JOHANNES KEPLER
UNIVERSITY LINZ

06-01-2016, MADRID

The Privacy Protecting Aspect of Indirect Questioning Designs

Andreas Quatember

1. Introduction

Indirect Questioning Designs have been developed as alternatives to the common direct questioning technique

In particular, for surveys on sensitive subjects such as alcoholism, doping, illegal employment, harassment at work, domestic violence and so forth

Such subjects lead to a further increase of nonresponse and the occurrence of untruthful answering

An example of a parameter of interest may be the proportion π_A of people of a population U of size N belonging to a subpopulation U_A bearing a certain attribute A

The Horvitz-Thompson (HT) based estimator of proportion π_A is given by

$$\pi_{A,HT} = \frac{1}{N} \cdot \sum\nolimits_s y_k \cdot d_k \,.$$

with y_k indicating unit k's membership of U_A and design-weights d_k ($k \in s$).

The theoretical variance of $\pi_{A,HT}$ is given by

$$V(\pi_{A,HT}) = \frac{1}{N^2} \cdot \sum\sum\nolimits_U \Delta_{kl} \cdot y_k \cdot d_k \cdot y_l \cdot d_l$$

(cf. Särndal et al. 1992, *Model Assisted Survey Sampling*, p.43).

Untruthful answering and nonresponse split up the set s of sampling units into three different sets: s_t, s_u, s_m

For s_u and s_m being non-empty sets,

$$\pi_{A,HT} = \frac{1}{N} \cdot \sum_s y_k \cdot d_k = \frac{1}{N} \cdot \left(\sum_{s_t} y_k \cdot d_k + \sum_{s_u} y_k \cdot d_k + \sum_{s_m} y_k \cdot d_k \right)$$

applies.

This means that parts of the y-values needed for $\pi_{A,HT}$ are not available

The best way of dealing with the problem is to avoid nonresponse and untruthful answering

After data collection, the statistical methods of weighting adjustment and/or data imputation can be applied

2. A Generalized Randomized Questioning Design

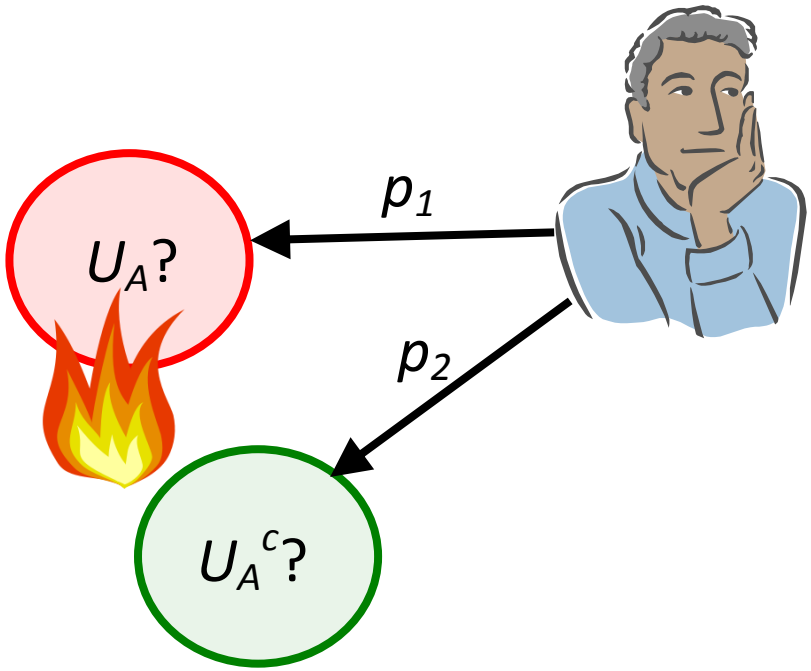
Alternatives to the direct questioning may be considered to reduce both rates through an increase in privacy protection: Indirect Questioning techniques such as Randomized Questioning Designs (RQ)

In an RQ-Design, respondents reply to only one of different questions or instructions, of which only one addresses the sensitive attribute

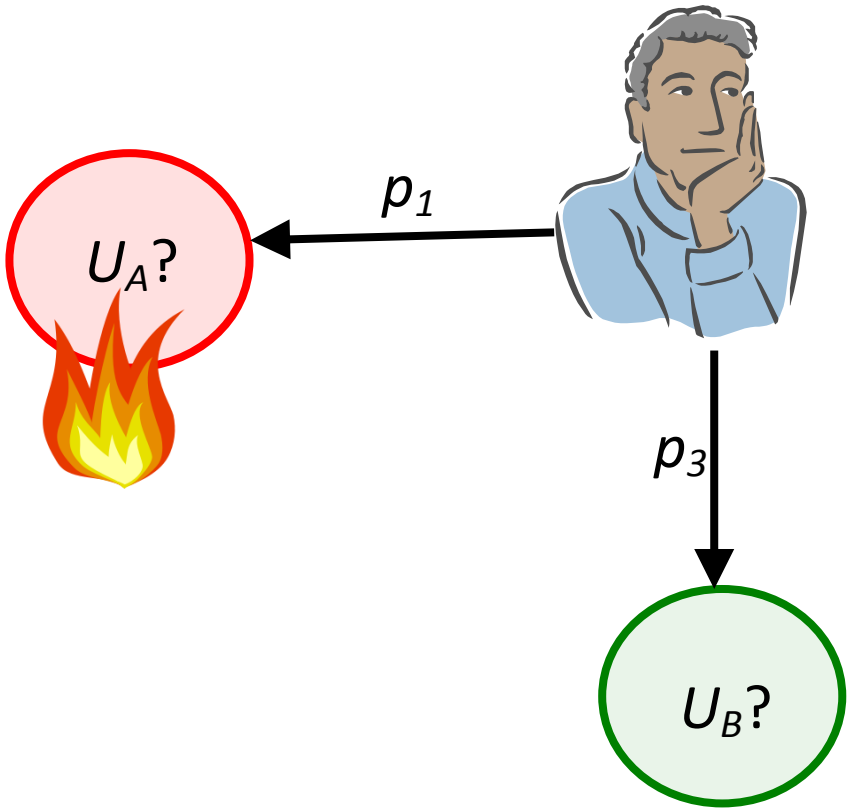
The question actually asked is randomly chosen according to an applied randomization device

Hence, neither the interviewer nor the experimenter knows on which question the answer was actually given

The known probabilities of asking the different questions/instructions allow to make inferences about unknown parameters of interest

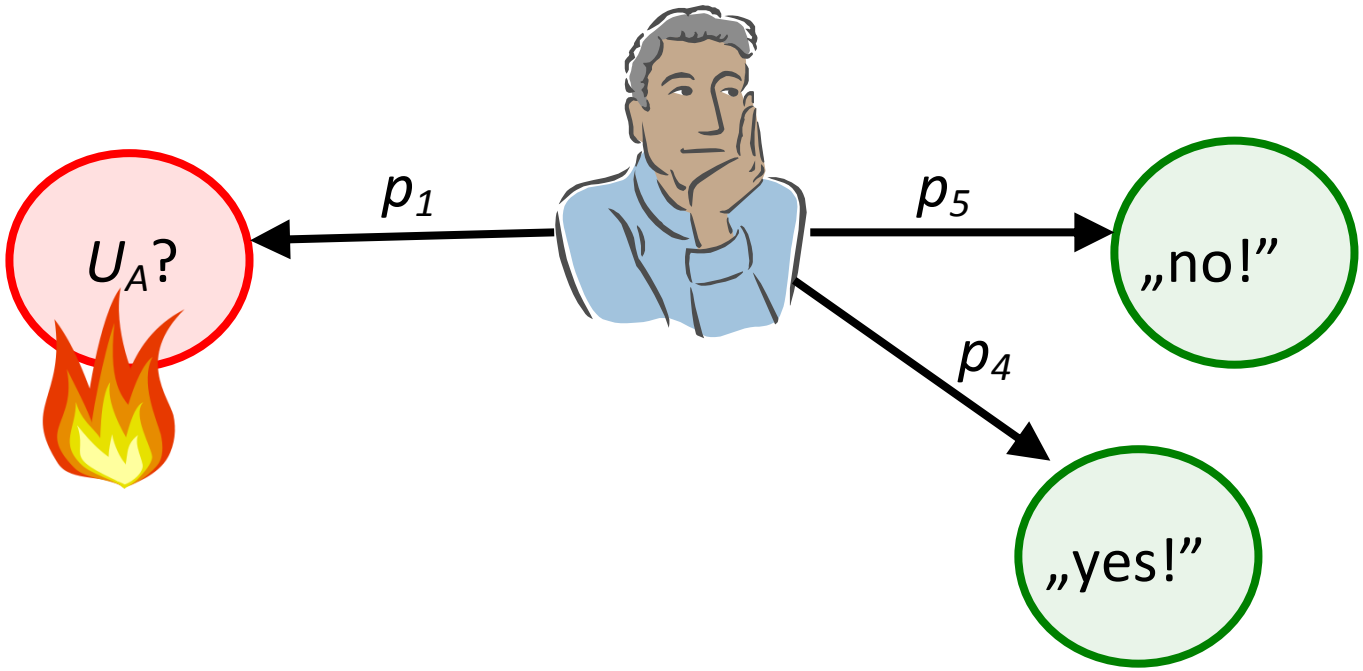


Design parameters: p_1, p_2 (Warner 1965, *JASA*)

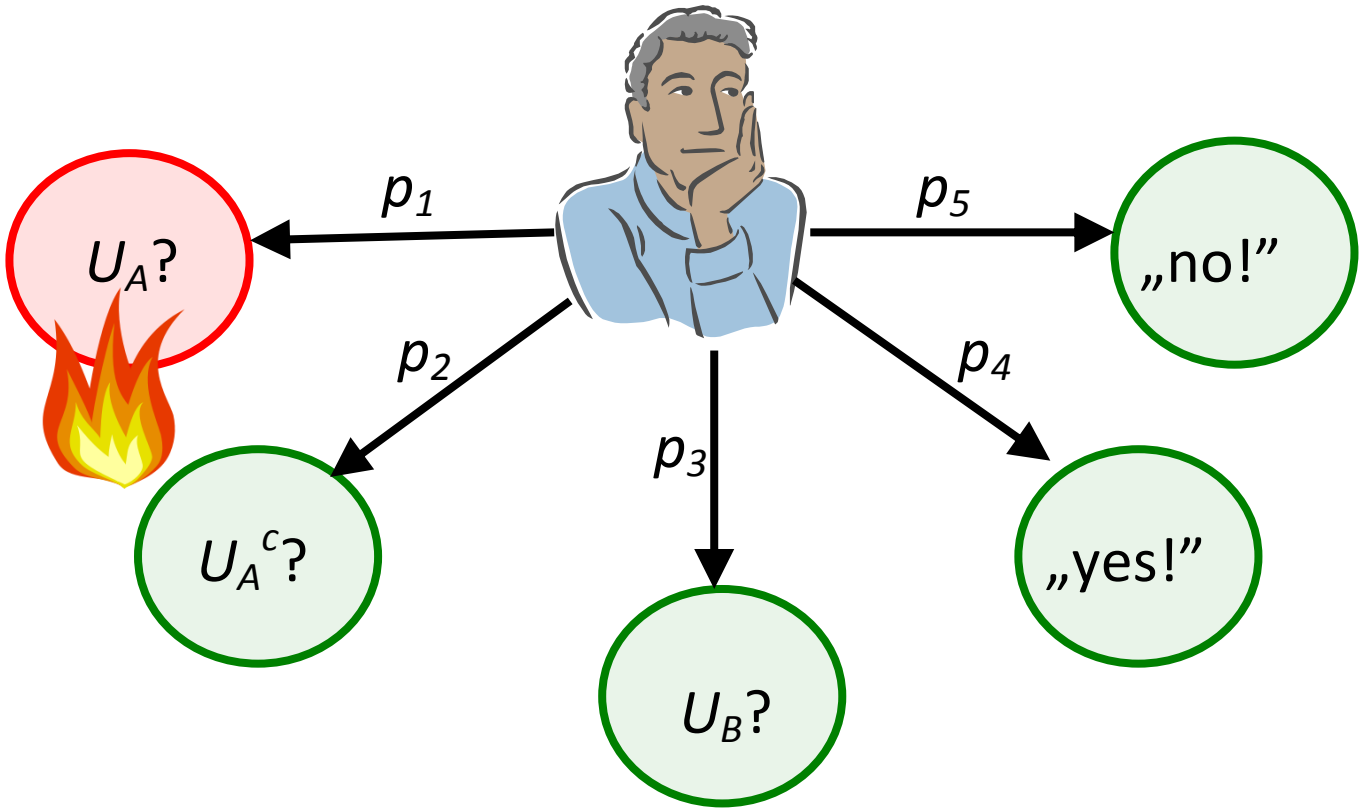


Design parameters: π_B, p_1, p_3 (Horvitz et al. 1967, *Proc. of Sect. on Surv. Res. Meth.*, ASA)

U_B … non-sensitive group of relative size π_B



Design parameters: p_1, p_4, p_5 (Boruch 1971, *The Americ.Soc.*)



Unified theoretical approach with design parameters: π_B, p_1, p_2, … , p_5 (Quatember 2009, *Surv.Meth.*)

[Note that this is not an invitation to use all five questions/instructions in the same questioning design]

Let the variable z indicate a "yes"-answer. Then, for given y_k $P(z_k = 1)$ is given by

$$P(z_k = 1) = p_1 \cdot y_k + p_2 \cdot (1 - y_k) + p_3 \cdot \pi_B + p_4 .$$

Hence, with $u \equiv p_2 + p_3 \cdot \pi_B + p_4$ and $v = p_1 - p_2$,

$$y_k^m = \frac{z_k - u}{v}$$

is an unbiased masked value for y_k. Therefore, applying the HT-principle, Quatember (2009) suggested

$$\pi_{A,RQ} = \frac{1}{N} \cdot \sum\nolimits_s y_k^m \cdot d_k$$

as unbiased generalized estimator of π_A (for the unified approaches regarding categorical or quantitative variables, see Quatember 2015, *Pseudo-Populations - A Basic Concept in Statistical Surveys*, Springer).

The theoretical variance of $\pi_{A,RQ}$ is given by

$$V(\pi_{A,RQ}) = V(\pi_{A,HT}) + C$$

with

$$C = \frac{1}{N^2} \cdot \sum_U \left(\frac{u(1-u)}{v^2} + \frac{1-2u-v}{v} \cdot y_k \right) \cdot d_k$$

This factor C can be considered as the costs for the higher privacy protection in terms of accuracy

3. Objectively Calculated Privacy Protection vs. Accuracy

To describe the dependence of the accuracy of $\pi_{A,RQ}$ on the level of privacy protection, the following measure of privacy protection is considered with regard to a "yes"-answer ($z_k = 1$):

$$PP_1 = \frac{min[P(z_k = 1 | k \in U_A), P(z_k = 1 | k \in U_{A^c})]}{max[P(z_k = 1 | k \in U_A), P(z_k = 1 | k \in U_{A^c})]}$$

(cf. Quatember 2015, *MASA*). For the conditional probabilities,

$$P(z_k = 1 | k \in U_A) = p_1 + p_3 \cdot \pi_B + p_4 = u + v$$

and

$$P(z_k = 1 | k \in U_{A^c}) = p_2 + p_3 \cdot \pi_B + p_4 = u$$

applies, respectively.

Therefore, PP_1 can be expressed by

$$PP_1 = \frac{u}{u+v}.$$

Regarding a "no"-answer ($z_k = 0$), the measure yields

$$PP_0 = \frac{min[P(z_k = 0 | k \in U_A), P(z_k = 0 | k \in U_{A^c})]}{max[P(z_k = 0 | k \in U_A), P(z_k = 0 | k \in U_{A^c})]} = \frac{1-u-v}{1-u}$$

For the ratios PP_1 and PP_0, $0 \leq PP_i \leq 1$ applies ($i = 1,0$) with

- zero when for the certain answer i the privacy of the respondents is not protected at all by the questioning design
- one, when respondents' privacy is totally protected

After straightforward calculations, term C in $V(\pi_{A,RQ}) = V(\pi_{A,HT}) + C$ can be expressed as a function $f(PP_1, PP_0)$ of these measures by

$$C = \frac{1}{N^2} \cdot \frac{1}{(1-PP_1)(1-PP_0)} \cdot \left[PP_0 \cdot \sum_U y_k \cdot d_k \cdot + PP_1 \cdot \sum_U (1-y_k) \cdot d_k \cdot \right]$$

For given sampling scheme and *y*-distribution in *U*, the efficiency of an RQ strategy depends solely on the privacy protection offered by the questioning design.

4. Subjectively Perceived Privacy Protection

Features that increase the subjectively perceived privacy protection of a survey are survey mode, self-administration of questionnaires, …

Additional perceived privacy protection might be achieved with an RQ-Design by more-than-one stage designs

Quatember (2012) proved that a multi-stage design can objectively not perform better than its one-stage version at the same level of privacy protection

The choice of an adequate randomization device might also have an effect on perceived privacy protection

For instance, a randomization requiring no physical device makes use of the Newcomb-Benford distribution (cf. Diekmann 2012, Soc.Meth.& Res.):

“Think of a person of whom you recall the house number: If the first digit is from 1 to 6, then answer truthfully on the sensitive question, if it is 7 or 8, then …”

There is a discrepency between the probabilities as perceived by the respondents and the actual probabilities

For instance, the leading digit being from 1 to 6 has a probability of 0.843

For further information:

Chaudhuri A. and Christofides T.C. (2013), *Indirect questioning in sample surveys*, Springer, Heidelberg.

Chaudhuri A. and Christofides T.C. and Rao C.R. (eds.) (2016), *Handbook of Statistics* (Volume 34): Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, Elsevier, Amsterdam (in print).

Quatember A. (2014), A randomized response design for a polychotomous sensitive population and its application to opinion polls, *Model Assisted Statistics and Applications*, 9, pp. 11-23.

Quatember A. (2016), A Mixture of True and Randomized Responses in the Estimation of the Number of People Having a Certain Attribute, In: Chaudhuri A. at al. (eds.) (2016), *Handbook of Statistics* (Volume 34, Elsevier, Amsterdam, pp. 91-103.

Thank you very much for your attention!