Arijit Chaudhuri · Tasos C. Christofides

Indirect Questioning in Sample Surveys

Springer

Indirect Questioning in Sample Surveys

Arijit Chaudhuri • Tasos C. Christofides

Indirect Questioning in Sample Surveys

Springer

Arijit Chaudhuri
Applied Statistics Unit
Indian Statistical Institute
Kolkata, India

Tasos C. Christofides
Department of Mathematics and Statistics
University of Cyprus
Nicosia, Cyprus

ISBN 978-3-642-36275-0 ISBN 978-3-642-36276-7 (eBook)
DOI 10.1007/978-3-642-36276-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013946203

Mathematics Subject Classification (2010): 62D05, 62P25

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Bulu
AC

To Liana, Andrea, Christoforos
TCC

Preface

Asking questions about sensitive and stigmatizing characteristics in surveys of human populations is not an easy matter. Gathering information on issues like sexual orientation, drunkenness, HIV positivity, experience in induced abortion, maltreatment of spouse, habits of wilful tax evasion, bribery, cheating, and fraud by means of direct questions and conventional survey methodology is likely to produce large nonsampling errors particularly due to nonresponse. People are not willing to provide information which might be considered as incriminating and stigmatizing. In cases they agreed to participate in such a survey, it is very reasonable to assume that many of them give false answers and provide misleading information.

Warner (1965) was the first to offer a way out as a pioneer with his Randomized Response Technique. A participant in a survey employing his technique, using a so-called randomization device, provides information from which it is not possible to infer whether he/she has the stigmatizing characteristic and thus his/her privacy is protected. However, based on the information collected from all participants, it is possible to make inferences about the prevalence of the stigmatizing attribute. This principle, namely that the information provided by a participant is not adequate to make inferences about his/her status as related to the sensitive characteristic but the information collected from all participants together is sufficient to estimate certain parameters of the population, is the one which governs all indirect questioning techniques devised so far.

Prospective readers may be familiar with the three treatises, namely (1) Randomized Response and Indirect Questioning Techniques in Surveys (Chapman & Hall, CRC Press, Boca Raton, Florida, USA, 2011) by Arijit Chaudhuri, (2) Randomized Response: Theory and Techniques (Marcel Dekker, NY. USA, 1988) by Arijit Chaudhuri and Rahul Mukerjee, and (3) Randomized Response: A method for Sensitive Surveys (Sage, London, 1986) by J.A.Fox and P.E.Tracy.

Warner and most of his followers did not clarify if their theories are related to a theoretical or a survey population of labeled individuals. Consequently most of the published works including (2) and (3) above dealt with analysis confined to simple random sampling with replacement alone. A few published papers and Chap. 7 in Chaudhuri and Mukerjee (1988) considered labeled finite survey populations

and general sampling schemes allowing selection without replacement and even selection with unequal or varying probabilities. The monograph (1) noted above provides a comprehensive review opening an avenue for further research in theory and practice in randomized response. It is a research publication out and out. Its emphasis is on thrashing out the point that for every randomized response technique employed in respect of the people selected in a sample, no matter how, data analysis is possible to throw up unbiased estimators for the proportion of people bearing a sensitive attribute in a community throwing up estimated measures of accuracy in estimation only provided that every person is given a positive inclusion-probability in a sample and that every pair of distinct persons also has a positive inclusion-probability in a sample. Chaudhuri (2011) and Chaudhuri and Mukerjee (1988) covered estimation of survey population totals of stigmatizing variables. In addition, taking account of certain emerging criticisms of randomized response techniques in general, alternative data-gathering procedures in indirect manners are also briefly studied by Chaudhuri's (2011) text.

However, recognizing that the monographs above involve a good deal of analytical sophistication not quite tasteful to social scientists enjoying less pleasure in their perusal but really more interested in the essentials of these Indirect Techniques for gathering sensitive data, the present monograph attempts at presenting a compendium of useful techniques with straightforward analytical tools in rather condensed forms. Although randomized response techniques account for the lion's share of indirect questioning, more recent approaches move away from the idea of using a randomization device. This monograph attempts to give the most basic and important aspects of indirect questioning. In addition to randomized response and other indirect questioning approaches such as the item count technique, the nominative technique, and the three-card method which have been known for quite some time, this monograph contains modern approaches such as non-randomized techniques and surveys with negative questions not to be found in any of the three monographs mentioned above.

In this book, the issue of the protection of privacy has a prominent place. But here we just do not view it as a concept on which one builds mathematical formulas and numerical indicators. We put emphasis on the perceived protection of privacy, i.e., the protection of privacy as how participants perceive it. Although the book does not offer any solutions to the issue of quantification of the perceived protection of privacy, we firmly believe that it will provide incentives to researchers, in particular social scientists to join forces with mathematical statisticians on this important issue.

Kolkata, India — Arijit Chaudhuri
Nicosia, Cyprus — Tasos C. Christofides
July 2013

Acknowledgments

I am indebted to the Director, Indian Statistical Institute and my colleagues in Applied Statistics Unit for offering me certain requisite spare time to complete writing this monograph.

Arijit Chaudhuri

I am indebted to the University of Cyprus for creating the academic environment that allows faculty members to author manuscripts such as this one.

Tasos C. Christofides

Contents

Chapter 1
A Plea for Indirect Questioning: Stigmatizing Issues of Social Relevance

Abstract Collecting data on human populations by means of sample surveys is not an easy task. Survey practitioners often experience difficulties in collecting reliable data due to various sources of nonsampling error and in particular due to nonresponse. In case the issues under investigation are of sensitive nature, such as issues on sexual orientation, tax evasion, or involvement in criminal activities, people are reluctant to participate, and even if they agree to participate, false or misleading answers are given by many of them. Indirect questioning techniques offer a solution to this problem. These are techniques designed in such a way that the information provided by a participant is not incriminating and thus his/her privacy is protected. However, based on the information collected from all participants, the investigator is able to estimate parameters of interest related to the sensitive characteristic. In this chapter we make a case in favor of the use of indirect questioning techniques. We briefly discuss hypothetical as well as real examples where the methodology presented in this book can be implemented.

1.1 Introduction

With time advancing, human civilization is rapidly progressing. Keeping pace with it, many social taboos are quickly disappearing. Yet, the society seems not to be permissive enough. Many practices are still found not to capture social approbation. For example, social scientists deem it discourteous to ask a stranger chosen in a sample if he/she is a habitual gambler or a tax evader or an exorbitantly drunken driver of a motor vehicle or engaged in any one or more similar illegal and/or unethical practices. Overcoming the delicacy, even if one plucks enough courage to put up a brave face to enquire about such traits in a chosen respondent, honest answers are frequently in short supply. People fight shy and either refuse to answer, or the responses often are suspected to be far from the truth in their revelations. Warner (1965) first published a technique of indirect questioning. This inaugurated an era of fruitful coverage of data on sensitive items in meaningful

A. Chaudhuri and T.C. Christofides, *Indirect Questioning in Sample Surveys*,
DOI 10.1007/978-3-642-36276-7_1, © Springer-Verlag Berlin Heidelberg 2013

studies. In addition to asking questions of sensitive nature, nowadays the issue of personal data protection makes it necessary to employ techniques which guarantee that no would be in a position to make inferences about the status or personal data of an individual, even if the status in question is not stigmatizing at all.

Our purpose is to develop a handbook of procedures to estimate parameters relating to items bearing social stigmas for human subjects. It is intended to be a compendium on how to gather sensitive information in sample surveys from persons by asking indirect questions or by employing certain techniques that essentially mask one's answer. The use of indirect questioning is for the sole purpose of protecting a respondent's privacy and thus enhancing the chances that the respondent would be willing to participate in such a survey and provide honest answers. It is reasonable to assume that on sensitive issues like tax evasion, sexual orientation, gambling, student academic dishonesty, illegal drug use, or criminal activities, people are reluctant to reveal information. Interviewer's assurances that the information furnished would be treated as strictly confidential are just not enough. Even in cases that one agrees to participate in a survey on sensitive issues, there is no guarantee that the information provided is correct. It is very human that people would provide untruthful answers just to be on the safe side. It is for this reason that the need for indirect questioning techniques arises.

Warner (1965) is the first researcher who came up with such a technique termed the Randomized Response Technique (RRT). Assume that by A we denote the sensitive or stigmatizing characteristic. Each person picked up at random is offered a box full of a number of cards identical in shape, size, color, weight, thickness, and in every other possible respect, but a fraction p ($0 < p < 1$, $p \neq 0.5$) of them are marked as A and the rest marked as A^c, the complement of A. The person is requested, outside of view of the interviewer, to randomly draw a card from the box out, after thoroughly shaking it and to truthfully say "Yes" if the mark on the card picked coincides with his/her status about the sensitive characteristic, i.e., to say "Yes" if he/she belongs to the sensitive group and the card picked up is marked A, or if he/she does not have the sensitive attribute and the card picked up is marked as A^c. Otherwise the respondent must respond "No." The respondent is of course not to divulge the card type to the enquirer and he/she is advised to put the card back to the box after truthfully declaring "Yes" or "No" to say if his/her real trait "matches" the card type drawn or "does not match" it. Hopefully the person so addressed is supposed to cooperate because the enquirer cannot be sure if "Yes" is the reply from a person bearing A or bearing the complement A^c as a matter of fact. It is important to emphasize that clear instructions must be given to the participants before they apply any randomized response procedure, or any indirect questioning technique for that matter. In addition, one should make sure that the participants are convinced that the procedure protects their privacy and their status related to the sensitive characteristic. Here we may add that the randomization device does not have to be a box of cards such as the one described above but could be any other device which can be used in such a way so that the respondent responds (with a "Yes" or "No") with probability p to the statement

(I) I have the characteristic A

and with probability $1 - p$ to the statement

(II) I have the characteristic A^c.

Such a device could be a standard deck of cards or even a fair coin or fair die appropriately used. Based on the responses obtained from all the participants, the person in charge of the survey is able to provide estimates for the prevalence of the stigmatizing characteristic as well as other measures associated with it.

Warner's technique has been followed by numerous other procedures. In all such cases, the objective remains the same: To estimate quantities related to sensitive attributes and at the same time to protect the privacy of the participants.

1.2 Real and Hypothetical Examples to Justify the Need for Indirect Methods

In the Netherlands, Scheers (1992), Kerkvliet (1994), van der Heijden and van Gils (1996), van der Heijden, van Gils, Bouts, and Hox (2000), Umesh and Peterson (1991) among others, like Maddala (1983), have been working long in examining efficacies of rival competitive survey techniques of specific nature. Those techniques were aiming at gathering useful data relating to sensitive issues and estimating proportions of people in the communities with propensities to indulge in practicing illegal, immoral, and unlawful practices, or practices considered to be having some cost, for instance not supporting the regime in a dictatorship.

van der Heijden et al. (2000) discuss the following case. From Police files, information was gathered about the people enjoying unemployment benefits while being not eligible. The curiosity was about how many of them would admit, on enquiry, of their complicity in this offensive act. Another couple of stigmatizing habits they considered were students' consumption of marijuana and cheating in examinations. The survey techniques they illustrate as employed are (1) Face-to-Face interviewing by the investigators, (2) Computer-Assisted Self-Interviewing, and (3) RRTs introduced by Warner (1965), the Unrelated Question Randomized Response Model by Abul-Ela, Greenberg, and Horvitz (1967), and Greenberg, Abul-Ela, Simmons, and Horvitz (1969), Kuk's (1990) Randomized Response Technique and the Forced Response Randomized Response Technique introduced by Boruch (1972). In order to improve upon the efficiency levels and also to identify factors that induce truthful answers to queries, they also made use of covariates like age, sex, racial trait, literacy levels, following Maddala (1983) among others.

In a recent study, Dietz et al. (2013) used an RRT approach to estimate the prevalence of cognitive enhancing drug use among university students in Germany. Based on their findings they argue that direct survey techniques used in the past have underestimated the use of those drugs. In a related study, Franke et al. (2013)

by means of the same RRT used in Dietz et al. (2013), estimated the prevalence of pharmacological cognitive enhancement or mood enhancement drugs among surgeons. It is known that surgeons often make use of such drugs in order to combat fatigue, distress and concentration deficits. But it is also known that this particular drug use may lead to addiction and overestimation of the surgeon's own capabilities, thus putting patients at risk.

Kuha and Jackson (2013) analyzed data on the illegal behavior of buying stolen goods. Data were obtained by applying the Item Count Technique with the use of an "item count question" included in the Euro-Justis Survey. The Item Count Technique, to be presented in Chap. 6, seems to be gaining a lot of momentum and it appears to be popular among social survey practitioners.

Karlan and Zinman (2012) have also implemented the Item Count Technique in order to estimate how clients of microfinance institutions spent their loan proceeds, thus providing an application of indirect questioning in the area of economics.

Jan, Jerke, and Krumpal (2012) used the Crosswise Model, one of the so called non randomized response models to measure plagiarism at Swiss and German universities. On the issue of student plagiarism is the paper of Coutts, Jann, Krumpal, and Naeher (2011) where three indirect questioning techniques, the RRT, the Item Count Technique and the Crosswise Model are evaluated for measuring the prevalence of plagiarism in student papers. The academic disintegrity of students and in particular medical students is measured by means of the RRT in Hejri, Zendehdel, Asghari, Fotouhi, and Rashidian (2013).

Ecology and the environment are areas where indirect questioning techniques have found application. In John, Edwards-Jones, Gibbons, and Jones (2010) two indirect questioning techniques, the RRT and the Nominative Technique are presented as methods to estimate the prevalence of rule breaking in conservation. In Blank and Gavin (2009) the RRT is used to estimate the extend of illegal fishing in Northern California.

The prevalence of illegal drug use by professional athletes is not easy to measure by conventional survey research methods. Thus, indirect questioning techniques have been used instead. Striegel, Simon, Hansel, Niess, and Ulrich (2006), Striegel, Ulrich and Simon (2010) and Pitsch, Emrich, and Klein (2007) use the RRT to measure the prevalence of doping among elite athletes. In Chap. 6, we provide model questionnaires for the Item Count Technique which could be used for the same purpose.

In some cases, an opinion expressed even in modern democratic societies might be considered of sensitive or stigmatizing nature. For example a person may have difficulties expressing his/her opinion about an ethnic group or another group (different from his/her own) in the same society. Social scientists find it convenient to employ indirect questioning techniques to gather information on how members of a certain group view another group. Research on racism, sexism or xenophobia may find indirect questioning techniques as an invaluable tool. In a recent study in Germany, described in Krumpal (2012), it is documented that Randomized Response is an effective technique eliciting socially undesirable opinions and

provides more accurate prevalence estimates of xenophobia and anti-Semitism than direct questioning.

Quantitative characteristics like number of induced abortion experienced so far by women interviewees, amounts gained or lost in gambling, amounts underreported in income tax returns, amounts surreptitiously earned in excess of legitimate earnings through kickbacks and bribes, numbers of days of drunken driving, amounts spent on items shameful enough to be hidden from the spouses, are some of the quantitative features socially needed to be examined, if actually rampant in a civil society. Many more of course may also be easily named. Indirect questioning tactics seem to be necessary and should be adequately explored by the social scientists indeed.

Not much is known about actual coverage of successfully applied procedures in statistical estimation of parameters relating to quantitative sensitive procedures. However, extensive theoretical research is known to have been carried out over the years to cover such issues. One can mention the work of Greenberg, Kuebler, Abernathy, and Horvitz (1971), Sen (1974), Chaudhuri (1987), Arnab (1995), Singh, Mahmood and Tracy (2001), Bar-Lev, Bobovitch and Boukai (2004), Huang, Lan and Kuo (2006), Saha (2008), Pal (2008), Bouza (2009) and Diana and Perri (2011) among others. Chaudhuri's (2011) monograph can be used as a reference for the case of quantitative sensitive attributes.

Whatever is presented so far in this introductory chapter may appear nebulous. However, we believe that things will become clear and the importance of the methods presented in this monograph will be greatly appreciated by social survey practitioners and mathematical statisticians.

References

Abul-Ela, Abdel-Latif, A., Greenberg, B.G., Horvitz, D.G. (1967). A multi-proportion randomized response model. *Journal of the American Statistical Association*, *62*, 990–1008.

Arnab, R. (1995). Optimal estimation of a finite population total under randomized response surveys. *Statistics*, *27*, 175–180.

Bar-Lev Shaul, K., Bobovitch, E., Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, *60*, 255–260.

Blank, S., & Gavin, M. (2009). The randomized response technique as a tool for estimating non-compliance rates in fisheries: a case study of illegal red abalone (*Haliotis rufescens*) fishing in Northern California. *Environmental Conservation*, *36*, 112–119.

Boruch, R.F. (1972). Relations among statistical methods for assuring confidentiality of social research data. *Social Science Research*, *1*, 403–414.

Bouza, C.N. (2009). Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character. *Metrika*, *70*, 267–277.

Chaudhuri, A. (1987). Randomized response surveys of finite populations: a unified approach with quantitative data. *Journal of Statistical Planning and Inference*, *15*, 157–165.

Chaudhuri, A. (2011). *Randomized response and indirect questioning techniques in surveys*. Boca Raton: Chapman & Hall, CRC Press, Taylor & Francis Group.