Institute of Applied Statistics

CESS BAMBERG 2018
CONFERENCE OF EUROPEAN STATISTICS STAKEHOLDERS

Bamberg, 10-18-2018

STS03: Protection of Privacy in Public Surveys

The two Different Aspects of Privacy Protection in Indirect Questioning Designs

Andreas Quatember
Johannes Kepler University Linz (Austria)

The theory of indirect questioning (IQ) techniques has been developed for different population characteristics, types of variables, and probability sampling schemes (Chaudhuri and Christofides 2013, Chaudhuri et al. 2016)

Their practical use is well documented (Gonzales-Ocantos et al. 2012, Moshagen et al. 2012, Kirchner et al. 2013, De Hon 2014, Malesky et al. 2015, or Corbacho et al. 2016)

Their effect was also investigated in several studies (cf. Lensvelt-Mulders et al. 2005, 2006, Holbrook and Krosnick 2010, Coutts et al. 2011, Krumpal 2012, Jann et al. 2012, Wolter and Preisendörfer 2013, Rosenfeld et al. 2016, Höglinger and Diekmann 2017)

An implementation of Warner's randomized response (RR) technique (1965), of which respondents understand the instructions for use better than of others (Höglinger et al. 2014, 24), is the crosswise model (Yu et al. 2008):

Q1 (the randomizing question): Think of a person, whose birth date you know. Is the birth date within the interval from … to … (design probability p)? [yes/no]

Q2 (the sensitive question): Are you a member of group A? [yes/no]

The answer the respondent has actually to provide is:

Are your answers on Q1 and Q2 the same? [yes/no]

Understanding of the principle correlates significantly with the development of trust in the strategy’s privacy protection (Höglinger et al. 2014, 25)

One can choose the design probability *p* according to own preferences, experiences, assumptions of the sensitivity level of the variable under study, and recommendations from the literature (cf. Greenberg et al. 1969, Fidler and Kleinknecht 1977, Soeken and Macready 1982, Edgell et al. 1982, 95f, Quatember 2009, Höglinger and Diekmann 2017, Online Appendix)

The design probability *p* determines how strong the privacy of respondents is objectively protected

Objectively offered privacy protection

In the literature:

… technique grants respondents full response privacy

… answer to the sensitive question remains completely private

… embarrassing fact in a completely secret way

… the procedure guarantees anonymity …

… a given answer does not reveal the true answer …

… the design protects the anonymity of respondents’ answers

… the respondent’s anonymity is guaranteed

… these surveys guarantee respondent confidentiality …

… the respondent can reveal critical information without fear

Objectively offered privacy protection / accuracy

In the literature:

… technique grants respondents full response privacy

… answer to the sensitive question remains completely private

… embarrassing fact in a completely secret way

… the procedure guarantees anonymity …

… a given answer does not reveal the true answer …

… the design protects the anonymity of respondents' answers

… the respondent's anonymity is guaranteed

… these surveys guarantee respondent confidentiality …

… the respondent can reveal critical information without fear

Objectively offered privacy protection / accuracy

In the literature:

… technique grants respondents full response privacy

… answer to the sensitive question remains completely private

… embarrassing fact in a completely secret way

… the procedure guarantees anonymity …

… a given answer does not reveal the true answer …

… the design protects the anonymity of respondents' answers

… the respondent's anonymity is guaranteed

… these surveys guarantee respondent confidentiality …

… the respondent can reveal critical information without fear

NOT TRUE

Of course, privacy is not protected completely, but at a certain level!

Q1 (the randomizing question): Think of a person, whose birth date you know. Is the birth date within the interval from 1st of January to 30th of December ($p \approx 0.997$)? [yes/no]

Q2 (the sensitive question): Are you a member of group *A*? [yes/no]

The answer the respondent has actually to provide is:

Are your answers on Q1 and Q2 the same? [yes/no]

Of course, privacy is not protected completely, but at a certain level!

Q1 (the randomizing question): Think of a person, whose birth date you know. Is the birth date within the interval from 1st of January to 19th of October ($p \approx 0.800$)? [yes/no]

Q2 (the sensitive question): Are you a member of group *A*? [yes/no]

The answer the respondent has actually to provide is:

Are your answers on Q1 and Q2 the same? [yes/no]

A measure of privacy protection should be considered

Such a measure with respect to a "yes"-answer (z_k = 1) of respondent *k* may be given by

$$P_{1k} = \frac{\min[\Pr("yes" | k \in A), \Pr("yes" | k \in A^C)]}{\max[\Pr("yes" | k \in A), \Pr("yes" | k \in A^C)]}$$

Regarding a "no"-answer, the measure yields

$$P_{0k} = \frac{\min[\Pr("no" | k \in A), \Pr("no" | k \in A^C)]}{\max[\Pr("no" | k \in A), \Pr("no" | k \in A^C)]}$$

($\mathbf{0} \leq P_{ik} \leq \mathbf{1}$; $i = 0,1$; $k \in U$)

For our strategy, it applies for all $k \in U$ that

$$P_{1k} = P_1 = \frac{1-p}{p} = P_{0k} = P_0$$

Q1 (the randomizing question): Think of a person, whose birth date you know. Is the birth date within the interval from 1st of January to 30th of December ($p \approx 0.997$)? [yes/no]

$$P_1 \approx \frac{0.003}{0.997} \approx 0.003$$

For our strategy it applies for all $k \in U$ that

$$P_{1k} = P_1 = \frac{1-p_1}{p_1} = P_{0k} = P_0$$

Q1 (the randomizing question): Think of a person, whose birth date you know. Is the birth date within the interval from 1st of January to 19th of October ($p \approx 0.800$)? [yes/no]

$$P_1 \approx \frac{0.2}{0.8} = 0.25$$

A level of around 0.25 corresponds to results from experiments with respect to the maximum avoidance of refusals and untruthful answering

Under the assumption of full cooperation, it can be shown that the increase of variance V_+ compared to the direct questioning only depends on the level of privacy protection

For SI sampling and Warner's strategy:

$$V_+ = f(P_1) = \frac{1}{n} \cdot \frac{P_1}{(1-P_1)^2}$$

($P_1 \neq 1$; Quatember 2018)

Subjectively perceived privacy protection

The privacy protection objectively measured may differ from the privacy protection subjectively perceived by the respondents (Chaudhuri and Christofides 2013, 169):

> "*Common sense mandates that the perceived protection of privacy is crucial in deciding to participate in a survey dealing with sensitive issues. In fact it is gaining ground the opinion that the perception of privacy protection should also be considered when the protection of privacy offered by various indirect questioning techniques is examined*"

Crosswise model of Warner's RR design:

Q1 (the randomizing question): Throw three dice. Is their sum within the set {8,9,10,11,12,13,14,15,17}? [yes/no]

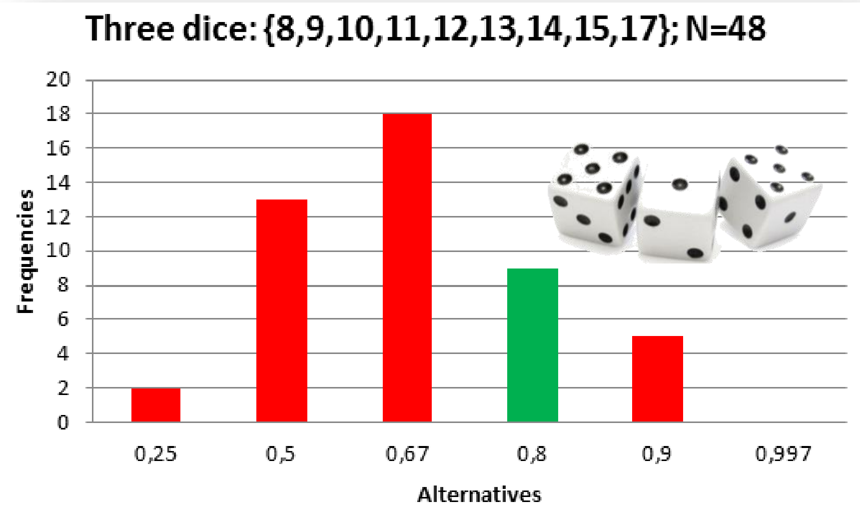
The true design probability p equals 0.81 and the objectively measured privacy protection PP_1 results in

$$P_1 = (1 - p)/p \approx 0.24$$

Crosswise model of Warner's RR design:

Q1 (the randomizing question): Throw three dice. Is their sum within the set {8,9,10,11,12,13,14,15,17}? [yes/no]

Respondents k who would erroneously derive the probability p by observing that 9 of 16 possible outcomes are elements of this set, might think that $pp_k = 9/16 \approx 0.56 < 0.81 = p$



Crosswise model of Warner's RR design:

Q1 (the randomizing question): Throw three dice. Is their sum within the set {8,9,10,11,12,13,14,15,17}? [yes/no]

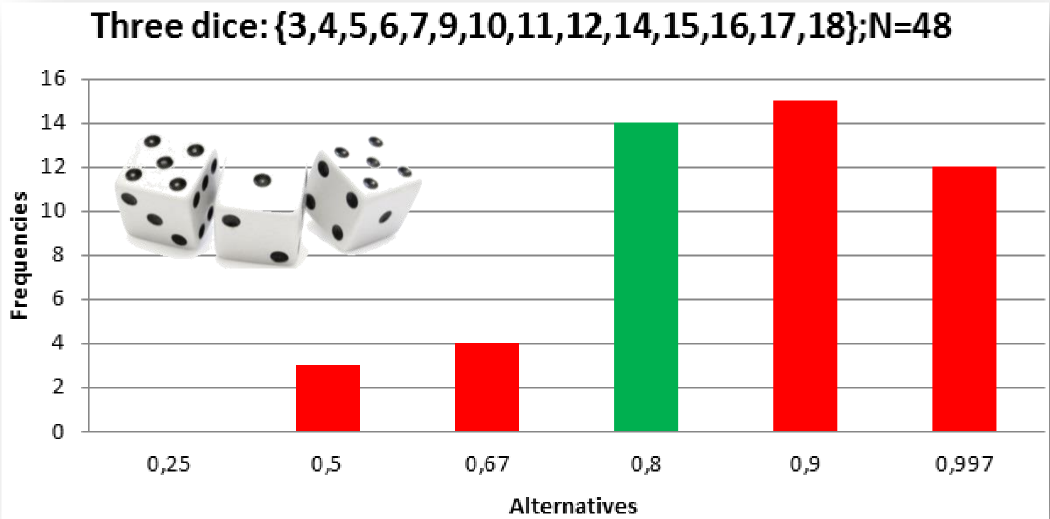
Respondents k who would erroneously derive the probability p by observing that 9 of 16 possible outcomes are elements of this set, might think that $pp_k = 9/16 \approx 0.56 < 0.81 = p$

For such survey units, the subjectively perceived privacy protection PP_{1k} may be calculated by

$$PP_{1k} = (1 - pp_k)/pp_k \approx 0.78 >> 0.24 = P_1$$

In such a case, the respondent should have a higher response propensity

The randomization instruction might also work in the opposite direction, when a respondent perceives a lower privacy protection compared to the real one!

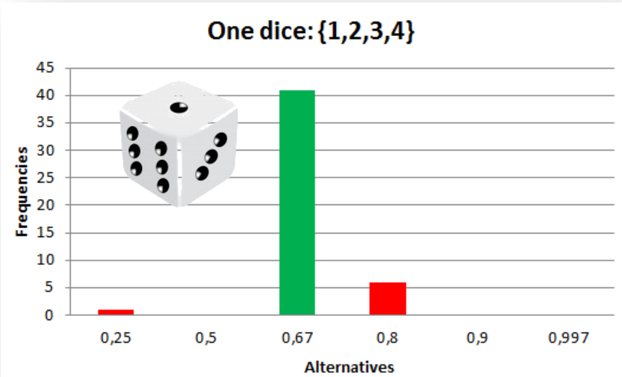


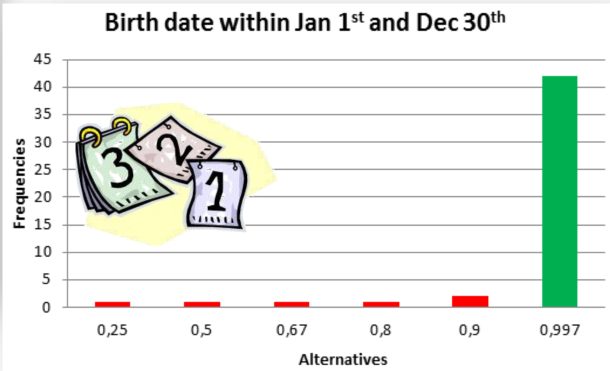
For respondent *k*, these differences can be formalized:

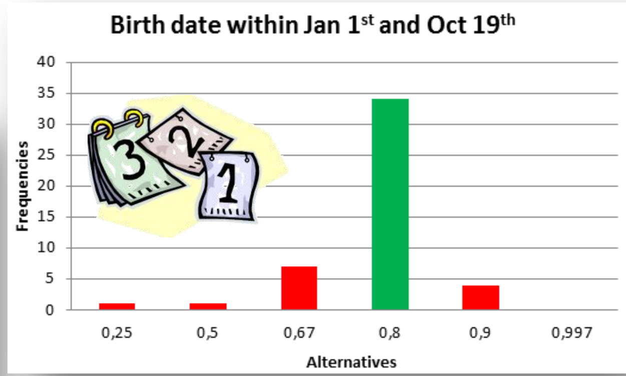
$$\Delta_{1k} = \Delta_{0k} = PP_{1k} - P_1$$

Assuming that the objective measure P_1 was reasonably fixed to allow maximum cooperation and high estimation efficiency, Δ_{1k} must not be negative for any $k \in U$

$\Delta_{1k} = 0$ might always apply when the possible randomization outcomes are uniformly distributed (one dice, birth dates)







Conclusion

The general point is that

- the privacy protection objectively offered by a questioning design directly affects the efficiency of the estimation,
- the privacy protection subjectively perceived by the respondents affects the survey units' willingness to cooperate

That is what indirect questioning designs are all about

Therefore, users have to pay attention to this fact, when choosing adequate randomization instructions to avoid that $\Delta_{1k} = P_{1k} - P_1 < 0$ and/or $\Delta_{0k} = PP_{0k} - P_0 < 0$ applies

Q1 (the randomizing question): Is your birth date within the interval from 1st of January to 19th of October? [yes/no]

Q2 (the sensitive question): Do you feel that this talk was interesting? [yes/no]

The answer you really have to provide is:

Are your answers on Q1 and Q2 the same? [yes/no]

References:

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

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

Corbacho, A., Gingerich, D.W., Oliveros, V., Ruiz-Vega, M.: Corruption as a Self-Fulfilling Prophecy: Evidence from a Survey Experiment in Costa Rica, in: American Journal of Political Science, Vol. 60(4), 1077-1092 (2016)

Coutts, E., Jann, B.: Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). Sociological Methods & Research 40(1), 169-193 (2011)

De Hon, D.: De Nederlandse topsporter en het anti-dopingbeleid 2014 – 2015. Doping Autoriteit. http://www.dopingautoriteit.nl/media/files/2015/Topsportonderzoek_doping_2015-07-21_DEF.pdf (2015). Accessed 21 March 2018

Diekmann, A.: Making use of "Benford's Law" for the randomized response technique. Sociological Methods & Research 41(2), 325-334 (2012)

Edgell, S.E., Himmelfarb, S., Duchan, K.L.: Validity of Forced Responses in a Randomized Response Model, in: Sociological Methods & Research, Vol. 11(1), 89-100 (1982)

Fidler, D.S., Kleinknecht, R.E.: Randomized response versus direct questioning: Two data collection methods for sensitive information. Psychological Bulletin 84(5), 1045-1049 (1977)

Gonzalez-Ocantos E., Kiewiet de Jonge, C., Melendez, C., Osorio, J., Nickerson, D.W.: Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua, in: American Journal of Political Science, Vol. 56(1), 202-217 (2012)

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R., Horvitz, D.G.: The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association 64(326), 520-539 (1969).

Groves R.M., Fowler, F.J., Couper, M.P., Lepkowski, G.M., Singer, E., Tourangeau, R.: Survey Methodology. John Wiley & Sons, Hoboken (2004)

Höglinger, M., Diekmann, A.: Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. Political Analysis 25, 131-137, https://doi.org/10.1017/pan.2016.5 (2017). Accessed 28 September 2018

Höglinger, M., Jann, B., Diekmann, A.: Sensitive Questions in Online Surveys: An Experimental Evaluation of the Randomized Response Technique and the Crosswise Model. University of Berne Social Sciences Working Paper No. 9 (2014)

Holbrook, A.L., Krosnick , J.A.: Measuring Voter Turnout by Using the Randomized Response Technique. The Public Opinion Quarterly 74(2), 328-343 (2010)

Jann, B., Jerke, J., Krumpal, I.: Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism. The Public Opinion Quarterly 76(1), 32-49 (2012)

Kirchner A., Krumpal, I., Trappmann, M., von Hermanni, H.: Messung und Erklärung von Schwarzarbeit in Deutschland - Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit, in: Zeitschrift für Soziologie, Vol. 42(4), 291-314 (2013)

Krumpal, I.: Estimating the Prevalence of Xenophobia and Anti-Semitism in Germany: A Comparison of Randomized Response and Direct Questioning. Social Science Research 41(6): 1387-1403 (2012)

Lensvelt-Mulders, G.J.L.M., Hox, J.J., van der Heijden, P.G.M.: How to Improve the Efficiency of Randomised Response Designs. Quality & Quantity 39, 253-265 (2005)

Lensvelt-Mulders G.J.L.M., van der Heijden, P.G.M., Laudy O., van Gils, G.: A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. Journal of the Royal Statistical Society A 169(2), 305-318 (2006)

Malesky E.J., Gueorguiev, D.D., Jensen, N.M.: Monopoly Money: Foreign Investment and bribery in Vietnam, a Survey Experiment, in: American Journal of Political Science, Vol. 59(2), 419-439 (2015)

Moshagen, M., Musch, J., Erdfelder, E.: A stochastic lie detector. Behavioral Research 44, 222-231 (2012)

Quatember A.: A standardization of randomized response strategies. Survey Methodology 35(*2*), 143-152 (2009)

Quatember, A.: A discussion of the two different aspects of privacy protection in indirect questioning designs. Quality & Quantity, https://doi.org/10.1007/s11135-018-0751-4 (2018). Accessed 28 September 2018

Rosenfeld B., Imai, K., Shapiro, J.N.: An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions, in: American Journal of Political Science, Vol. 60(3), 783-802 (2016)

Singer, E., van Hoewyk, J., Neugebauer, R.J.: Attitudes and behaviour: The impact of privacy and confidentiality concerns on participation in the 2000 Census. The Public opinion Quarterly 67(3), 368-384 (2003)

Soeken, K.L., Macready, G.B.: Respondents' perceived protection when using randomized response. Psychological Bulletin 92(2), 487-489 (1982)

Warner S.L.: Randomized response: A survey technique for eliminating evasive answer bias, in: Journal of the American Statistical Association, Vol. 60, 63–69 (1965)

Wolter, F., Preisendörfer. P.: Asking Sensitive Questions: An Evaluation of the Randomized Response Technique Versus Direct Questioning Using Individual Validation Data. Sociological Methods & Research 42(3), 321-353 (2013)

Yu, J.-W., Tian, G.-L., Tang, M.-L.: Two new models for survey sampling with sensitive characteristic: design and analysis. Metrika 67, 251-263 (2008)

Thank you for your appreciated attention!