Bias

Fabian C. Okafor[1]

INTRODUCTION

Most of the statistics needed for national planning are derived from large scale sample surveys with households as reporting units both in the developed and developing countries. Examples of these surveys are Income and Expenditure, Employment, Food consumption and Nutrition, Agriculture, Health, Education, Establishment, etc. The required information on these topics is obtained from the selected households or firms in an Establishment survey. The households to be canvassed are usually selected by probability sampling. As it happens in all surveys some selected units may not be contacted, or fail to respond to the interview or when they do provide distorted or inaccurate responses. These introduce some element of bias in the estimates.

Survey planners and analysts in Nigeria have devoted much more attention to sampling errors at the expense of non-sampling errors (non-response and response errors). Sampling error is the degree to which the sample estimate differs from the average value of the characteristic due to chance. The present discussion will be centered on non-sampling error, which may present serious deficiencies in the statistics and render the survey useless. According to Platek and Gray (1986), "Non-response has been generally recognized as important measure of the quality of data since it affects the estimates by introducing a possible bias in the estimates and an increase in sampling variance because of reduced sample." They continued by saying that "in a practical way, the size of non-response may indicate the operational problems and provide an insight into the reliability of survey data." There is need therefore to study the nature and effect of non-response in the surveys conducted in the country.

NON-SAMPLING ERROR

Non-sampling error is due to other causes apart from chance factor, i.e. inductive process of inference from a probability sample. This type of error is found both in sample survey and complete enumeration. Non-sampling error can occur anywhere from the planning to analysis stage of the survey. At the planning stage, we have such factors like ambiguous definitions and concepts in designing a questionnaire, length of a questionnaire, omission or duplication of some units due to use of obsolete frame. At executive stage we have such factors as inability to locate some of the units of enquiry probably due to civil disturbances, flood disaster, security problem, etc; memory lapse on the part of the respondent; inadequate training of enumerators, etc. The factors at the analysis stage include carelessness in editing of the completed questionnaire, tabulation and printing of final results among others. Non-sampling error can be classified into non-response and response errors.

Non-response Error

We define non-response as failure to collect data from a sample unit in the target population. It may occur because of refusal of some units in the sample to return the completed questionnaire or to grant interview in the case of face-to-face interview or through non-contact. The first kind is called unit (total) non-response. Non-response also occurs when a unit provides information to some but not all questions in the questionnaire. This is called item non-response. Item non-response may be as a result of irrelevant or sensitive questions in the questionnaire; question not understood or through fatigue or lack of knowledge.

The size of non-response is an indication of how reliable the survey data are.

[1] Dept of Statistics, University of Nigeria, Nsukka; e-mail: fc_okafor@yahoo.com

Response Error

Response error is the difference between the true value of a characteristic and the actual value supplied by the respondent or recorded by the interviewer. Response error occurs as a result of faulty instrument, the respondent or the interviewer, or as a result of interplay of the three.

More details on response error can be found in Moser and Kalton (1979, P 378).

PROBLEM OF NON-RESPONSE AND RESPONSE ERRORS

Non-response has effect on the survey estimates because of potential bias introduced in the estimates. However, "the response bias may not be nearly as serious relative to the sampling error for small samples as it is for large samples" (Platek and Gray, 1986). Let ϕ be the population proportion of the respondents and 1-ϕ that of the non-respondents. Suppose the interest is in the estimation of the population mean, μ. If only the respondents' values are used to estimate the mean, the estimate will be $\square_{1.}$ Let the average value of this estimate be μ_1; hence the bias introduced due to non-response is given by

$$\mu_1 - \mu = \mu_1 - \{\phi\mu_1 + (1-\phi)\mu_2\} = (1-\phi)(\mu_1 - \mu_2) \quad (1)$$

This is the measure of non-response bias.

Let y_i be the survey value and x_i the true value of the variable of interest; then $y_i = x_i + e_i$, and $y_i - x_i = e_i$ is the individual response error. The average of all individual response errors given by $\overline{Y} - \overline{X} = \overline{e}$ is called the response bias.

As we have seen, non-response and response errors introduce bias in the estimates and create distortion in the survey results. Furthermore, non-response rate is an essential factor in assessing the reliability of survey results. When the non-response rate is high and vary from one area to the other, the bias may still remain even after adjusting for non-response. Therefore, it is very important to include the response rate in every survey publication. Actually, estimates from the analysis of only the respondents' records do not apply to the target population, but are representative of the population of individuals who would respond to the survey, especially when the characteristics of the respondents differ much from those of the non-respondents.

DETECTION AND CONTROL OF REPONSE ERROR

There is yet no satisfactory method of ascertaining the size of response errors in the individual sample survey or census. But what is whispered among those who have dealt with surveys in Nigeria is that the greatest cause of response error is deliberate falsification and fabrication of survey data by the respondents and interviewers respectively. It is not an easy task to ascertain whether a respondent has given a true answer or not during interview. The usual practice is to use some checks which mainly determine the discrepancy in responses. These checks include the following.

- Record Checks: We can detect errors in response through the use of records or documents, e.g. birth records, marriage directories. In Nigeria we have not fully imbibed the culture of record keeping and official records are not readily available because of official secrecy and unnecessary bottleneck. Even if the individual records are available, it will not be easy to match sample respondents with official records. In some cases some names may not be available in the records. Moreover, records may not indicate the true value unless one is sure that the information in the record is more accurate than the answers supplied by the respondent during the survey.
- Consistency Checks: Consistency checks are normally inbuilt in the questionnaire. In consistency check, information sought from the individual in a particular issue is obtained by asking the relevant question in two or more different ways located at different points in the questionnaire. This will serve as a check on

the quality of data collected. This notwithstanding, respondents determined to give false answers may still do so.

- Re-interviewing: This is also known as quality check or post enumeration survey in census. It involves re-interviewing a sub-sample of individuals in order to ascertain the extent the first response is consistent with that of the second interview. In re-interview one could repeat the same survey questions or use more detailed questions and better trained and experienced interviewers.

For the control of response error, the procedure to be adopted must be carefully considered before the survey. One of the methods available to the sample survey expert for reducing response error is the careful selection, training and supervision of interviewers. Another is the use of well designed questionnaire with precise and clear instructions.

In Nigeria, most survey bodies pay more attention to sample size and sampling error neglecting non-sampling error, which in most cases may be serious as to distort the estimates from the survey as was earlier pointed out.

ASSESSING THE EXTENT OF NON-RESPONSE BIAS

The question is how do we assess or determine the extent of non-response bias? I have not come across any research carried out in Nigeria to assess the extent and nature of response and non-response bias. We rely on results from other countries, which should not the case. We have to bear in mind that decreasing non-response rate may not necessarily always reduce the non-response bias; but low response rate may yield statistics with large non-response bias. The methods for assessing the extent on non-response bias are found in http://www2.chass.ncsu.edu/garson/pa765/sampling.htm and the UN Handbook of Household Surveys, Part One and summarized below. These are:

- **Population comparison:** Here survey averages are compared with those from other sources like recent population census. This method is useful in comparing variables like gender, age distribution, race, income and occupation. The idea is to identify variables where the sample mean differs from the population mean. The setback in this method is that the deviation may be due also to the sample units selected and canvassed in addition to non-response bias.
- **Intensive post-sampling:** A sample of non-respondents is selected and intensive effort is made to obtain interview on this sub-sample. The difference between the estimates from the respondents and the sub-sample of non-respondents is a measure of the non-response bias, i.e.

$$\frac{n_r}{n}(\bar{y}_r - \bar{y}_{nr}) \qquad (2)$$

- **Matching:** A sub-sample of the respondent and non-respondent sample cases is matched with the current population census. An indication of the non-response bias is given by the difference between census data for respondent and non-respondent sample cases.
- **Call backs:** A comparison of the survey results for sample units interviewed after considerable call backs (e.g. after the third visit) with those interviewed at the first attempt provides some measure of non-response bias.
- **Adjustment:** Another method of assessing non-response bias is to compare the unadjusted estimates based on the respondents with those estimates obtained after adjusting for non-respondents (Ekholm and Laaksonen, 1991).

HANDLING NON-RESPONSE

Non Theoretical Approach

Some of these approaches discussed here came out at the interactive session during the workshop on non-response in surveys organized by Central Bank of Nigeria in September 2010. One of such approaches is developing a rapport with the community or the respondents through social engagements, like attending naming ceremonies, traditional marriages, funeral condolences and sending season's greetings, etc. For example, it was reported of a case of an interviewer who obtained cooperation by merely giving attention to a child in the family; asking about his welfare. Response rate could be increased through advocacy using radio jingles, print media and television to sensitize people

about an oncoming survey. Also community heads or chiefs when contacted and educated on the nature and benefit of a survey can help to get cooperation and thereby improve on the response rate.

The manner of dressing determines whether one can obtain interview or not. One who is going to interview a company director must be if possible dress like one and come into the premises with a good car. In short and interviewer should dress well if possible better than the respondents. There was a case where a female interviewer went to a community in a pair of trousers; she never succeeded in obtaining interview from any one. But when she dressed like a typical traditional woman with a wrapper and a matching blouse the situation changed. The view of the community was that she came to corrupt their girls with her type of dressing.

Another way of increasing cooperation is through the use of incentives either material or financial. People are of the opinion hat incentives could be given but only if the respondent fails to respond after two or three visits. Incentives may be in the form of company calendar, diary or reports of past surveys, especially in an establishment survey. Corporate award may be given to an establishment that cooperates regularly and effectively in surveys. Financial incentive may help in some cases but may be counter productive. There is a case of a gateman who refused an interviewer entry to an establishment; but immediately money exchanged hands the gate flew wide open. However, when people get used to financial gift, they fail to respond unless some amount of money is given. This happened in a community where the interviewer who used to give monetary incentive was changed with someone who does not give. At the next round of survey the interviewer met brick wall; response rate dropped drastically.

Finally, use of locals with knowledge of the terrain and good command of the local language and with necessary academic qualification will definitely improve cooperation on the part of the respondents. For lack of space and time we shall not mention all other methods of increasing the response rate.

Theoretical Approaches

There are several methods of handling non-response depending on whether it is total or item non-response. Total non-response is handled by using weighting adjustments whereas item non-response is taken care of by imputation.

Weighting Adjustment

"The aim of any weighting adjustment is to increase the weights of the respondents so as to compensate for the non-respondents" (Okafor, 2002, p358). Proponents of weighting adjustment argue that it is useful for adjusting for biases in the sample. Those against are of the opinion that it makes little or no difference to conclusions. According to Bourque and Clark (1992, p60) "the use of weights does not substantially change estimates of the sample mean unless non-respondents are appreciably different from respondents and there is a substantial proportion of non-respondents." We shall now summarize the methods used for weighting adjustments. For more details the reader is referred to Lohr (1999, p266) and Kalton and Kasprzyk (1986).

1. **Weighting adjustments overall:** Let π_i be the probability of selecting unit *i* in the sample. The sample weight (raising factor) is $w_i = \frac{1}{\pi_i}$, $i = 1,2,\ldots, n$. Now the estimate of response probability is $\varphi = \sum_{i=1}^{n} \alpha_i w_i \Big/ \sum_{i=1}^{n} w_i$; α_i takes the value one if the i^{th} unit respond and zero otherwise. The new weight for the respondents will be $w_i^* = \frac{1}{\pi_i \varphi}$ and the estimator of the population total of the variable of interest *y* will be $\hat{t} = \sum_{i=1}^{n} \alpha_i w_i^* y_i$. (3)
2. **Weighting- class adjustment:** The sample units are grouped into classes using auxiliary variable available for all sample units like sex, race, and geographical area. Each class response probability is estimated and used to adjust for class non-response.
3. **Population weighting adjustment:** The sample respondents are classified into strata using known auxiliary variable. Each stratum has a known population distribution (proportion), W_h from past census result, which is

used as weight for each respondent. The advantage of the population weighting adjustment is that it reduces the non-response bias as well as non-coverage error.

4. **Sample weighting adjustment:** The total sample units are grouped into strata using information available for both respondents and non-respondents in the sample. The sample distribution (proportion) $w_h = n_h / n$ is used as weight in each stratum. For simple random sampling the estimator of the mean is

$$\bar{y}^* = \sum_{h=1}^{H} w_h \frac{1}{n_{rh}} \sum_{i=1}^{n} \alpha_i y_i \qquad (4)$$

5. **Post-stratification using weights:** As described by Lohr (1999, p268), this method uses the ratio of N_h (population units in stratum h) and its estimator $\sum_{j=1}^{n_h} \alpha_j w_{jh}$. With this ratio, the modified weight is given by $w_i^* = \frac{w_i N_h}{\sum_{j=1}^{n_h} \alpha_j w_{jh}}$.

The assumptions in weighting adjustments are that the respondents and non-respondents are similar and each unit is equally likely to participate in the survey. These assumptions are never always true in practice. Hence, weighting may not completely eliminate all non-response bias.

Imputation

Imputation entails assigning values to missing responses making use of available auxiliary information on the sample units. It is used in handling item non-response in order to reduce non-response bias among other reasons. Imputation method has advantages and disadvantages; for details see Kalton and Kaspryzk (1986). Some of the imputation methods available in the literature include:

- Mean Imputation: This involves imputing to the missing response the mean of the respondents for the particular item.
- Random Imputation: A respondent is randomly chosen from among all the respondents; the value of the response for this individual is assigned to the missing response for a given item. In order to preserve the multivariate nature of the data, values from the same donor are used for all missing values of a non-respondent. Sometimes more than one value is imputed for every missing item by carrying out *k* greater or equal to two imputations. This is called multiple imputation proposed in Rubin (1978) and illustrated in Rubin (1986) and Iwebo (2008). The multiple imputations help one to estimate the additional variance due to imputation.
- Hot-deck Imputation: In hot-deck imputation the value of one of the respondents is assigned to the missing response. The problem with this is that the value of one donor can be used for many missing values in a particular item. Versions of hot-deck imputation are sequential and nearest-neighbour hot-deck imputations. In nearest-neighbour hot-deck imputation a distance function (for the auxiliary variables) is used to impute a value of a respondent to the non-respondent. Chen and Shao (2000) explained this for a bivariate sample values. Let all *x*-values be observed for all sample units. Suppose $y_1,...,y_r$ are observed values for the respondents and the remaining *n-r* values are missing. Then the nearest-neighbour hot-deck imputes a missing y_j by y_i, where $1 \le i \le r$ and i is the nearest neighbour of j measured by satisfying the condition $|x_i - x_j| = \min_{1 \le i \le r} |x_i - x_j|$.
- Regression Imputation: Logistic regression forms part of this technique. Using the data from the respondents, the values of the variable to be imputed are regressed on some auxiliary variables available for all the sample units. The regression equation so obtained is used to predict for the missing item responses. Residuals could be added to the predicted values, if desired, to introduce randomness. Logistic

regression is used in imputing for dichotomous variable, that takes the value 1 or 0, based on the logistic regression function given by $\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$.

Its logit is $\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$. The estimates of α and β, and known x value are used to obtain the estimated probability $\hat{\pi}(x)$. A random number is drawn from a uniform distribution [0,1], say τ. If $\hat{\pi}(x_i) > \tau$, then y_i takes the value 1, and 0 otherwise.

NON-RESPONSE BIAS AND NON-RESPONSE RATE

"The relationship between the bias and the size of non-response while perhaps more important is less obvious since it depends on both the magnitude of non-response and the differences in the characteristics between respondents and non-respondents" (Platek and Gray, 1986). In fact there is no clear-cut simple relationship between non-response rate and non-response bias. However, in survey involving face-to-face interview a response rate of between 70 – 85% is regarded as acceptable. But below 70% there is a chance of non-response bias in the estimates. According to Groves (2006), the non-response bias is a function of how correlated the survey variable is to the response propensity (likelihood, probability of being a respondent). It has also been argued that decreasing non-response rates may not necessarily lead to decrease in non-response bias; and that the non-response rate is not a very good measure of the size of the non-response bias. In a review of literature on the study of non-response rate and non-response bias, Groves (2006) concluded that, for a particular survey "if the non-response rate were reduced by methods more attractive to the higher-income persons, then the relative non-response bias might decrease dramatically. If the non-response rate were reduced by methods equally attractive to higher-and lower-income persons, then the bias might be reduced as a simple function of how the non-response rate declined. However, if the non-response rate were reduced by methods more attractive to the lower-income persons, then the non-response bias might actually increase".

There is yet no conclusive study on the linkage between non-response rate and non-response bias. It depends on the variables of interest and statistics obtained and the response propensity of an individual. Non-response rate, however, is indirectly related to non-response bias in some estimates. "In short, non-response bias is a phenomenon much more complex than mere non-response rates" Groves (2006). Finally, in conclusion, "a rational approach to the problem of controlling non-sampling errors will, therefore be to try to reduce them as much as possible to levels at which the results will be usable for the purpose in view, but not to such extent as will render the efforts and costs to become incommensurate with the improvements achieved" (Murthy 1967, p467).

REFERENCES

Bourgue, L.B. and Clark, V.A. (1992): Processing Data: The Survey Example. *Thousand Oaks CA Sage Publications.*

Chen, J. and Shao, J. (2000): Nearest-Neighbour Imputation for Survey Data. *Journal of Official Statistics 16, 583 - 599.*

Ekholm, A. and Laaksonen, S. (1991): Weighting via Response Modeling in the Finnish Household Budget Survey. *Journal of Official Statistics 7: 1325-37*

Groves, R.M. (2006): Non-response Rates and Non-response Bias in Household Surveys. *Public Opinion Quarterly. 70 (5) 646 – 675.*

Iwebo, E. J. (2008): Imputing for Missing Values in the Estimation of the Population Mean of Post University Matriculation Score. *Unpublished MSc project, Dept of Statistics, University of Nigeria, Nsukka.*

Kalton, G and Kasprzyk, D. (1982): Imputing for Missing Survey Responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association, 22 – 31*

---------- (1986): The Treatment of Missing Survey Data. *Survey Methodology, 12, 1, 1 – 16.*

Lohr, S.L. (1999): Sampling: Design and Analysis. *Duxbery Press, Pacific Grove, CA, USA.*

Moser, C.A. and Kalton, G. (1979): Survey Methods in Social Investigation. 2nd Edn. Heinemann Educational Books, London

Murthy M.N. (1967): Sampling Theory and Methods. *Statistical Publishing Society, Calcutta, India.*

Okafor, F.C. (2002): Sample Survey Theory with Applications. *Afor-Orbis Publications Ltd, Nsukka, Nigeria*

Platek, R. and Gray, G.B. (1986): On the Definitions of Response Rate. *Survey Methodology, 12, 1, 17 – 27.*

Rubin, D.B. (1978): Multiple Imputations in Sample Surveys – a Phenomenological Bayesian Approach to Non-response. *Proceedings of the Survey Research Methods Section of the American Statistical Asso. 20 -34.*

…….(1986): Basic Ideas of Multiple Imputation for Non-response. *Survey Methodology, 12, 1, 37 – 47.*

Sampling: *http://www2.chass.ncsu.edu/garson/pa765/sampling.htm*