FROM Meunier Catherine catherine.meunier@insee.fr

SAMPLING – SURVEYS on WEB

HOW ADEQUATE ARE THE WEB SURVEYS in RELATION WITH THE OTHER KINDS OF SURVEY? (E.g. phone-survey, etc)

1. Surveys on Web

The development of the Internet in the last decades is very-very fast and so it has led to a new type of survey data collection: *the web survey*.
Web surveys are becoming increasingly popular. This is not surprising. A web survey is a simple means of getting access to a large group of potential respondents. Questionnaires are distributed [3], [2], at very low costs. No interviewers are needed, and there are no mailing and printing costs. Surveys can be launched very quickly. Little time is lost between the moment the questionnaire is ready [3] and the start of the fieldwork. Web surveys also offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation and movies).
According to the above the web survey seems to be a fast, cheap and attractive means of collecting large amounts of data. This raises the question whether or not the web surveys can be used for data collection in general population surveys. At first view, the web surveys seem to have much in common with other types of surveys. We can say that It is just another mode of data collection. Questions are not asked face-to-face [3] or by phone, but over the Internet. There are, however, some methodological issues. Some of them are discussed in this place.

2. Under-coverage

Web surveys (on general population or not) may suffer from under-coverage because the target population is usually much wider than the Internet population (frame population [3], [1]). For example, according to data from Eurostat, the statistical bureau of the European Union, 54% of the households in the EU had access to Internet in 2007. There were large variations between countries. The countries with the highest %-ges of Internet access were The Netherlands (83%), Sweden (79%) and Denmark (78%). Internet access was lowest in Bulgaria (19%), Romania (22%) and Greece (25%). Internet access will even be lower in many other countries in the world. Even more problematic is that Internet access is distributed over the Population with large sizes of variance. A typical law for many countries is that the elderly, the low-educated and the ethnic minorities are severely under-represented among those having access to Internet. To obtain insight in the impact of under-coverage on estimates, suppose a simple random sample is selected from the Internet population. Let the target population of the survey consist of N persons. Associated with each person k is a value Y_k of the target variable Y. The aim of the web survey is assumed to be estimation of the population mean $\bar{X} = \frac{1}{N}\sum_{m=1}^{N} X_m$ of the target variable X.

The population $\boldsymbol{U}$ is divided into 2 sub-populations $\boldsymbol{U_1}$ and $\boldsymbol{U_2}$ of size N_1 & N_2 of persons, having access to Internet or no, respectively. The sub-population $\boldsymbol{U_1}$ will be called the Internet population The sample mean $\bar{x}_i$ is an unbiased estimator of the mean $\bar{X}_i$ of the Internet population [1], [2], [3], but not necessarily of the mean of the target population. Bethlehem (2009) shows that the bias of this estimator is equal to

$$B(\bar{x}_{HT}) = E(\bar{x}_{HT}) - \bar{X} = \bar{X}_1 - \bar{X} = \frac{N_2}{N}(\bar{X}_1 - \bar{X}_2) \qquad (2.1)$$

The magnitude of this bias is determined by two factors. The first factor is the relative size N_2/N of the sub-population without Internet. Therefore the bias decreases as Internet coverage increases. The second factor is the difference $\bar{X}_1 - \bar{X}_2$ between the 2 means of the Internet-population and the non-Internet-population. *The more the mean of the target variable differs for these two sub-populations, the larger the bias will be.* Since the Internet coverage is steadily increasing, the factor N_2/N is decreasing. This has a bias reducing effect. However, it is not clear whether the contrast also decreases. To the contrary, it is not unlikely that the (small) group of people without Internet will be more and more different from the rest of the population and so, a kind of bias may still remain.

3. Self-selection

The basics of probability sampling as applied now in official statistics have been laid down by Horvitz and Thompson in their seminal paper in 1952. They prove that unbiased estimators of population characteristics can always be constructed, provided samples are selected by means of probability sampling and every element in the population has a known and strictly positive probability of being selected. Unfortunately, many web surveys are based on some form of self-selection. The survey is simply put on the web. Participation requires in the first place that respondents are aware of the existence of a survey. They have to accidentally visit the website, or they have to follow up a banner, e-mail message, or a telephone call. In the second place, they have to make the decision to fill in the questionnaire on the Internet. The survey researcher is not in control of the selection process. All this means that each element k in the population has unknown probability ρ_k of participating in the survey, for k = 1, 2, ..., N. Bethlehem (2009) shows that the expected value of the sample mean is equal to $E(\bar{y}) \approx \bar{Y}^* = \frac{1}{N}\sum_{k=1}^{N}\frac{\rho_k}{\bar{\rho}}Y_k$ (3.1)

where $\bar{\rho}$ is the mean of all response propensities. The bias of this estimator is equal to

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{Cov(\rho, Y)}{\bar{\rho}} = \frac{R_{\rho,Y} \cdot S_\rho \cdot S_Y}{\bar{\rho}} \ , \qquad (3.2)$$

in which $Cov(\rho, Y)$ is the covariance between the target variable and the response probabilities, $R_{\rho Y}$ is the correlation coefficient, S_ρ is the standard deviation of the response probabilities, and S_Y is the standard deviation of the target variable. It can be shown that in the worst case (S_ρ assumes it maximum value and the correlation $R_{\rho Y}$ is equal to either +1 or -1) the absolute value of the bias is equal to

$$\left|B_{\max}(\bar{y})\right| = S_Y\sqrt{\frac{1}{\bar{\rho}} - 1} \ . \qquad (3.3)$$

Bethlehem (1988) shows the formula (3.2) also applies in the situation in which a probability sample has been drawn, and subsequently no response occurs during the fieldwork. Consequently, expression (3.3) provides a means to compare potential biases in various survey designs. For example, regular surveys of Statistics Netherlands are all based on probability sampling. Their response rates are around 70%. This means the absolute maximum bias is equal to $0.65 \times S_y$. One of the largest self-selection web surveys in The Netherlands was the public opinion survey described on the website 21minuten.nl. Within a period of six weeks in 2006 about 170,000 people completed the online questionnaire. The target population of this survey was not defined, as everyone could participate. If it is assumed the target population consists of all Dutch from the age

of 18, the average response propensity is equal to 170,000 / 12,800,000 = 0.0133. Hence, the absolute maximum bias is equal to $8.61 \times S_y$. It can be concluded that the bias of the large web survey can be a factor 13 larger than he bias of the smaller probability survey.
General population web surveys based on self-selection are unacceptable if objective is to obtain accurate estimates of population characteristics. Proper probability sampling is required. This is not easy to implement as it requires a sampling frame of e-mail addresses. This is usually not available. The way out is to recruit sample persons using a different such as mail, telephone or face-to-face. For example, a letter can be sent to all sample persons containing the Internet address of the survey questionnaire and a unique identification code.

4. Weighting adjustment

It is often attempted to correct for a bias due to under-coverage or self-selection by applying some kind of weighting adjustment technique. Weighting requires auxiliary variables. These variables must have been measured in the survey, and moreover information on their population distribution (or complete sample distribution) must be available. By comparing the population distribution of an auxiliary variable with its response distribution, it can be assessed whether or not the sample is representative for the population (with respect to this variable). If these distributions differ considerably, one must conclude that the sample is selective. To correct this, adjustment weights can be computed. Weights are assigned to all records of observed elements. Estimates of population characteristics can now be obtained by using the weighted values instead of the unweighted values.
Post-stratification is a well-known and often used weighting method. To carry out post-stratification, one or more qualitative auxiliary variables are needed. Together they divide the target population into a number of strata (i.e. sub-populations). Identical adjustment weights are assigned to all elements in the same stratum. The bias of the estimate based on weighted data will be small if there is (on average) no difference between participants and non-participants. This is the case if there is a strong relationship between the target variable and the stratification variables. This situation is referred to in the literature as Missing at Random (MAR). The variation in the values of the target variable manifests itself between strata but not within strata. In other words, the strata are homogeneous with respect to the target variable. Unfortunately, such auxiliary variables are not very often available, or there is only a weak correlation.

5. Measurement errors

Traditionally, general population surveys are face-to-face surveys or telephone surveys.
Can a web survey be an alternative for such surveys? With respect to data collection, there is a substantial difference between face-to-face and telephone surveys on the one hand and web surveys on the other. Interviewers carry out the fieldwork for face-to-face and telephone surveys. They are important in convincing people to participate in the survey, and they also can assist in completing the questionnaire. There are no interviewers in a web survey. It is a self-administered survey. Therefore quality of collected data may be lower due to higher nonresponse rates and more errors in answering questions. However, response to sensitive questions is higher without interviewers. Moreover, respondents may be more willing to give truthful but socially undesirable answers in a web survey.
CAPI and CATI are the computer-assisted forms of face-to-face and telephone

interviewing. Computer-assisted interviewing (CAI) has the advantage that error checking can be implemented. It means that answers to questions are checked for consistency. Errors can be detected during the interview, and therefore also corrected during the interview. It has been shown CAI can improve the quality of the collected data. The question is whether error checking should be implemented in a web survey? What happens if respondents are confronted with error messages? Maybe they just correct their mistakes, but it may also happen that they will become annoyed and stop answering questions. There may be a trade-off here between nonresponse and data quality. Further research should make clear what the best approach is.

6. Mixed-mode surveys

A web survey can be one of the modes in a mixed-mode data collection approach. Each mode of data collection (face-to-face, telephone, mail, web, etc) has its advantages and disadvantages. Mixing data collection modes provides an opportunity to compensate for the weakness of each individual mode. This can reduce costs and at the same time increase response rates and data quality. Several mixed-mode data collection strategies are possible.. An interesting approach from the point of view of reducing costs is to start with a web surveys. Non-respondents are followed up by the next cheapest mode (CATI), and finally remaining non-respondents by the most expensive mode CAPI). Another strategy could be to let respondents select their preferred data collection mode.
A major concern for mixed-mode data collection is that data quality may affected by the occurrence of mode effects. This is the phenomenon that asking a person the same question in different data collection modes would lead to different answers. An example is asking a closed question with a substantial number of answer options. The respondent in a face-to-face survey would be presented a show card with all possible answers. In case of a telephone survey, the interviewer would read all possibilities to the respondents. Research indicates that this results in a preference for the last options in the list (recency effect). Respondents offered a drop-down list in a web survey have a preference for answers early in the list (primacy effect).

7. References and interesting texts

[1] Bethlehem, J.G. (2009), Applied Survey Methods, A Statistical Perspective. John Wiley & Sons, Hoboken, NJ.
[2] Couper, M. P. (2008), Designing Effective Web Surveys. Cambridge University Press, New York, USA.
[3] Farmakis, N. (2009), Surveys and Ethics, A & P Cristodoulidi, Thessaloniki (in Greek)