Proceedings of Q2006
European Conference on Quality in Survey Statistics

Data Driven Identification of Sources of Errors for Improving Survey Quality

Christin Schäfer, Hartmut Bömermann, Ricarda Nauenburg and Karsten Wenzel[1]

1. Introduction

1.1 Motivation

Official statistics is faced with the dilemma to ensure a high data quality by working under a high pressure of time and under the restrictions that follow from decreasing budgets. To find a trade-off between high quality, time and money is *the* challenging task.

Actions that ensure the quality of a survey are taken prior and posterior the conduction of the survey. Prior decisions have to be made e.g. about the most appropriate frame, the interview mode, the design of the questionnaire, furthermore interviewer have to be trained, software and hardware must be prepared and so on. After the survey the data are validated with respect to a predefined set of rules. For inquiry and verification of given or missed answers interviewees are contacted.

The focus of this work lies on the posterior strategies for quality control. Most of the standard methods used for this task consume a high amount of time, like the effort that has to be made when contacting an interviewee again. At the same time this procedure is expensive by blocking many time of the employees of the bureaus of statistics. Furthermore it improves only one aspect of the survey data, that is the completeness. Many other discrepancies can not be detected in this way.

We propose to use the given full data set and apply some data analysis methods to find hints for sources of error in the whole survey process from the data collection with all aspects of field work including possible differences between different questionings, to problems introduced by the software and the transfer from analogue to digital data. The used methods can be implemented as part of the standard data pre-processing, the elapsed time is short, thus we offer a cheap, automatic and fast tool for quality control.

[1] Christin Schäfer, Fraunhofer Institute FIRST.IDA, Kekuléstr.7, Berlin, Germany, 12435 (christin.schaefer@first.fraunhofer.de), Hartmut Bömermann, Ricarda Nauenburg and Karsten Wenzel, Statistisches Landesamt Berlin, Alt-Friedrichsfelde 60, Berlin, Germany, 10315.

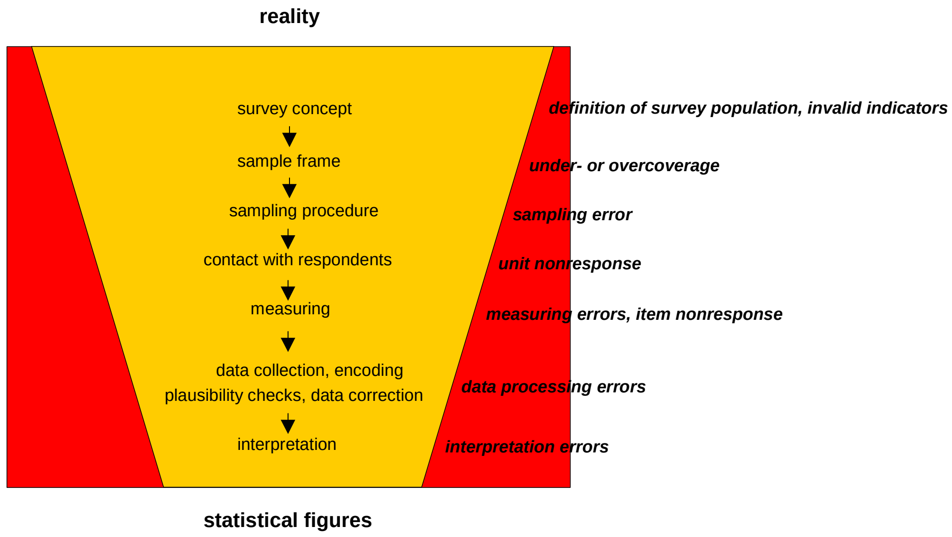

Figure 1: types of errors in surveys following a concept of Radermacher and Körner (2005).

1.2 Quality control in surveys

Quality control and quality assurance in surveys -- especially in official statistics -- have gained a lot of attention in the recent years. The decrease in time and money on the one hand, the increase in requirements on the other have accelerated this development. The conferences Q 2001 in Stockholm, Q 2004 in Mainz and Q 2006 in Cardiff display the importance and the need of quality control in modern official statistics. The conference proceedings give an overview on all aspects of the topic.

The objective is to guarantee the quality of the survey data to be controlled on a high level. The problem with this goal is that there exists no definition of quality that everyone agrees on. How to control quality that is not even well defined? One workaround is to concentrate of the detection of possible sources of errors. Every error has an influence on the quality of the data. Therefore the detection of errors is a first step in a quality control process. The next steps are the actions to be taken when a certain error occurs. The decision about the most appropriate action can not be learnt from the data; the survey team must take it.

There are many other possible sources of errors that can occur during the run of a big survey. Figure 1 presents the general stages of a survey and specifies typical errors at every stage that have a negative effect on the data quality.

1.3 Micro census Berlin 2004

The Micro census is a representative one per cent sample survey of the German population. It is conducted every year since 1957 as the biggest national multi-purpose survey. The micro census offers not only information on the structure of the population, their economic and social situation, their education, and their position in the labour market, but is the basis for adjusting sample designs and weighting schemes of many other national surveys. It is a widely exploited source of information for politicians and scientists. There is another point that underlines the importance of the micro census: in Germany, the EU Labour Force Survey is integrated into the micro census questionnaire (Schwarz 2001). A high data quality is therefore essential for the German micro census (Statistische Ämter des Bundes und der Länder, 2003, Statistisches Bundesamt, 2005, Quality Concept for Official Statistics: Encyclopaedia of Statistical Sciences).

By now, quality assurance for the micro census survey takes place at several stages of data sampling. For example, respondents being in the sample are obligated by law to give information on almost all survey questions. This considerably decreases the share of total or partly refused interviews. Since it is a common fact that face-to-face interviews are of higher quality than telephone interviews or self-filled questionnaires, an interviewer will have to make up to three visits in an attempt to conduct the interview in case the selected person to interview is not at home. If there is no possibility for a personal contact, members of the micro census staff try to call the respective respondents or send them a questionnaire for self-completion. Paper-pencil-interviews (PAPI) are step by step substituted by computer aided personal interviews (CAPI). The questionnaire is now stored on a laptop and a software takes care for the interview filters or the range of possible answer codes. The so-called „Blaise“ software helps the interviewer to follow complicated filter questions correctly, and eliminates the bulk of typing errors. Furthermore, it avoids another error source, because the data entry from paper to electronic media is already done during the interview. Thus it can be assumed that CAPI interviews provide the highest data quality. After being collected, the data undergo several plausibility controls to detect inconsistencies. In such cases the data are made consistent either by a computer program or the respondents are asked again by phone to confirm their answers. However, most of these currently employed methods of quality assurance are costly in time and manpower.

The project used the Berlin data of the micro census 2004. In Berlin, every year about 18000 households and 34000 persons are asked to answer the survey. Around 170 interviewers visit the respondents to carry out the face-to face interviews. Figure 2 shows the scheme of data collecting applied in Berlin.

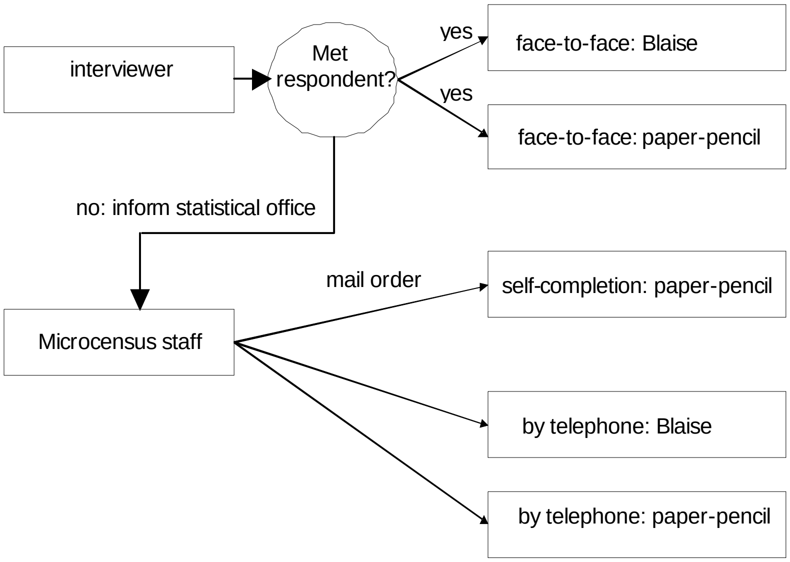
interviewer
Met respondent?
yes
face-to-face: Blaise
yes
face-to-face: paper-pencil
no: inform statistical office
mail order
self-completion: paper-pencil
Microcensus staff
by telephone: Blaise
by telephone: paper-pencil

Figure 2: Berlin micro census: scheme of data collection.

2. Method

2.1 Idea

The conduction of a survey - seen as a process - is quite complex, consists of many steps and involves specific methodologies and guidelines. All steps of this process have been subject of extensive scrutiny, and an impressive corpus of specific *savoir-faire* is available for each of them. There are many aspects taken into consideration for careful survey planning (e.g. steps to ensure correct population sampling etc.). In other words, the *a priori* quality control is already in place.

However, since it is impossible to give a guarantee that even the best defined process runs without any error, there is the need to assess quality *a posteriori* the survey and to identify the sources of error and the process steps at which they had occurred. This information can then be passed to field knowledge experts, that can analyze and interpret the significance of the results and draw consequences on the data collecting procedure. Therefore, quality assurance is a constant cooperation between *a priori* and *a posteriori* controls.

As mentioned above our focus lies on a posteriori quality control. The question occurs: how to assess the a posteriori quality? There is no clearly established quality criterion to check for the quality of the final survey data. This seems to be inherent to the very nature of surveys: the quality of the survey is, in vague terms, defined by how truthful the end survey data is to the target population. The quality here concerns the *quality of information*. But to check the validity of the survey information, we would need some reliable, reference information available in the first place about the target of the survey.

Nevertheless it is still possible to perform a relevant a posteriori quality control of the survey data *using only the survey output data itself*. This can be achieved by checking *indirect* criteria, i.e., we do not verify directly the quality of the information on the target population, but we concentrate in detecting `symptoms' of flaws in the data collection of processing.

In contrast to validation, where every single data point can be partly modified to ensure certain rules are satisfied that are supposed to improve quality, we tackle a different question, whose focus is not to pinpoint problems or aberrations at the level of a single data point, but rather to detect specific circumstances of the data collection and processing under which the resulting data is of poor quality. In effect, the detected anomalies can then concern whole subsets of the survey data. The goal is to identify the source of the anomaly in the data processing chain.

To this end we propose to use meta-data, that is additional data that comes with each data point and reflects the way this data point was collected (typically, geographical information, interview mode, etc ...). For example, in idealized data collecting, the data itself should be independent of the interview mode. If there is a flaw with a specific interview mode so that the associated data is biased or otherwise corrupted, we would like to detect it and identify this interview mode.

Below, we focus on checking meta-data that should not have any influence on the final data and expose a methodology to detect and identify sources of deviation.

2.2 General framework

Let us represent one data point of the survey in an abstract way as a couple of random variables (X,Y), where X is the survey data output *per se* and Y is some additional information - some variable from the set of meta-data. Let us assume that (X,Y) is drawn according to a joint distribution P from which we have obtained a finite sample $S=\{(X_i,Y_i), i=1,...,N\}$. The distribution P represents the result of the entire data collecting process. As such, this includes randomness coming from the initial population, as well as additional noise or distortion.

The test idea is to check whether or not there is a dependency between X and Y. Under the null hypothesis, that is the ideal case of flawless data collection, the distribution of X should be independent of Y. Assuming Y is a discrete variable, the problem reduces to comparing different conditional distributions $P(X|Y=y)$. Using the available data, one is interested in comparing different subsamples of the form $S_y = \{(X,Y) \in S, Y=y\}$, more precisely, the goal is to test whether all subsamples S_y follow the same distribution. Given the generality of our framework, there is a vast choice of methods that can be applied to achieve the wanted goal.

3. Results

The reported results are obtained by a first preliminary study of the conceptual idea. Nevertheless the impact of the results is very high. For the investigation we choose the meta-variable Y = `interview mode'. As depicted in Figure 2 we have to distinguish between four different modi: PAPI and CAPI, self-completion and telephone interview, combining PAPI and CAPI telephone interviews to one mode.

Since the data set does not contain any metric variables it is reasonable to pre-process the data as follows: For every variable we introduce as many new variables as different values of the variable are observed in the data. For every data point the one new dimension that codes the observed value is set to one, while the others are zero. One further dimension codes missing values.

Now the test problem is to compare the parameters of binomial distributions. That is, for every new dimension we first conduct a global test over the four populations introduced be Y = `interview mode'. If we have to reject the null hypothesis of equal binomial parameters, we start a pairwise comparison. From previous studies it is known, that due to a sampling bias the results for telephone and self-completion interviews differ from face-to-face interviews. Therefore, we restrict the pairwise comparison to the PAPI vs CAPI case.

Note, that with this approach one has to face all problems connected to multiple testing. But since our aim is to get a ranking of possible sources of error, to invest time and money most efficient, we accept some false discoveries.

After running the test procedure for all new dimensions, we select those with rejected global null hypothesis. These dimensions are sorted by the value of the test statistic for the PAPI vs CAPI comparison, from large to small values, yielding a ranking of potential sources of error.

Together with the micro census faculty we could now identify sources of errors in the following domains:

- *Incomplete coding scheme*: Due to incompleteness in the coding scheme for some questions it is e.g. impossible to distinguish between the answer `inapplicable' and `no answer'.
- *Inconsistent coding scheme*: For example the standard code for the answer `no' is the digit `8'. Questions asking for some quantity, like the number of children, code `no' with the digit `0'. We found a surprisingly high number of people with 8 children in the data set.
- *Unclear formulations of questions*: In Berlin some questions ask for the `part' of Berlin, that is `East' or `West', while others focus on the federal state like `Berlin' or for example `Brandenburg'. The formulation of some questions is ambiguous, leaving

room for many errors.

- *Dynamic filter guidance*: We found an improvement of the interviews completeness with the Blaise-program where dynamic filter guidance is offered in comparison to the face-to-face interviews, that is lead by the interviewer.
- *Different quality of plausibility*: A certain percentage of respondents are ask a further EU-supplement program. The additionally given answers can be used to validate answers from the standard program, leading to different qualities in the plausibility for the standard questions.
- *Difference in interpretation and differentiation* between leaded face-to-face interviews and such filled out by the interviewee on its own.

4. Conclusion

To improve the quality of the micro census 2004 and learn about flaws in the survey process we applied a data driven procedure taken place a posteriori the survey. Even if we have run the methods in the most simple way, the value of the outcome of the investigation can not be underestimated. The results of the analysis are used to improve the surveys quality in many directions in a quite efficient way. Furthermore we can define new quality assurance guidelines for the survey process.

The Berlin Statistical Office has signalised its intention to order a professional software tool that works on the basis of the new method.

References

Diekmann, A. (2002). *Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung*. Manuskript 06/2002, Institut für Technikfolgen-Abschätzung (ITA), Wien.

Koch, A. (1995). *Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994*. ZUMA Nachrichten, 36: 89–105.

Lynn, Peter (2004), *Editorial: Measuring and communicating survey quality.* Journal of the Royal Statistical Society A,167 (4): 575–578.

Marshall, Eliot (2000): *How Prevalent Is Fraud? That's a Million-Dollar Question*. Science magazine, 290: 1662–1663.

Radermacher, Walter, Körner, Thomas (2005). *Fehlende und fehlerhafte Daten in der amtlichen Statistik – neue Herausforderungen und Lösungsansätze*, Statistische Woche 2005, Braunschweig

Reuband, K.-H. (1990). *Interviews, die keine sind - 'Erfolge' und 'Misserfolge' beim*

Fälschen von Interviews. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 4:706–733.

Schäfer, C., J.-P. Schräpler, K.-R. Müller, G.G. Wagner (2005). *Automatic Identification of Faked and Fraudulent Interviews in the German SOEP*, Schmollers Jahrbuch, Duncker & Humblot, Berlin, 125: 183–193.

Schnell, R. (1991). *Der Einfluss gefälschter Interviews auf Survey-Ergebnisse*. Zeitschrift für Soziologie, 20(1):25–35.

Schräpler, J.-P., G.G. Wagner (2005). *Characteristics and Impact of Faked Interviews in Surveys - An analysis of genuine fakes in the raw data of SOEP*. Allgemeines Statistisches Archiv, 89 (1): 7 - 20.

Schwarz, N. (2001). *The German Microcensus.* Schmollers Jahrbuch, Duncker & Humblot, Berlin, 121: 649-654.

Statistische Ämter des Bundes und der Länder (2003). *Qualitätsstandards in der amtlichen Statistik* (URL: //www.destatis.de/allg/d/ueber/q_stand.htm).

Statistisches Bundesamt (2005): Qualitätsbericht Mikrozensus (URL: //www.destatis.de/download/qualitaetsberichte/qualitaetsbericht_mikrozensus.pdf)

Quality Concept for Official Statistics: Encyclopedia of Statistical Sciences. John Wiley & Sons, New York, Update Vol. 3: 621-629.