The Online Statistical Survey As An Alternative To Traditional Methods

SIMONA DINU
Department of Fundamental Sciences and Humanities
Constanta Maritime University
Mircea cel Batran Street, no.104, Constanta
Postal Code 900663, Romania,
sedinu@yahoo.com

MADALINA NITOIU
Department of Statistics and Econometrics (Associate Professor)
Academy of Economic Studies
Calea Dorobantilor 15-17, District 1, Bucharest,
Postal Code 010572, Romania
madalina_nitoiu@yahoo.com

CATALIN POMAZAN
Department of Engineering Sciences in the Electrical Field
Constanta Maritime University
Mircea cel Batran Street, no.104, Constanta
Postal Code 900663, Romania,
catalinpomazan@gmail.com

Abstract: Some of the most important innovations were the Internet and Web technologies with a great impact on society. Statistics, as a social science, has only benefits from it. Lately, the online sampling technique has greatly expanded. Each website that has a certain importance includes questionnaires in various forms. These range from a simple question to complex questions and are part of everyday life for those who have Internet access. The main question is how feasible are the results derived from these samples. The main problem is the representativeness because a nonreprezentative sampling is useless. This paper aims to analyze the online survey methodologies and their effectiveness by presenting their advantages and disadvantages.

Key-Words: Internet, online survey, questionnaire, standard error.

1 Introduction

It is known that the advent of the Internet has disturbed the relationships between the economic agents and has changed the concept of communication. Due to the increasingly coverage area and the huge volume of information involved, the traditional research methods can be adapted to be used in the electronic environment.

The Internet offers unique methodological opportunities that can be viewed from two perspectives:

- *Internet as a source of documentation: o*rganized as a network, it allows access in a short time to a wealth of information;

- *Internet as a medium for data collection:* in this case it appears as an alternative to the traditional methods of data collection: by total or by selective observation.

In the modern society, the acute need of a large amount of current information makes the administrative sources and studies carried on exhaustive research to be insufficient. Since the use of surveys allows obtaining briefly complex information, this method has penetrated almost all fields of human activity.

The major technological leap and the expansion of the Internet as a medium of communication or as a mean of doing business have caused the adaptation of classical methods for data

collection. Thus, the online surveys have become increasingly used.

Similar to traditional surveys, the respondents should be assured that the data they provide to complete a questionnaire shall be confidential. In an online survey, the danger of data interception rises, so the organizers of such a survey should look more carefully at protecting data after it is collected. It is recommended to encrypt data and also the existence of a firewall for the server that keeps them in memory.

2 LITERATURE REVIEW

Online surveys have many strengths and weaknesses. These are extensively described in the literature (e.g. Fricker and Schonlau [2], Malhotra [4], McDaniel and Gates [5], Tingling et al. [8], Wilson and Laskey [9]). Thus, the online survey is quick, inexpensive and attractive, by collecting large amounts of data at a very low price. Therefore, more and more questionnaires are organized like this.

However, the question is whether an online survey is equally attractive in terms of quality, because there are methodological problems. These problems are caused by the use of the Internet as a tool for selection of respondents [6].

Always, the goal of a survey is to gather well-defined information on target population. For this purpose a sample is selected from this population. Sampling methodology was developed over a period of more than one hundred years. It is based on the fundamental principles of probability theory in sampling.

As a result, the precision of the estimators can be quantified and controlled. The probabilistic sampling principle has been applied successfully in the Academic Journal of Statistics of 1940 and expanded in consumer market research.

3 THEORETICAL FRAMEWORK

At first glance, online surveys have much in common with other types of surveys. These represent just another way of collecting data. Questions are not required face to face or by phone, but on the Internet. Samples are not built through the probabilistic sampling, but they rely on the self-selection of respondents. This can have a major impact on the survey results.

Also, there is another methodological problem of these web surveys, namely that they are not based on probabilistic samples, and do not cover the whole area concerned. The coverage area could be a serious problem for web surveys. If the target population consists of all people with an Internet connection, there is no problem. However, usually the target population is larger than that on the Internet, because there are many people who do not have Internet access.

3.1 A random sample of the Internet population

One argument is that the above mentioned issue could disappear with the increase of people connected to the Internet. However, it is not obvious. Bethlehem [1] shows that the influence owed to estimator coverage area for the average of the population "y" of variable "Y" is equal to:

$$B(\bar{y}_I) = E(\bar{y}_I) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) \quad (1)$$

The estimator $\bar{y}_I$ is the average of the survey based on the observations collected from the respondent population on the Internet. $\bar{Y}_I$ and $\bar{Y}_{NI}$ are the averages of the population on the Internet and on the non-Internet. In addition, N is the total population size and N_{NI} is the non-Internet population size.

The magnitude of this influence is determined by two factors. The first factor is the relative importance of total population size against the non-Internet population, N_{NI}/N. The influence is decreasing, because a small part of the population hasn't Internet access. The second factor is the contrast $\bar{Y}_I - \bar{Y}_{NI}$ between the Internet population and the non-internet population. Moreover, the averages of target variables differ for these two subpopulations. An increase of the Internet population coverage will reduce the slope for the factor N_{NI}/N. However, the contrast is not necessarily decreasing due to the increase of the Internet coverage area. Thus, taking into account the combined effect of these two factors, there is no guarantee that the increase of the Internet coverage area will reduce the uncovered area.

3.2 The auto-selection of the Internet population

Many web surveys are not based on the sampling probability. The questionnaire is simply posted on the Internet. The respondents are those people who happened to be at that time on the Internet and visit the website that posts the survey. They decide whether or not to participate in the study. These surveys are called self-managed

surveys. The problem is that the researcher does not control the selection process. The selection probabilities are unknown and, moreover, they are considerably lower than the sampling probability within the traditional surveys. Therefore, estimates can not be determined with precision.

The participation in a self-managed Web survey requires respondents first to acknowledge the existence of the survey (they may accidentally visit that website or watch various banners or mail messages). Then they may decide to complete the questionnaire on the Internet. All of these mean that each element k of the population does not know the probability of survey participants, for *k = 1,2,...,N*.

The responding elements may designate the series of those N indicators r, r_1, ..., r_N, assuming that there are N participants. The expected values $\rho_k = E(r_k)$ will be called the tendency of responses for the element k.

The random variables r_1, r_2, ..., r_N are independent. The sample selection process has a Poisson sampling form. However, in practical applications the selection probability of a Poisson process is unknown. The number of respondents is equal to:

$$n = \sum_{k=1}^{N} r_k \tag{2}$$

The mean of the estimated population is:

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{N} r_k \cdot Y_k \tag{3}$$

This estimator implicitly assumes that each element of the population has the same probability of participation in the survey. It can be shown that the expected value is approximately equal to the ratio of the estimated values of random variables. Therefore:

$$E(\bar{y}) \approx \bar{Y}_I^* = \frac{1}{N \cdot \bar{\rho}}\sum_{k=1}^{N} \rho_k \cdot I_k \cdot Y_k \tag{4}$$

where $\bar{\rho}$ is the average of all of probable responses.

Usually, the estimated value of the sample mean is not equal to the population mean. This happens only if the slope disappears and all the answers of the Internet population are equal. In this case, the auto-selection results in a representative sample because all elements have the same selection probability.

Bethlehem [1] showed that the mean of the estimated population can be written as:

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{C_{\rho Y}}{\bar{\rho}} \tag{5}$$

where $C_{\rho Y}$ is the covariance between the target variable values and the probability of responses:

$$C_{\rho Y} = \frac{1}{N}\sum_{k=1}^{N} (\rho_k - \bar{\rho})(Y_k - \bar{Y}) \tag{6}$$

The sample mean slope (as an estimator of the population mean) is determined by two factors:
- The average response probability. The more people who take part in the survey, the higher is the average response probability.
- The relationship between the target variable and the respondent behavior. The higher correlation between the values of target variables and the probability of response, the higher is the slope.

S_ρ is the standard deviation of the response probabilities. With the response probability mean $\bar{\rho}$, then the standard maximum value S_ρ can not exceed the:

$$S_\rho \le \sqrt{\bar{\rho}(1-\bar{\rho})} \tag{7}$$

3.3 Post-stratification

To perform post-stratification, one or more qualitative auxiliary variable is needed. The model below considers only one variable. Extending to multiple variables is essentially the same.

Suppose that there is an auxiliary variable X with L categories. Thus the target population is divided into L layers. Layers are denoted by subsets U_1, U_2,..., U_L of the population U. The number of elements of the target population in layer U_h is N_h, h = 1, 2,..., L. Population size N is: $N = N_1 + N_2 +... + N_L$.

Suppose an auto-selected sample is selected from the Internet population. If n_h is the number of respondents in the layer h, then $n = n_1 + n_2 +... + n_L$. After post-stratification, the estimator takes the following form:

$$\bar{y}_{PS} = \frac{1}{N}\sum_{h=1}^{L} N_h \cdot \bar{y}_h = \sum_{h=1}^{L} W_h \cdot \bar{y}_h \tag{8}$$

where $\bar{y}_h$ is the sample mean in the layer h and $W_h = N_h/N$ is the relative size of the layer h.

These circumstances can be achieved if there is a strong relationship between the target

variable Y and the stratified variable X, when the variation in Y values occurs between layers, but will not be manifested inside layers. In other words, the layers are homogeneous in terms of the target variable.

As a conclusion we can say that the post-stratification application is successfully applied, if appropriate auxiliary variables can be found.

These variables must satisfy three conditions:

- They must be measured in the survey;
- The population distribution (N_1, N_2, ..., N_L) must be known;
- They must produce homogeneous layers.

Unfortunately, these variables are rarely available, or have only a weak correlation. If these variables are not available, one should consider undertaking a reference survey. This reference survey is based on a small random sample, if the data collection takes place in a different way than on the Web, for instance computer assisted interviewing or assisted interviewing by phone.

To explore this, it is assumed that there is at least one auxiliary qualitative variable, both in the web survey and in the reference survey, and that this variable has a strong correlation with the target variable of the survey.

Then, using a form of post-stratification, we use the means of layers that are estimated based on the web data, respectively using the reference survey data.

The post-stratification estimator is written as follows:

$$\bar{y}_{RS} = \sum_{h=1}^{L} \frac{m_h}{m} \cdot \bar{y}_h \qquad (9)$$

where $\bar{y}_h$ is the mean of layer h of target population for the survey based on estimations (for h = 1, 2, ..., L), and m_h/m is the relative size of the sample in the layer h for the reference sample (for h = 1, 2, ..., L).

Under the conditions described above, quantity m_h/m is an impartially estimation of $W_h = N_H / N$ (where N_H is the population in layer h and N is the total population).

A strong relationship between the target variable and the auxiliary variable used to calculate the means show that there is little or no variation of the target variable within the layers. Therefore, the correlation between the target variable and the behavioral response will be small, and the same is true for the standard deviation of the targeted variable.

4 SIMULATION STUDY

It is very important to have quick access to the research information. Time is valuable and should not be wasted searching for information in e-mails or in research reports from archive. All can be calculated with specific applications for the management of the research indicators [7].

It is important to fully benefit from the research data and to be able to perform analysis and further processing when needed. The proposed data processing application with capacities of analysis for the data categories, it is easy to use.

Consider a study based on a population of 30000 people, with five variables:

• *Age* in two categories: young people (65%) and elders (35%).

• *Nationality* in two categories: natives (80%) and the non-natives (20%).

• *Internet access* in two categories Yes and No. The probability of having Internet access depends on two variables: age and nationality. Recent studies and literature indicate the following: Internet access decreases with age, music listening on the Internet decreases with age and reading newspapers on the Internet increases with age. According to these aspects, we considered the following values for this study: for natives, we considered the probability of having Internet access 90% (for young people) and 50% (for the elderly).

• *Frequency of listening to music on the internet* (per week). Of course, this frequency is 0 for those who do not have access to Internet. For those with Internet access, frequency depends on age. For young people, the frequency was an integer randomly chosen from the interval [0, 25]. For elderly, the frequency was an integer randomly chosen from the interval [0, 5].

• Frequency of reading newspapers on the Internet (weekly). This frequency depends on age only. For young people, the frequency was an integer randomly chosen from the interval [0, 3]. For elderly, the frequency was an integer randomly chosen from the interval [4, 7].

Reading newspapers and listening to music on the Internet were used as target variable.

There is a direct relationship between the target variable: reading newspapers and the auxiliary variable: age, and there is no a direct relationship between age and the lack of internet access. Estimates will be biased, but correction with variable age should contribute to reducing prejudice.

In the estimation process, "sample mean" becomes an estimator for the parameter: population mean. The most appropriate method to estimate is through interval estimation. It offers the opportunity

to show up whether the parameter belongs to that interval with a desired confidence level, set in advance.

Interval estimation associated with a fixed confidence level (90%, 95% etc.) is called a confidence interval for the estimated parameter. In this sense, the confidence interval for the parameter I is:

$$\bar{y} - z_{\alpha/2} \cdot S(\bar{y}) \le I \le \bar{y} + z_{\alpha/2} \cdot S(\bar{y}) \qquad (10)$$

where:
$z_{\alpha/2}$ = value of the normal distribution table (distribution z) corresponding to the established confidence level (90%, 95%, 99% etc.);
$S(\bar{y})$ = standard deviation from the mean;
$\bar{y}$ = population mean.

Thus, if we choose a confidence level of 95%, this means at the same time, an error of 5%, i.e. a 5% mischance of having a sample mean that is not placed within the confidence interval.

As stated above, the sample mean has an normal distribution with mean $\bar{y}$, (unknown mean of the population, one that we wish to estimate), and the standard deviation $S(\bar{y})$. But if the distribution of I is normal with mean $\bar{y}$ and standard deviation $S(\bar{y})$, then it is true formula:

$$P(\bar{y} - z_{\alpha/2} \cdot S(\bar{y}) \le I \le \bar{y} + z_{\alpha/2} \cdot S(\bar{y})) \approx 0.95 \quad (11)$$

i.e., the probability that the random variable I to be between the mean plus or minus 1.96 standard deviations is 0.95, almost one.

The standard error for Internet population is smaller as the sample size increases. This causes the confidence intervals decrease. When drawing a simple random sample of the target population, the sample mean has (approximately) a normal distribution. Then, the 95% confidence interval for the population mean is equal to:

$$I = [\bar{Y} - 1.96 \cdot S(\bar{Y}); \bar{Y} + 1.96 \cdot S(\bar{Y})] \qquad (12)$$

where $S(\bar{Y})$ is the standard error of the sample mean.

The "Survey" application allows recording statistical data presented in this paper. The interface is designed in Visual Basic using Microsoft Visual Studio 2010.

The database for storing the application data is memorized in an Access file. We chose this solution so that data can be stored in a relatively compact format, allowing an easy transfer along with the application, in parallel with the advantages of using Transact-SQL language for querying and processing the stored information.

One important goal was to design the application to be user friendly. The Startup Form (Fig. 1) displays the number of surveys recorded in the database. From here, by pressing the "Survey Form" button the user can add a new survey in the database using the form shown in Fig. 2, or, by pressing the "Survey results" button the Survey Results Form (Fig. 3) will be shown.

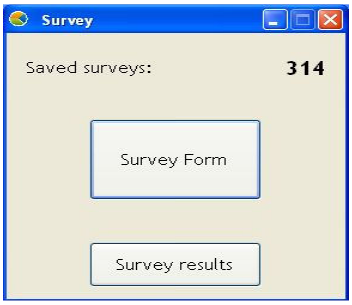


Fig. 1: The Start-up Form

In the Survey Form the user will fill the age, the nationality and the answers to the target questions. If the nationality can not be selected from the previously recorded data, the user can add the new value by selecting "Other" in the nationality's combo box and filling the "Add nationality" text box.

When the "SAVE" button is pressed, the application checks if all required data is filled in correctly and if so, the data is recorded in the database and the form is closed.

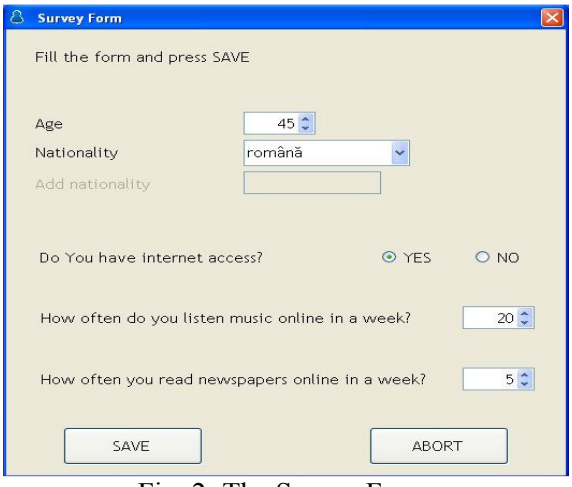


Fig. 2: The Survey Form

Based on the information stored in the database, in the "Survey Results Form" the user will input the sample population size and the confidence level to calculate the sampling error by pressing the "COMPUTE" button.

ISBN: 978-1-61804-189-0

The "DELETE" button allows deleting the data in this form for a recalculation of the sampling error. The "RESULTS TABLE" button displays the Results Table Form (Fig. 4).

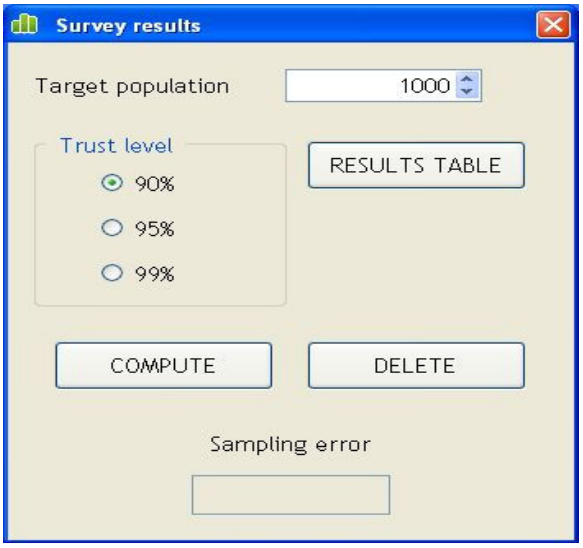
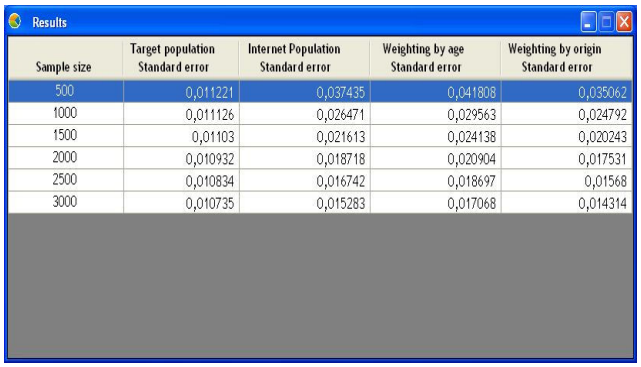


Fig. 3: The Survey Results Form

The "Results Table Form" shows sampling error values for a value of 95% of confidence level.



Sample size	Target population Standard error	Internet Population Standard error	Weighting by age Standard error	Weighting by origin Standard error
500	0,011221	0,037435	0,041808	0,035062
1000	0,011126	0,026471	0,029563	0,024792
1500	0,01103	0,021613	0,024138	0,020243
2000	0,010932	0,018718	0,020904	0,017531
2500	0,010834	0,016742	0,018697	0,01568
3000	0,010735	0,015283	0,017068	0,014314

Fig. 4: The Results Table Form

5 CONCLUSION

The main objective of this research was to develop interdisciplinary ways of using statistics to evaluate the behavior of web users.

The proposed reference survey is a way to eliminate the estimates in web surveys. One of the advantages of this survey is the fact that the auxiliary variables can be used for weighting so that to be correlated with the target variables, or with the missing ones. In this way, the correction will be more effective.

From the "Results Table Form" the following conclusions can be drawn:

- The standard error of the estimate of the reference survey is substantially higher than the standard error of the post-stratification estimator. This was expected. It was shown in section 3.3 that the standard error of the estimate of the reference survey is mainly determined by the sample size of the reference survey. For example, the standard error for a web survey of size 500 is equal to 0.0112. If the web sample size increased from 500 to 3000, the standard error is reduced from 0.0112 to 0.0107.
- The same is observed for the other targeted variables (Internet population, age, nationality).

In the future we plan to extend this research considering more targeted variables to achieve a more detailed evaluation of web user behavior.

References:

[1] Bethlehem, J.G., *Applied Survey Methods – A Statistical perspective*, John Wiley & Sons, Hoboken, NJ, USA, 2009.

[2] Fricker, R.D. Jr, Schonlau, M., Advantages and disadvantages of internet research surveys: evidence from the literature, *Field Methods,* Vol. 14 No. 4, 2002.

[3] Groves, R.M., Brick, J.M., Couper, M., Kalsbeek, W., *Alternative practical measures of representativeness of survey respondent pools,* Survey Practice, 2008. (http://surveypractice.org/2008/10/30/issues-facing-thefield/#more-302)

[4] Malhotra, N.K., *Marketing Research: An Applied Orientation*, 4th ed., Prentice Hall, Englewood Cliffs, NJ, 2004.

[5] McDaniel, C., Gates, R., *Marketing Research*, 6th ed., John Wiley & Sons, New York, NY, 2005.

[6] Porojan, D., *Statistica şi teoria sondajului*, Editura Şansa, Bucureşti, 1993.

[7] Schonlau, M., Van Soest, A., Kapteyn, A., Are "webographic" or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods*, Vol. 1, No. 3, 2007, pp. 155-163.

[8] Tingling, P., Parent, M., Wade, M., Extending the capabilities of internet-based research: lessons from the field, *Internet Research*, Vol. 13 No. 3, 2003, pp. 223-35.

[9] Wilson, A., Laskey, N., Internet-based marketing research: a serious alternative to traditional research methods?, *Marketing Intelligence & Planning*, Vol. 21 No. 2, 2003, pp. 79-84.