Assessing Sensitive Topics in Surveys

Sat Gupta
University of North Carolina - Greensboro
Professor of Statistics
sngupta@uncg.edu

UNCG QMS Series
May 7, 2018

Outline

- Sensitive Topics
- Respondent Privacy
- Data Confidentiality

- Randomized Response Models (RRT Models)
- Full, Partial, or Optional Randomized Response Models

- Sensitive Topics
  - Drugs
  - Violence
  - Sexual Behavior
  - Social Attitude
  - Etc.
- Social Desirability the Main Culprit

- Mailed Surveys
  - Poor Response rate
- Face-to-Face Surveys
  - SDB
  - Response rate very high
- Online Surveys
  - Self Deception

Respondent Privacy – Front-end Problem

- Sensitive topics
- Social desirability bias

- Non response or inaccurate response likely if respondent privacy is not guaranteed

Data Confidentiality – Back-end Problem

- Maintain confidentiality of record level data. Not much worry at aggregate level
- Anonymity violation
- Ethical/Legal Issues
- It is not enough to delete names/subject ID's

Medical Data Compromised

Sweeny L (2002) k-anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge based Systems* 10: 557-570

Randomized Response Models
Data Masking

- Allow respondents to give scrambled response in order to protect their privacy

- Unscrambling possible at aggregate level but not at individual level

Full, Partial, Optional RRT Models

- Full RRT Models – Warner (1965, 1971, *JASA*), Greenberg et al. (1969, 1971, *JASA*)

  All respondents provide scrambled response

- Partial RRT Models – Mangat & Singh (1990, *Biometrika*)

  Only some of the respondents provide scrambled response

- Optional RRT Models – Gupta et al. (2002, *JSPI*)

  The respondent decides whether to give a truthful response or a scrambled response

Model Efficiency

Amount of uncertainty in estimating the important parameters from randomized data

$$Model\ Efficiency = \frac{1}{Var\ (\hat{\theta})}$$

Respondent Privacy

Privacy Level = $E(Z - Y)^2$

Z = Scrambled Response

Y = Unscrambled True Response

Warner (1965) – Indirect Questioning Model for Binary Response

Ask some respondents direct question and some indirect version of the same question randomly

- Did you file a correct tax return last year?
- Did you intentionally file an incorrect last year?

Question 1 is asked with probability *p* and Question 2 with probability *1-p*

$$p \neq 1/2$$

$$p_y = p\pi + (1-p)(1-\pi)$$

p_y = Probability of "Yes" response

p= Proportion of cards with direct question

$$\hat{\pi} = \frac{\frac{n_1}{n} - (1-p)}{2p-1}$$

n_1= Number of "yes" responses in a sample of size *n*

$$Var\left(\hat{\pi}\right) = \frac{\pi\left(1-\pi\right)}{n} + \frac{p\left(1-p\right)}{n\left(2p-1\right)^2}$$

Greenberg et al. (1969) Unrelated Question Binary Model

A proportion *p* of the respondents are asked the real question "Did you intentionally file an incorrect return last year"

Rest are asked an unrelated question like "were you born in the month of January or February"

$$p_y = p\pi + (1-p)\pi_U, \qquad \pi_U = 2/12$$

$$\hat{\pi} = \frac{\frac{n_1}{n} - (1-p)\pi_U}{p}$$

$$Var(\hat{\pi}) = \frac{p_y(1 - p_y)}{np^2}$$

Warner's Additive RRT Model – Quantitative Response

$$Z = Y + S$$

Y = True Response

S = Scrambling Variable (with zero mean)

$$\mu_Z = \mu_Y + \mu_S$$

$$\hat{\mu} = \bar{Z}$$

$\bar{Z}$= Sample Mean of scrambled responses

$$Var(\hat{\mu}) = \frac{\sigma_Y^2}{n} + \frac{\sigma_S^2}{n}$$

Quantitative Data Scrambling

- **Helps with both – respondent privacy and data confidentiality**

- Too much scrambling or too little scrambling

- Think of two data scrambling models for variable *Y*. *S* is a scrambling variable and θ is a constant

$$Z = Y + S$$
$$Z = Y + \theta S$$

- Confidentiality is higher when θ is larger
- Data quality is better when θ is smaller
- Same dilemma as in reliability vs. precision in confidence intervals

Greenberg Unrelated Question Quantitative Model (1971)

$$Z = \begin{cases} Y \; with \; probability \; p \\ U \; with \; probability \; (1-p) \end{cases}$$

$$E(Z) = p\mu_Y + (1-p)\mu_U$$

$$\hat{\mu} = \frac{\bar{Z} - (1-p)\mu_U}{p}$$

$$Var(\hat{\mu}) = \frac{Var(\bar{Z})}{np^2}$$

Recent Applications of RRT

Ostapczuk et al. (2009): ***European Journal of Social Psychology***

A randomized-response investigation of the education effect in attitudes towards foreigners

Spears- Gill et al. (2013): ***Springer Proceedings in Mathematics and Statistics*****,**

Estimates of risky sexual behaviors among college students

Chhabra et al. (2016): ***North Carolina Journal of Mathematics and Statistics***

Prevalence of sexual abuse of female college students by acquaintances

Thank You!

Questions/Comments