Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

www.czasopisma.uni.lodz.pl/foe/
4(349) 2020

DOI: http://dx.doi.org/10.18778/0208-6018.349.03

Aleksandra Kupis-Fijałkowska

University of Łódź, Faculty of Economics and Sociology, Department of Statistical Methods
aleksandra.kupis-fijalkowska@uni.lodz.pl

Selected Problems of Quality Assessment in Internet Surveys – a Statistical Perspective

Abstract: The paper presents selected problems related to the quality assessment from the statistical perspective of survey data based on Internet sources. Internet access is consequently expanding all over the world. In parallel with the running development of other new technologies, it is pervading daily life and business activities more and more. It also has influenced surveys practice to a large extent as a research tool for collecting both primary and secondary data, and it also challenges surveys to research the Internet population. Moreover, as the Internet and its entities are able to register all activities that are performed on the web, issues related to big data and organic data processing as well as their applications arise. As a result of decreasing response rates and increasing survey costs, Internet data collection is constantly growing. Due to many advantages, Internet surveys are used widely and this process seems to be inevitable. However, it needs to be emphasised that Internet surveys are developing in practice faster than the methodology in this area. Hence, a lot of problems can be identified, especially when considering the quality of data based on Internet sources. The following issues are discussed as the most far-reaching in the prism of statistical survey methodology: determination of the sampling frame, self-selection and related estimates bias, as well as under/over-coverage.

Keywords: Internet survey, online survey, survey quality, survey error

JEL: C8

1. Introduction

Internet coverage and its penetration rate are constantly growing. In parallel, interest in and usage of Internet sources and resources are increasing. This includes scientists, researchers, students, and occasional users. There are many postulates regarding today's surveys, however, the requirement of providing up to date data, delivered as fast as possible with the lowest possible cost, is the main reason that seems to encourage support for offline modes with Internet data collection or transfer of surveys completely to the web space (Bethlehem, 2010; Bethlehem, Biffignandi, 2011; Tourangeau, Conrad, Couper, 2013; Schonlau, Couper, 2017; Kalton, 2018). "The use of Internet for collecting survey-type data has grown enormously in recent years. […] However, the quality of the estimates produced is questionable" as G. Kalton wrote (Kalton, 2018: S12). And the theory of statistics is challenged to assess that quality as well as to make recommendations what statistical methods can be applied to improve the quality of the results. As it is invertible, the methodology must commensurate to the progress that occurs in practice. Researchers and recipients must be aware of its properties and a great deal of attention should be paid to the quality assessment (Szreder, 2017). The issue is complex, as it affects many areas of survey methodology (Schonlau, Cooper, 2017; de Leeuw, 2018).

2. Internet coverage & Internet population – introduction and influence on surveys

Internet coverage is constantly growing all over the world, its penetration is becoming wider and deeper. The continuous development of new technologies strengthens the effect of omnipresence of the Internet. Its applications and meaning are expanding for both individuals and corporate users. As the development of the information society is progressing, data demand is increasing. The Internet has had a significant impact on surveys by providing broader possibilities in the data collection process with lower costs. It has become a communication tool, a medium, and an easily accessible source of data. It is a social and business space now: individual users, social media, e-commerce; banking and accounting portals; news; government, public institutions, non-government organisations; corporations and enterprises. A hitherto unknown new dimension of human and business life has been created and the boundary between reality and virtuality is blurred now. The term of virtual society (understood as a sub-population of entities that have and use Internet access) has been introduced and, from the scientific point of view, a new collectivity has come to life: the Internet population – the population of Internet users. When studying the Official Statistics reports and different organisations' elaborations dedicated to the Internet and Internet surveys, it can be observed that as Internet coverage is rising, also the surveys conducted via and on the Internet are gaining in popularity. To illustrate it based on an example, a case of Poland will be presented. Currently, 84% households in Poland have Internet access. In Table 1 detailed statistics are listed for EU countries.

Table 1. Households – level of Internet access [%] in EU countries in 2010–2018

GEO/TIME	2010	2011	2012	2013	2014	2015	2016	2017	2018
European Union	70	73	76	79	81	83	85	87	89
Belgium	73	77	78	80	83	82	85	86	87
Bulgaria	33	45	51	54	57	59	64	67	72
Czechia	61	67	73	73	78	79	82	83	86
Denmark	86	90	92	93	93	92	94	97	93
Germany	82	83	85	88	89	90	92	93	94
Estonia	67	69	74	79	83	88	86	88	90
Ireland	72	78	81	82	82	85	87	88	89
Greece	46	50	54	56	66	68	69	71	76
Spain	58	63	67	70	74	79	82	83	86
France	74	76	80	82	83	83	86	86	89
Croatia	56	61	66	65	68	77	77	76	82
Italy	59	62	63	69	73	75	79	81	84
Cyprus	54	57	62	65	69	71	74	79	86
Latvia	60	64	69	72	73	76	77	79	82
Lithuania	61	60	60	65	66	68	72	75	78
Luxembourg	90	91	93	94	96	97	97	97	93
Hungary	58	63	67	70	73	76	79	82	83
Malta	70	75	77	78	80	81	81	85	84
Netherlands	91	94	94	95	96	96	97	98	98
Austria	73	75	79	81	81	82	85	89	89
Poland	63	67	70	72	75	76	80	82	84
Portugal	54	58	61	62	65	70	74	77	79
Romania	42	47	54	58	61	68	72	76	81
Slovenia	68	73	74	76	77	78	78	82	87
Slovakia	67	71	75	78	78	79	81	81	81
Finland	81	84	87	89	90	90	92	94	94
Sweden	88	91	92	93	90	91	94	95	92
United Kingdom	80	83	87	88	90	91	93	94	95
Iceland	92	93	95	96	96	na	na	98	99
Norway	90	92	93	94	93	97	97	97	96

Source: Eurostat, 2018

According to Statistics Poland's report "Information society in Poland. Results of statistical surveys in the years 2014–2018", in 2018 in Poland: 97.4% households with children and 95.6% of all enterprises had Internet access. Figure 1 presents how the Internet coverage has been growing in Poland since 2000.

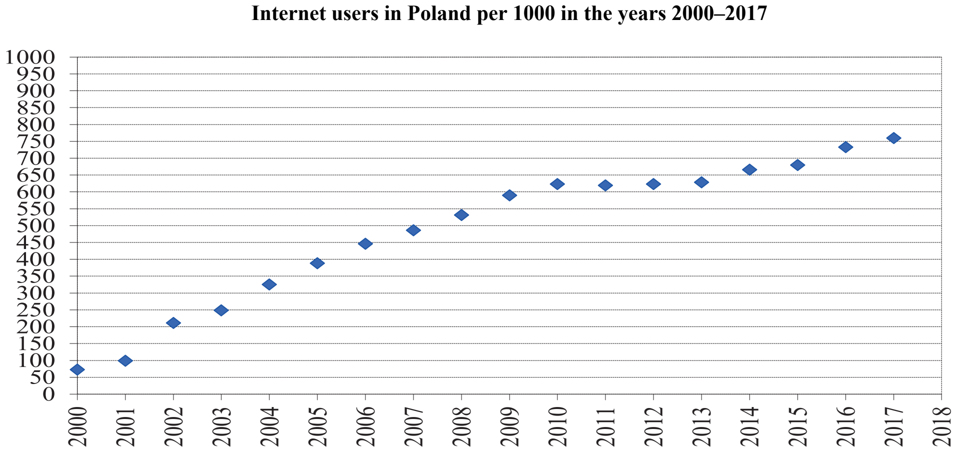

Figure 1. Internet users in Poland per 1000 population from 2000 to 2017

Source: own elaboration based on Statistics Poland data

The presented above descriptive statistics for Poland and different world regions regarding Internet access provide a good proof of the growing power of the Internet. There are no doubts that era of digitisation has come, and it is a natural consequence that surveys have to reach into online sources (Bethlehem, 2010; Callegaro, Manfreda, Vehovar, 2015).

It is important to mention smartphone users statistics here, as these devices connect to the web, which intensifies Internet penetration. The Polish Internet Survey by Gemius S. A. (2018) provides monthly data about Internet users as well as most popular websites and applications. According to the November 2018 report, 23.4 million Internet users in Poland were connecting via smartphone. And, according to the already mentioned report of Statistics Poland “Information society in Poland. Results of statistical surveys in the years 2014–2018”, in 2018, in Poland, 47.4% of individuals had access to the Internet via a mobile phone or smartphone.

As Internet access expands, the ratio of research based on Internet surveys is constantly growing, The Polish Society of Market and Opinion Researchers in its Yearbooks presents each year statistics about the situation of market and opinion research in Poland.

In 2008, 2.9% respondents in the market and opinion research in Poland were contacted by CAWI, 5 years later in 2012 it was nearly 25% and in 2017 the figure exceeded 50%, so more than half of all respondents were interviewed this way. In comparison to CATI, it can be observed that in the years 2008–2014 this mode was oscillating around 1/3 of all modes, and in 2015 a trend change occurred and it decreased to 27%, and then respectively to 25% in 2016 and to 21% in 2017.

Table 2. The CAWI and CATI modes by respondents [%] in the Polish market research and opinion sector from 2008 to 2017

Year/Method	CAWI	CATI
2008	2.9	29.6
2009	7.1	36.1
2010	18.3	33.4
2011	21.3	33.0
2012	24.6	32.5
2013	23.5	30.0
2014	30.3	31.0
2015	36.0	27.0
2016	48.0	25.0
2017	53.0	21.0

Source: own elaboration based on Yearbooks 2011/2012 to 2018/2019 – data of the Polish Society of Market and Opinion Researchers

The Official Statistics also recognises a great opportunity to conduct research on the Internet. For example, in Poland, in the National Census 2011, Statistics Poland used a mixed mode for its data collection process and Poles had the ability to decide on online self-interviewing (CAII) – all together around 12% respondents preferred this way of contact. Internet surveys sources, same as Big Data, have a huge potential to support official statistics, probably complementarily (Szreder, 2015; de Leeuw, 2018), however, using Internet based data sources for official statistics purposes at the moment is under discussion: scientists and statistical experts working groups are investigating the potential of available e-sources (Beręsewicz, Szymkowiak, 2015). Methodological studies are being carried out how to merge this type of sources to the official statistics area and how it could work with current legal regulations and good practices.

In summary, the development of new technologies has already influenced the survey execution process, and it is expected by many authors that a deeper influence will be observed in the future (de Leeuw, 2018; Kalton, 2018).

The terms "Internet survey" and "web survey" can be used interchangeably or can be understood differently, as they may be considered in different context/meanings, i.e. the mode of contact, the mode of response, or they may refer to the population of Internet users. Bethlehem and Biffignandi (Bethlehem, Biffignandi, 2011) proposed the following definitions:

> ***Internet survey*** is a general term for various forms of data collection via the Internet (i.e. a web survey, an e-mail survey), also all forms of data collection that use the Internet to transfer questionnaires and collected data between entities of interest;

> ***Web survey*** is a form of data collection via the Internet in which respondents complete questionnaires on the World Wide Web, the questionnaire is accessed by means of a link to a web page.

For the purpose of this article, the definitions given by Bethlehem and Biffignandi apply.

Bethlehem and Biffignandi (2011) also introduced the definition of self-selection survey, which will be referred to later in this paper, as:

> ***Self-selection survey*** is a survey for which the sample has been recruited by means of self-selection, hence users can decide whether or not to participate in the survey.

Many approaches can be found in the literature in the context of the mentioned definitions of the analysed terms (Bethlehem, Biffignandi, 2011; Tourangeau et al., 2013; Fielding, Lee, Blank, 2017), and new concepts, more detailed, are proposed as well. For example, in the prism of self-selection issue and entity responsible for data maintenance, Beręsewicz (Beręsewicz, 2015; 2017) introduced the following Internet data source (IDS) definition:

> ***Internet data source (IDS)*** is a self-selected (non-probabilistic) sample that is created through the Internet and maintained by entities external to NSIs and administrative regulations.

3. Internet surveys – benefits and problems

The Internet is already a successful tool for surveys, the main reason lies in many technical opportunities which it gives to researchers. It offers a broad spectrum of new tools, for example, in-time dynamic question adjustments, reaction time or mimics can be measured, or new multimedia tools are available: animations, movies, sound, high contrast interface, or online eye tracking. Regarding conducting surveys not by but on the Internet, its popularity is caused, as already mentioned, by growing Internet coverage and by the phenomenon that a large part of human life moves to the web relations building, shopping, paying bills, e-medicine, e-pharmacy, watching nature and entertainment places via cameras, voting, etc. Also, modern business depends more and more on the web and a lot of enterprises cooperate more online than offline.

Some advantages and disadvantages were already mentioned but here is a synthetic list of the most important ones (based on: Fricker, Schonlau, 2002; Bethlehem, Biffignandi, 2012; Tourangeau et al., 2013; Fielding, Lee, Blank, 2017).

The most visible benefits:

- quicker and cheaper data collection (at all stages of the data collection process);
- simplicity in comparison to other modes and attractive multimedia forms;
- quick respondent selection on the basis of required features (questionnaires can be filled with already available information, i.e. digital traces);
- no interviewer effect, higher individualisation;
- less intrusive and suffer less from social desirability effects;
- immediately sent and answered questionnaires, quick follow-ups and reminders;
- dynamic sequences of questions adapted to the specific respondent, which results in lower respondent burden and introduction of small modifications;
- reduction of the number of missing responses and partial answers as well as data entry errors;
- lower time and space respondent burden, the response burden can be easily monitored as server-side and client-side information is available;
- a new understanding of individual's anonymity and intimacy (it allows researchers to reach niche populations' opinions easier and investigate rare features more effectively).

And, respectively the list of the most visible disadvantages looks as follows:

- inability to construct a comprehensive sampling frame (can't identify all members of the Internet population and hence unable to apply the assignment rule[1]) that results in sample selection limitation as well as a lack of representativity and biased estimations;
- self-selection;
- coverage problems;
- low response rates;
- problem with bias measurement and quality assessment;
- technological exclusion and problem with respondents' computer skills;
- technical problems can occur;
- inability to confirm respondents' identity;
- "professional" respondents, multiple participation;
- unusual real-time situations can create problems resulting in discontinuation of answering.

In summary, from the statistical point of view, a lack of representativity (from the perspective of the probabilistic survey theory) is the main cause of reducing quality: the inability to define the sampling frame means that selection methods are extremely reduced, and in majority of cases the target population differs from the survey population (coverage problems).

[1] It is possible only for specific websites and if the page administrator keeps a registry of users.

4. Internet surveys data quality – a statistical perspective

To reach the most possible reliable information, it is crucial to make a solid research design as well as to choose and apply data collecting methods properly. It allows researchers to know all the details through the survey realisation process and be aware of all existing complications as well as possible error sources. Preceding the further considerations, the classic theory of survey sampling should be presented. Generally, in the early fifties of the twentieth century, the methodology of survey sampling was completed and became a common practice for the official statistics systems, as well as scientific and private sector research (Bethlehem, 2009). If Internet surveys are taken into consideration, fundamental principles of probability sampling and survey theory are not applied (Bethlehem, 2009), which results in the lack of representativity that generates low quality data. Especially, in the context of growing web surveys popularity, the obtained results are published frequently and their recipients are getting more familiarised with this type of surveys, so the results might be perceived as reliable, while they are not. It is observed that full information about the data collection process, problems, and their consequences is not revealed. In the context of probability sampling approach attributes, there are a lot of methodological issues to be solved in the nearest and further future. There are three main problems from the statistical point of view: Internet under/over-coverage, determination of the sampling frame and respondents' self-selection. All of the aforementioned issues result in a lack of (full) representativity, and thereby do not reflect the exact nature of the phenomena studied, so the quality is not sufficient. At the same time, some statistical tools exist and their implementation can improve the quality by toning down discrepancies, low precision and poor accuracy effects.

Probability sampling is crucial to obtaining the most possible reliable information. Selection of data collecting methods and a high quality survey execution process are crucial as well. A lot of surveys suffer from a lack of representativity, which causes the reliability of the collected data to be lower than it could.

The first three of the disadvantages listed above are the main methodological problems in web surveys from the statistical perspective, due to the generated bias: estimations based on the collected material differ significantly from the population parameters and no valuable inferences can be drawn about the researched phenomenon. Hence, the main objective of the conducted survey – obtaining reliable information – is not achieved. The bias in general can be caused by many errors that can occur in the survey execution process (Figure 2).

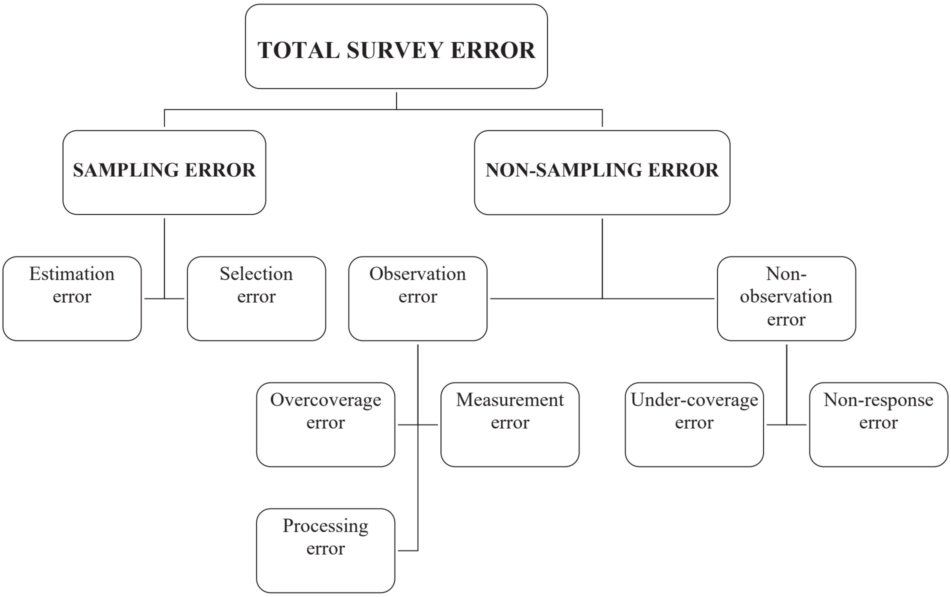

Figure 2. Taxonomy of survey errors

Source: Bethlehem, 2010: 164

Let us consider selected problems concerning data quality when using the Internet for collecting survey-type data that have the most far-reaching consequences from the statistical perspective.

From the prism of the statistical survey theory, it should be done in the context of the presented above breakdown of errors: undercoverage and selection errors that occur here. The first type of errors is the consequence of the inability to build the sampling frame, so no proper selection method can be applied. Hence, basically no proper random sample is selected and a self-selection situation occurs. It means that the respondent has to be aware of the existence of the questionnaire and has to decide to fill it. The other error source is obvious: not all elements of the target population have Internet access. Hence, there is no chance those units can be contacted and interviewed (Bethlehem, 2010).

A short statistical investigation will be introduced now in order to present how the bias caused by undercoverage error can be measured (Bethlehem, 2010). Let us consider the target population of N fully identifiable elements (each element k is labelled; $k = 1, 2, 3, \ldots, N$) and the target variable Y, where for each element k, a value Y_k exists. Let us assume that the web survey aims to estimate the value of the population simple mean for the target variable Y given as:

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k \ . \tag{1}$$

The population U is divided into two subpopulations, U_I – all elements with Internet access and U_{NI} – all elements without Internet access. Let each element k be characterised by the I_k indicator which:

$$I_k = \begin{cases} 1 \text{ for } k \in U_I \\ 0 \text{ for } k \in U_{NI} \end{cases} . \tag{2}$$

Hence, the number of U_I (Internet population) is equal to:

$$N_I = \sum_{k=1}^{N} I_k . \tag{3}$$

Respectively N_{NI} denotes the U_{NI} (non-Internet population) number, where:

$$N = N_{NI} + N_I . \tag{4}$$

The mean of the target variable for the U_I population is equal to:

$$\overline{Y}_I = \frac{1}{N_I} \sum_{k=1}^{N} I_k Y_k \tag{5}$$

and the mean of the target variable for the U_{NI} population is equal to:

$$\overline{Y}_{NI} = \frac{1}{N_{NI}} \sum_{k=1}^{N} (1 - I_k) Y_k . \tag{6}$$

Let us assume now that the sampling frame can be constructed for the Internet population and a random sample (simple random sampling scheme without replacement) represented by the following series is selected:

$$s_1, s_2, s_3, \ldots, s_{N-1}, s_N \tag{7}$$

of N indicators, where the kth indicator s_k assumes 1 if element k is selected and 0 if it is not, for $k = 1, 2, 3, \ldots, N - 1, N$. Hence, the sample size is equal to:

$$n_I = s_1 + s_2 + s_3 + \ldots + s_{N-1} + s_N = \sum_{k=1}^{N} s_k . \tag{8}$$

The first-order inclusion probability of the kth element is defined by the following expected value:

$$\pi_k = E(s_k)\,. \tag{9}$$

The Horvitz-Thompson estimator for the mean of the U_I population is defined by:

$$\bar{y}_{HT} = \frac{1}{N_I}\sum_{k=1}^{N} s_k I_k \frac{Y_k}{\pi_k}\,. \tag{10}$$

The inclusion probability π_k for all elements outside the Internet population is equal to 0:

$$\pi_k = 0. \tag{11}$$

When we deal with a simple random sample from the Internet population, all inclusion probabilities are equal to:

$$\pi_k = \frac{n}{N_I}\,. \tag{12}$$

Hence, expression (10) reduces to:

$$\bar{y}_I = \frac{1}{n}\sum_{k=1}^{N} s_k I_k Y_k\,. \tag{13}$$

Expression (13) represents an unbiased estimator of the mean $\bar{Y}_I$ given by expression (5), but not necessarily of the mean $\bar{Y}$ given by expression (1).

Let us denote $B(\bar{y}_{HT})$ as the estimator bias, in the discussed situation, it is equal to:

$$B(\bar{y}_{HT}) = E(\bar{y}_{HT}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI})\,. \tag{14}$$

Expression (14) shows that the magnitude of this bias is determined by the following two factors:

– the relative size of $\frac{N_{NI}}{N}$ of the U_{NI} population, and the larger this proportion is, the higher bias occurs;

– the difference $(\overline{Y}_I - \overline{Y}_{NI})$, and the larger this difference is, the higher bias occurs.

As not everyone has web access, two sub-populations exist: the Internet population and the non-Internet population. Their structures can differ, for example, while considered through the prism of age, structures of the U_I and U_{NI} populations can be much different.

The next quality issue that should be discussed is the self-selection problem (Bethlehem, 2010). As the participation requires the awareness of the existence of the survey, and then the decision whether to participate in it or not, this means that each element k ($k = 1, 2, 3, \ldots, N - 1, N$) of the Internet population has unknown probability ρ_k of individuals participating in the survey. The responding elements are denoted by a vector:

$$r_1, r_2, r_3, \ldots, r_{N-1}, r_N, \tag{15}$$

where $r_k = 1$ if the k^{th} element responds and $r_k = 0$ if it does not, for $k = 1, 2, 3, \ldots, N - 1, N$. Let the probability of response of element k be given as the expected value $\rho_k = E(r_k)$.

Considering U_{NI}, all response probabilities for elements in the non-Internet population are 0.

The obtained sample size is denoted by:

$$n_S = r_1 + r_2 + r_3 + \ldots + r_{N-1} + r_N = \sum_{k=1}^{N} r_k \,. \tag{16}$$

If every element in the Internet population had the same probability of being included in the sample, then the estimator for the population mean would be expressed as:

$$\overline{y}_S = \frac{1}{n_S} \sum_{k=1}^{N} r_k Y_k \tag{17}$$

and its expected value would be approximately equal to:

$$E(\overline{y}_S) \approx \overline{Y}_I^* = \frac{1}{N_I \overline{\rho}} \sum_{k=1}^{N} \rho_k I_k Y_k, \tag{18}$$

where $\overline{\rho}$ is the mean of all response propensities in the Internet population.

It can be shown (Bethlehem, 2010) that the bias of the estimator given by (17) can be expressed as:

$$B(\overline{y}_S) = E(\overline{y}_S) - \overline{Y}_I \approx \overline{Y}_I^* - \overline{Y}_I = \frac{\operatorname{cov}(\rho, Y)}{\overline{\rho}} = \frac{R_{\rho Y} SD_\rho SD_Y}{\overline{\rho}}, \tag{19}$$

in which the covariance between the values of the target variable and the response probabilities in the Internet population is given as:

$$\operatorname{cov}(\rho, Y) = \frac{1}{N_I} \sum_{k=1}^{N} I_k (\rho_k - \overline{\rho})(Y_k - \overline{Y}) \tag{20}$$

and respectively:
$\overline{\rho}$ is the average response probability;
$R_{\rho,Y}$ is the correlation coefficient between the target variable and the response behaviour;
SD_ρ is the standard deviation of the response probabilities;
SD_Y is the standard deviation of the target variable.

In the case of self-selection, the bias is determined by the following factors:

- the average response probability;
- the variance of response probabilities;
- the relationship between the target variable and the response behaviour.

As the general population is considered, the bias of the sample mean consists of under-coverage and self-selection biases and can be expressed as:

$$B(\overline{y}_S) = E(\overline{y}_S) - \overline{Y} = E(\overline{y}_S) - \overline{Y}_I + \overline{Y}_I - \overline{Y} = \frac{N_{NI}}{N}(\overline{Y}_I - \overline{Y}_{NI}) + \frac{\operatorname{cov}(\rho, Y)}{\overline{\rho}}. \tag{21}$$

There are different methods to reduce the bias of the estimates in such cases and increase informativity of Internet survey results (Bethlehem, 2010). The most popular ones are weighting adjustment methods, including post-stratification weighting, weighting adjustment with a reference sample, propensity score adjustment, and rim weighting. However, it should be emphasised that only from the theoretical point of view those methods should be sufficient to deal with the bias. In practice, the application of those techniques does not result in the bias elimination but only allows for some reduction of it (Bethlehem, Biffignandi, 2012).

Internet surveys, in general, suffer from a problem of nonresponse (unit nonresponse or item nonresponse). It is the most recognised source of errors from the statistical point of view (Schouten et al., 2012). In the case of web surveys, Bethlehem (2012) has shown that the expression for the bias in the case of random sample affected by nonresponse is identical as (19), as the magnitude of the nonresponse bias is equal to:

$$B(\bar{y}_R) = \tilde{Y} - \bar{Y} = \frac{\operatorname{cov}(\rho, Y)}{\bar{\rho}} = \frac{R_{\rho Y} SD_{\rho} SD_{Y}}{\bar{\rho}}. \tag{22}$$

This means that in the case of web surveys, the bias generated by self-selection corresponds to the non-response one.

The non-response is recognised as a serious source of survey errors. The related bias of estimates is determined by two factors (Skinner et al., 2009):

– how respondents and non-respondents differ, on average, with respect to the target variable (the contrast between response and non-response);
– the number of responses in the survey (the response rate sets a bound to the maximal impact of non-response).

To assess the effects of non-response on the quality of estimators, both the response rate itself and the contrast (between respondents and non-respondents) should be investigated. It is discussed in the literature (Groves, Peytcheva, 2008; Schouten, Cobben, Bethlehem, 2009) that response rates by themselves are not sufficient indicators of the non-response bias. Schouten Cobben and Bethlehem (2009) found that increases in response rates due to follow-up efforts did not significantly improve response representativeness.

To complete the quality assessment based on the response rate, supplemental survey quality measures are proposed, including: R-indicators (Representativeness indicators), bias reduction indicators, Mahalanobis distance, response rates for key domains, or tracking key survey estimates.

Currently, in the context of Internet surveys, the R-indicators concept seems to be the most widely discussed in the literature as a supplemental quality measure to the response rate (Shlomo et. al., 2008). Although the response rate should be treated as the core indicator of the survey quality, it does not necessarily express all the aspects that influence the representativity of the survey results suffering from non-response. In this paper, the R-indicator as a measure based upon the variance of estimated response probabilities (Cobben, Schouten, 2005; 2007; Schouten, Cobben, Bethlehem, 2009) will be discussed.

Let us suppose that a sample survey is undertaken where a sample *s* is selected from a finite population *U*. The sizes of *s* and *U* are denoted *n* and *N*, respectively. The units in *U* are: $i = 1, 2, \ldots, N$. The sample is assumed to be drawn by the probability sampling design *p*(.) where the sample *s* is selected with probability *p*(*s*).

Let us denote s_i as the 0–1 sample indicator (if unit *i* is sampled, it takes the value 1 and 0 otherwise), r_i as the 0–1 response indicator for the unit *i* (if unit *i* is sampled and did respond, it takes the value 1 and 0 otherwise), so the set of respondents is given as *r* ($r \subset s \subset U$) and π_i as the first-order inclusion probability of unit *i*. Let us assume that no-response occurs.

Let ρ_i be the probability that the unit *i* responds when it is sampled. Let us consider that response propensity is motivated by a variable *X* (more than one could be assumed), then the expected conditional response propensity is given as:

$$\rho_i = \rho_X(x_i) = E(R_i \setminus x_i)\,. \tag{23}$$

In respect to the survey response, two definitions of representativeness (as a wide concept of the response representativeness, not in the understanding of the sampling theory) were introduced (Schouten, Cobben, Bethlehem, 2009) strong and week.

Definition (strong): A response subset is representative with respect to the sample if the response propensities ρ_i are the same for all units in the population:

$$\forall i \quad \rho_i = P[r_i = 1 \,|\, s_i = 1] = \rho \tag{24}$$

and if the response of a unit is independent of the response of all other units.

If a missing-data mechanism satisfies the strong definition, then the mechanism will correspond to Missing-Completely-at-Random (MCAR) with respect to all survey questions. The validity of the strong definition cannot be verified in practice, so a weak definition was proposed (Schouten, Cobben, Bethlehem, 2009).

Definition (weak): A response subset is representative of a categorical variable *X* with *H* categories if the average response propensity $\overline{\rho}$ over the categories is constant:

$$\overline{\rho} = \frac{1}{N_h}\sum_{k=1}^{N_h} \rho_{hk} = \rho, \text{for } h = 1, 2, ..., H, \tag{25}$$

where:

N_h is the population size of category *h*;

ρ_{hk} is the response propensity of the unit *k* in the class h and summation is over all units in this category.

The week definition corresponds to MCAR with respect to *X*, as distinguishing respondents from nonrespondents based on knowledge of *X* is not possible. Hence, regarding a week definition, the response propensities can be estimated within corresponding strata based on *X*, so the assumption of weak representativity can be verified in practice.

Schouten, Cobben and Bethlehem (2009) introduced the R-indicator for the evaluation of a representative response as a measure based upon the variance of estimated response probabilities.

Let us consider the hypothetical situation with all individual response propensities known – a strong definition could be tested and measurement of variability in the response propensities would be easy, and the more variation, the less representativity in the context of the strong definition.

Let ρ = (ρ$_1$, ρ$_2$, …, ρ$_N$)’ be a vector of response propensities, let 1 = (1, 1, …, 1)’ be the N – vector of “1”, and let $\rho_0 = 1 \times \rho$ be the vector of the average population propensity:

$$\bar{\rho} = \frac{1}{N}\sum_{i=1}^{N}\rho_i \ . \tag{26}$$

Any distance function d in $[0, 1]^N$ would suffice in order to measure the deviation from the strong representative response (the strong definition) by measuring the distance $d(\rho, \rho_0)$.

The Euclidean distance can be applied to the distance $d(\rho, \rho_0)$ and the measure proportional to the standard deviation of the response probabilities is given as:

$$SD(\rho) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2} \, . \tag{27}$$

When fixing the average response probability $\bar{\rho}$, the maximum possible variance value is obtained by letting $\bar{\rho}N$ of the response probabilities be equal to 1 and respectively $(1-\bar{\rho})N$ to value 0 (Cobben, 2009), hence:

$$SD(\rho) \le \sqrt{\bar{\rho}(1-\bar{\rho})} \ . \tag{28}$$

Moreover, for $\bar{\rho} = \frac{1}{2}$:

$$SD(\rho) \le \sqrt{\bar{\rho}(1-\bar{\rho})} \le \frac{1}{2} \, . \tag{29}$$

The R-indicator proposed by Schouten, Cobben and Bethlehem (2009) takes values in the interval [0, 1] with the value 1 being strong representativeness and the value 0 being the maximum deviation from the strong representativeness. The following indicator was defined:

$$R(\rho) = 1 - 2SD(\rho) \ \Rightarrow \ R(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2} \ . \tag{30}$$

The minimum value of (29) depends on the response rate, it has the 0 value for $\bar{\rho}=\frac{1}{2}$ and the 1 value for $\bar{\rho}=0$ or $\bar{\rho}=1$, as there is no variation observed in the response rate then. The R-indicator may be considered as a lack of association measure. From the quality perspective, it should be discussed as a measure of extent to which the survey response deviates from the representative response. R-indicators can be used to compare representativeness of different surveys, but cannot be used for identifying subgroups that are over and under represented. However, they can be supplemented by partial R-indicators corresponding to the weak definition (Schouten, Cobben, Bethlehem, 2009).

Let us denote estimated response propensity for each element *i* as $\hat{\rho}_i$.

Let $\hat{\bar{\rho}}$ be denoted as the weighted sample average of the estimated response propensities given as:

$$\hat{\bar{\rho}}=\frac{1}{N}\sum_{i=1}^{N}\hat{\rho}_i\frac{s_i}{\pi_i}, \tag{31}$$

where the inclusion weights are applied. If $\hat{\bar{\rho}}$ is introduced to the R formula given as (30), the following partial indicator can be defined:

$$\hat{R}(\rho)=1-2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\hat{\rho}_i-\hat{\bar{\rho}})^2}\ . \tag{32}$$

It should be emphasised that representativity is considered here in the sense of the representative response concept (Schouten, Cobben, Bethlehem, 2009), not the statistical sampling theory. Especially for the surveys based on Internet sources, this approach might be satisfying in assessing quality by measuring if and to what extent answers from a given survey are representative in the context of the entire population. The main advantages of R-indicators are: a simple scale of measurement, the assessment of sample representativeness and the nonresponse bias, as well as the identification of subgroups for nonresponse follow-up. The main limitations are: the auxiliary data availability as well as the fact that comparisons require identical auxiliary variables and that threshold values are not identified. However, the R-indicators seem to be successfully used as a quality assessment tool in tandem with the response rate, and they can help to improve quality during data collection as well as help to compare data representativeness in different modes. Partial R-indicators can be used to determine which subgroup(s) are contributing the most to a lack of sample representativeness, which can significantly support the adaptive survey approach.

References

Beręsewicz M. (2015), *On the representativeness of Internet data sources for the real estate market in Poland*, "Austrian Journal of Statistics", vol. 4(2), pp. 45–57.

Beręsewicz M. (2017), *A Two-Step Procedure to Measure Representativeness of Internet Data Sources*, "International Statistical Review", vol. 85(3), pp. 473–493.

Beręsewicz M., Szymkowiak M. (2015), *Big data w statystyce publicznej – nadzieje, osiągnięcia, wyzwania i zagrożenia*, "Ekonometria", vol. 2(48), pp. 9–21.

Bethlehem J. (2009), *The Rise of Survey Sampling*, Discussion Paper 09015, Statistics Netherlands, The Hague/Heerlen.

Bethlehem J. (2010), *Selection bias in web surveys*, "International Statistical Review", vol. 78(2), pp. 161–188.

Bethlehem J. (2012), *Using response probabilities for assessing representativity*, Discussion paper 201212, Statistics Netherlands, Voorburg/Heerleen.

Bethlehem J., Biffignandi S. (2011), *Handbook of Web Surveys*, Wiley, New York.

Callegaro M., Manfreda K., Vehovar V. (2015), *Web survey methodology*, Sage, London.

Cobben F. (2009), *Nonresponse in sample surveys: methods for analysis and adjustment*, Statistics Netherlands, The Hague/Heerlen.

Cobben F., Schouten B. (2005), *Bias measures for evaluating and facilitating flexible fieldwork strategies*, Paper presented at 16th International Workshop on Household Survey Nonresponse, August 28–31, Tällberg.

Cobben F., Schouten B. (2007), *An empirical validation of R-indicators*, Discussion paper, Statistics Netherlands, Voorburg.

Cochran W.G. (1977), *Sampling Techniques*, John Wiley & Sons, New York.

Eurostat (2018), *Households – level of internet access*, http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_ci_in_h&lang=en (accessed: 20.12.2018).

Fielding N.G., Lee R.M., Blank G. (2017), *The SAGE Handbook of Online Research Methods*, Sage, London.

Fricker R., Schonlau M. (2002), *Advantages and disadvantages of internet research surveys: Evidence from the literature*, "Field Methods", no. 15, pp. 347–367.

Gemius S.A. (2018), *Wyniki badania Gemius/PBI za listopad 2018*, https://www.gemius.pl/reklamodawcy-aktualnosci/wyniki-badania-gemiuspbi-za-listopad–2018.html (accessed: 28.12.2018).

Groves R.M., Peytcheva E. (2008), *The Impact of Nonresponse Rates on Nonresponse Bias*, "Public Opinion Quarterly", no. 72, pp. 1–23.

Kalton G. (2019), *Developments in Survey Research over the Past 60 Years: A Personal Perspective*, "International Statistical Review", no. 87(S1), pp. S10–S30.

Leeuw E.D. de (2018), *Mixed-Mode: Past, Present, and Future*, "Survey Research Methods", vol. 12(2), pp. 75–89.

Polish Society of Market and Opinion Researchers (2011), *Yearbook of Polish Society of Market and Opinion Researchers 2011/2012*, Polskie Towarzystwo Badaczy Rynku i Opinii, Warszawa.

Polish Society of Market and Opinion Researchers (2012), *Yearbook of Polish Society of Market and Opinion Researchers 2012/2013*, Polskie Towarzystwo Badaczy Rynku i Opinii, Warszawa.

Polish Society of Market and Opinion Researchers (2013), *Yearbook of Polish Society of Market and Opinion Researchers 2013/2014*, Polskie Towarzystwo Badaczy Rynku i Opinii, Warszawa.

Polish Society of Market and Opinion Researchers (2014), *Yearbook of Polish Society of Market and Opinion Researchers 2014/2015*, Polskie Towarzystwo Badaczy Rynku i Opinii, Warszawa.

Polish Society of Market and Opinion Researchers (2015), *Yearbook of Polish Society of Market and Opinion Researchers 2015/2016*, Polskie Towarzystwo Badaczy Rynku i Opinii, Warszawa.

Polish Society of Market and Opinion Researchers (2016), *Yearbook of Polish Society of Market and Opinion Researchers 2016/2017*, Polskie Towarzystwo Badaczy Rynku i Opinii, Warszawa.

Polish Society of Market and Opinion Researchers (2017), *Yearbook of Polish Society of Market and Opinion Researchers 2017/2018*, ORA & Funksters, Warszawa.

Polish Society of Market and Opinion Researchers (2018), *Yearbook of Polish Society of Market and Opinion Researchers 2018/2019*, Polskie Towarzystwo Badaczy Rynku i Opinii, Warszawa.

Schonlau M., Couper M.P. (2017), *Options for Conducting Web Surveys*, "Statistical Science", vol. 32(2), pp. 279–292.

Schouten B., Cobben F., Bethlehem J.G. (2009), *Indicators for the Representativeness of Survey Response*, „Survey Methodology", vol. 35(1), pp. 101–113.

Schouten B., Bethlehem J., Beullens K., Kleven O., Loosveldt G., Luiten A., Rutar K., Shlomo N. (2012), *Evaluating, Comparing, Monitoring and Improving Representativeness of Survey Response Through R-indicators and Partial R-indicators*, "International Statistical Review", vol. 80(3), pp. 382–399.

Shlomo N., Skinner C.J., Schouten B., Bethlehem J., Zhang L. (2008), *Statistical Properties of R-indicators*, Work Package 3, Deliverable 3.1, RISQ Project, 7th Framework Programme (FP7) of the European Union.

Skinner C., Shlomo N., Schouten B., Zhang L., Bethlehem J. (2009), *Measuring Survey Quality Through Representativeness Indicators Using Sample and Population Based Information*, NTTS Conference, 18–20 February 2009, Brussels.

Statistics Poland (2018a), *Information society in Poland. Results of statistical surveys in the years 2013–2017*, https://stat.gov.pl/en/topics/science-and-technology/information-society/information-society-in-poland-results-of-statistical-surveys-in-the-years–20132017,1,4.html (accessed: 20.12.2018).

Statistics Poland (2018b), *Information society in Poland. Results of statistical surveys in the years 2014–2018*, https://stat.gov.pl/en/topics/science-and-technology/information-society/information-society-in-poland-results-of-statistical-surveys-in-the-years–20142018,1,5.html (accessed: 20.12.2018).

Statistics Poland (2018c), *Local Data Bank*, https://bdl.stat.gov.pl/BDL/start (accessed: 17.12.2018).

Szreder M. (2015), *Big data wyzwaniem dla człowieka i statystyki*, "Wiadomości Statystyczne", vol. 8(651), pp. 1–11.

Szreder M. (2017), *Nowe źródła informacji i ich wykorzystywanie w podejmowaniu decyzji*, "Wiadomości Statystyczne", vol. 7(674), pp. 5–17.

Tourangeau R., Conrad F., Couper M. (2013), *The science of web surveys*, Oxford University Press, New York.

Wybrane problemy oceny jakości w badaniach internetowych – perspektywa statystyczna

Streszczenie: W artykule przedstawiono – z perspektywy statystycznej – wybrane problemy związane z oceną jakości badań opartych na źródłach internetowych.

Dostęp do internetu konsekwentnie poszerza się na całym świecie. Równolegle, wraz z rozwojem innych nowych technologii, przestrzeń internetowa przenika coraz bardziej codzienne życie społeczeństwa, a także funkcjonowanie firm. Wszechobecny internet wywarł także wpływ na badania rynku i opinii: jako narzędzie badawcze do zbierania danych pierwotnych i wtórnych oraz w kontekście badania populacji internetowej. Ponadto, ponieważ internet i jego podmioty rejestrują wszystkie działania podejmowane w sieci, pojawiła się kwestia związana z wykorzystaniem i analizą big data i danych organicznych. W połączeniu z problemem malejących stóp odpowiedzi w badaniach i z rosnącymi ich kosztami źródła internetowe, ze względu na wiele zalet, są w powszechnym użyciu. Coraz szersze wykorzystanie internetu i jego zasobów wydaje się nieuniknione. Należy jednak podkreślić, że w praktyce proces realizacji badań na podstawie źródeł internetowych wyprzedził prace metodologiczne. Można wskazać wiele problemów, szczególnie w kwestii jakości uzyskiwanych danych. Artykuł prezentuje wybrane z nich, istotne zwłaszcza z punktu widzenia statystyki: kwestie związane z poprawnym zdefiniowaniem operatu losowania, samodoborem, nadmiernym/niedostatecznym pokryciem i powiązanymi z nimi obciążeniami estymatorów.

Słowa kluczowe: badanie internetowe, badanie on-line, jakość badania, błędy w badaniach

JEL: C8

Received: 2019-01-18; verified: 2020-02-24. Accepted: 2020-09-01