Plagiarism in student papers and cheating on exams

Results from a survey using special techniques for sensitive questions

Ben Jann

University of Bern, jann@soz.unibe.ch

17th UK Stata Users' Group meeting

London, September 15/16, 2011

Outline

- Sensitive questions in survey research
- Some indirect approaches to elicit truthful answers
  - The Randomized Response Technique (RRT)
  - The Crosswise Model: A new alternative to RRT
  - Stata implementation of estimators
- Empirical application: Plagiarism and cheating on exams
- Conclusions

Eliciting truthful answers – not an easy task

- Direct questioning (DQ) does often not work ...

Karl-Theodor zu Guttenberg, February 2011
(German Minister of Defense at that time)

Eliciting truthful answers – not an easy task

- A couple days later, the University of Bayreuth revoked zu Guttenberg's doctoral degree because of plagiarism.
- Some examples for proportion of "liars" (respondents with a false negative response) in surveys that use direct questioning (estimates from validation studies):
  - Penal conviction: 42.5% (F2F, Wolter 2010)
  - Welfare and unemployment benefit fraud: 75% (F2F, van der Heijden et al. 2000)
  - Driving under influence: 54% (P&P, Locander et al. 1976)
  - Bankruptcy: 32% (P&P, Locander et al. 1976)

Some traditional measurement approaches

Asking the Embarrassing Question

By Allen H. Barton

University of Chicago

The pollster's greatest ingenuity has been devoted to finding ways to ask embarrassing questions in non-embarrassing ways. We give here examples of a number of these techniques, as applied to the question, "Did you kill your wife?"

1. The Casual Approach:

   "Do you happen to have murdered your wife?"

(POQ 22/1958: 67-68)

Some traditional measurement approaches

2. The Numbered Card:

   Would you please read off the number on this card which corresponds to what became of your wife?" (HAND CARD TO RESPONDENT)

   1. Natural death
   2. I killed her
   3. Other (What?)

   (GET CARD BACK FROM RESPONDENT BEFORE PROCEEDING!)

3. The Everybody Approach:

   "As you know, many people have been killing their wives these days. Do you happened to have killed yours?"

. . .

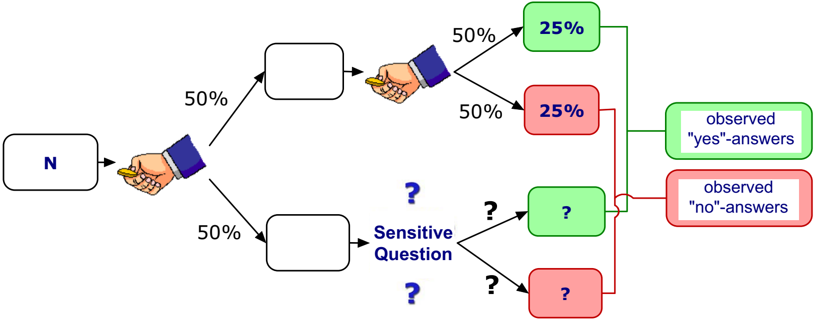
8. Putting the question at the end of the interview.

The Randomized Response Technique (RRT)

(Warner 1965; Fox and Tracy 1986)

- Main principle: privacy protection through randomization (i.e. add random noise to the answers)
- A *randomizing device*, the outcome of which is only known to the respondent, decides whether ...
  - the sensitive question has to be answered
  - or an automatic "yes" or "no" has to be given or a surrogate question has to be answered
- Since only the respondent knows the outcome of the randomization device, a "yes" cannot be interpreted as an admission of guilt.
- However, if the properties of the randomizing device are known, a prevalence estimate for the sensitive question can be derived.

Example (forced response RRT)



$$\Pr(\text{observed yes}) = \Pr(\text{sensitive question}) \cdot \pi + \Pr(\text{surrogate yes})$$

$$\Rightarrow \quad \pi = \frac{\Pr(\text{observed yes}) - \Pr(\text{surrogate yes})}{\Pr(\text{sensitive question})}$$

The Crosswise Model (CM): A new alternative to RRT

(Yu, Tian, and Tang 2008)

- Very simply idea: Ask a sensitive question and a nonsensitive question and let the respondent indicate whether ...
  - **A** the answers to the questions are the same (both "yes" or both "no")
  - **B** the answers are different (one "yes", the other "no")

		nonsensitive question: no	nonsensitive question: yes
sensitive question	no	**A**	**B**
	yes	**B**	**A**

  - Note: Questions must be uncorrelated and probability of "yes" must be unequal 0.5 for the nonsensitive question.

The Crosswise Model (CM): A new alternative to RRT

(Yu, Tian, and Tang 2008)

- Prevalence estimate:

$$\Pr(A) = (1 - \pi) \cdot (1 - \Pr(\text{nonsensitive yes})) + \pi \cdot \Pr(\text{nonsensitive yes})$$

$$\Rightarrow \quad \pi = \frac{\Pr(A) + \Pr(\text{nonsensitive yes}) - 1}{2 \cdot \Pr(\text{nonsensitive yes}) - 1}$$

  - Note: Crosswise Model is formally identical to Warner's original RRT model.

The Crosswise Model: Let's practice

- Two questions:
  1. Is your mother's birthday in January or February?
  2. Did you ever falsify your data or results?

     (e.g. edit data points or delete observations so that hypothesis is confirmed, falsify entire dataset, invent or manipulate reported results)

- Compare your answers: Are they the same or different?
  - Write "**A**" if they are the same (both Yes or both No)
  - Write "**B**" if they are different (one Yes, the other No)

The Crosswise Model: Let's practice

- Two questions:
  1. Is your mother's birthday in January or February?
  2. Did you ever falsify your data or results?
     (e.g. edit data points or delete observations so that hypothesis is confirmed, falsify entire dataset, invent or manipulate reported results)
- Compare your answers: Are they the same or different?
  - Write "**A**" if they are the same (both Yes or both No)
  - Write "**B**" if they are different (one Yes, the other No)

Results of survey:

- 24 A and 10 B (N = 34); 1 C
- Design parameter: Pr(nonsensitive yes) $= 2/12$
- Prevalence estimate:
  $\hat{\pi} = \frac{\widehat{\Pr(A)} + \Pr(\text{nonsensitive yes}) - 1}{2 \cdot \Pr(\text{nonsensitive yes}) - 1} = \frac{24/34 + 2/12 - 1}{2 \cdot 2/12 - 1} = 19.1\%$
- Standard error:
  $SE(\hat{\pi}) = \sqrt{\frac{\widehat{\Pr(A)} \cdot (1 - \widehat{\Pr(A)})}{(N-1) \cdot (2 \cdot \Pr(\text{nonsensitive yes}) - 1)^2}} = \sqrt{\frac{24/34 \cdot (1 - 24/34)}{(34-1) \cdot (2 \cdot 2/12 - 1)^2}} = 11.9\%$
- 95% confidence interval: $[-5.1\%, 43.3\%]$

Generalized estimator for RRT and CM

- Let

  Y_i response ($Y_i = 1$ if "yes" in RRT or "A" in CM, else $Y_i = 0$)

  λ_i probability of $Y_i = 1$

  π_i (unknown) prevalence of sensitive item

  p_i^{w} probability of being directed to the negated question in Warner's RRT (or prevalence of nonsensitive item in CM)

  p_i^{yes} overall probability of surrogate "yes"

  p_i^{no} overall probability of surrogate "no"

- Then

$$\lambda_i = (1 - p_i^{\mathrm{yes}} - p_i^{\mathrm{no}}) p_i^{\mathrm{w}} \pi_i + (1 - p_i^{\mathrm{yes}} - p_i^{\mathrm{no}})(1 - p_i^{\mathrm{w}})(1 - \pi_i) + p_i^{\mathrm{yes}}$$

  and hence

$$\pi_i = \frac{\lambda_i - (1 - p_i^{\mathrm{yes}} - p_i^{\mathrm{no}})(1 - p_i^{\mathrm{w}}) - p_i^{\mathrm{yes}}}{(2p_i^{\mathrm{w}} - 1)(1 - p_i^{\mathrm{yes}} - p_i^{\mathrm{no}})}$$

Two Stata commands (available from the SSC Archive)

`rrreg` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *regress_options*
`pwarner(`#|*varname*`)` `pyes(`#|*varname*`)` `pno(`#|*varname*`)`]

- Assumes $\pi_i = X_i'\beta$ and estimates β using least squares with transformed response

$$\tilde{Y}_i = \frac{Y_i - (1 - p_i^{\text{yes}} - p_i^{\text{no}})(1 - p_i^{\text{w}}) - p_i^{\text{yes}}}{(2p_i^{\text{w}} - 1)(1 - p_i^{\text{yes}} - p_i^{\text{no}})}$$

Two Stata commands (available from the SSC Archive)

`rrlogit` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *logit_options*
`pwarner(`# | *varname*`)` `pyes(`# | *varname*`)` `pno(`# | *varname*`)`]

- Assumes $\pi_i = e^{X_i'\beta}/(1 + e^{X_i'\beta})$ and estimates β using maximum likelihood with

$$\ln L = \sum_{i=1}^{n} \left\{ Y_i \ln(R_i) + (1 - Y_i) \ln(S_i) - \ln\left(1 + e^{X_i'\beta}\right) \right\}$$

where

$$R_i = c_i + q_i e^{X_i'\beta} \qquad c_i = (1 - p_i^{\text{yes}} - p_i^{\text{no}})(1 - p_i^{\text{w}}) + p_i^{\text{yes}}$$
$$S_i = (1 - c_i) + (1 - q_i) e^{X_i'\beta} \qquad q_i = (1 - p_i^{\text{yes}} - p_i^{\text{no}}) p_i^{\text{w}} + p_i^{\text{yes}}$$

Empirical application

- Web-Survey among student of University of Bern and ETH Zurich in Spring 2011
- Response rate 33%
- Comparing direct questioning to three variants of RRT and two variants of the Crosswise Model
  - *example*
- Sensitive questions on
  - copying from other students in exam (`copy`)
  - using crib notes in exam (`notes`)
  - taking drugs to enhance performance on exam (`drugs`)
  - partial plagiarism (`partial`)
  - sever plagiarism/ghostwriting (`severe`)
- Team: Marc Höglinger, Ben Jann, Andreas Diekmann

Data analysis

```
prog mylincom
    *! version 1.0.1  08jul2009  Ben Jann
    syntax anything(everything equalok), Name(name) [ Level(passthru) ]
    lincom `anything´, `level´
    tempname b b1 V
    mat `b1´ = r(estimate)
    mat coln `b1´ = lincom:`name´
    mat `b´ = e(b)
    local eqs: coleq `b´, quoted
    local eqs: subinstr local eqs `""_""´ `""main""´, all word
    mat coleq `b´ = `eqs´
    mat `b´ = `b´, `b1´
    mat `V´ = (e(V), J(rowsof(e(V)), 1, 0)) \ ///
        (J(1, colsof(e(V)), 0), r(se)^2)
    erepost b=`b´ V=`V´, rename
end
global sqvar copy notes drugs partial severe
forv i = 1/5 {
    local depvar: word `i´ of $sqvar
    di as res _n "==> depvar: `depvar´"
    rrreg `depvar´ DQ RRT CM, nocons hc2 pyes(pyesQ`i´) pno(pnoQ`i´) ///
        pwarner(pwarnQ`i´)
    mylincom RRT-DQ, name(RRT_DQ)
    mylincom CM-DQ, name(CM_DQ)
    eststo `depvar´
}
esttab, starkeep(lincom:) nonumb mti se b(1) compress transform(100*@ 100) ///
     eqlab("Level" "Difference") coef(RRT_DQ "RRT - DQ" CM_DQ "CM - DQ")
```

Results: Prevalence estimates

	copy	notes	drugs	partial	severe
Level					
DQ	17.5	8.8	3.4	2.5	1.5
	(1.2)	(0.9)	(0.6)	(0.6)	(0.5)
RRT	19.6	12.7	0.6	4.2	-0.6
	(1.2)	(1.1)	(1.0)	(1.2)	(1.1)
CM	27.2	15.0	9.9	8.2	3.0
	(2.0)	(1.9)	(1.9)	(2.1)	(2.0)
Difference					
RRT - DQ	2.1	3.9**	-2.8*	1.7	-2.1
	(1.7)	(1.5)	(1.1)	(1.3)	(1.2)
CM - DQ	9.7***	6.2**	6.5***	5.7**	1.5
	(2.3)	(2.1)	(2.0)	(2.2)	(2.1)
N	5726	5727	5711	4226	4224

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Results: Determinants of sensitive behavior

	copy	notes	drugs	partial	severe
Perceived prevalence	0.053***	0.043***	0.042***	0.051***	0.072***
	(0.003)	(0.003)	(0.006)	(0.006)	(0.014)
Perceived risk	-0.018***	-0.031***		0.002	-0.014
	(0.005)	(0.008)		(0.005)	(0.016)
Risk attitude	0.074*	0.088*	0.142*	0.079	0.104
	(0.032)	(0.036)	(0.064)	(0.070)	(0.137)
RRT	0.172	0.508**	-0.175	0.812*	-0.432
	(0.148)	(0.172)	(0.334)	(0.350)	(0.679)
CM	0.876***	0.774***	0.860**	1.515***	-0.711
	(0.178)	(0.212)	(0.322)	(0.377)	(1.895)
Constant	-3.475***	-3.718***	-4.997***	-5.226***	-5.390***
	(0.257)	(0.297)	(0.476)	(0.586)	(0.983)
N	4956	4973	5270	3696	3162

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Summary of results

- Crosswise Model clearly outperforms direct questioning (if we are ready to accept the "more-is-better assumption").
  - An exception is the last item (severe plagiarism), where prevalence is very low for all techniques.
- RRT, on the other hand, does not yield higher estimates than direct questioning
- Reasons for the failure of RRT
  - difficulties understanding RRT, no trust in RRT
  - "self-protective no" bias
    - respondents who are not guilty are reluctant to give a "yes" answer and, hence, do not comply with the instructions
    - in RRT there is a "dominant strategy": say "no", no matter what

Methodological conclusions

- The Randomized Response Technique does not seem to be a good method for self-administered surveys. Although we put a lot of effort into pretesting and finding good implementations, no convincing evidence could be found that RRT yields more valid estimates than direct questioning.
- The Crosswise Model is a promising alternative, since it does not suffer from some of the deficiencies of the RRT ("self-protective no" bias, complexity).
- Improvement of RRT estimates is possible by correcting for cheating respondents who do not comply with the instructions. Such estimates, however, have low efficiency.

Substantive conclusions

- A substantial proportion of students have cheated on an exam (copying: 20 to 25 percent, crib notes: around 15 percent)
- Using drugs to enhance performance on exams is not uncommon (about 10 percent)
- Rates for partial plagiarism (using a passage from someone else's work without providing proper citation) are about 10 percent. The prevalence of severe plagiarism (hand in someone else's work) is about 2 percent.
- These numbers may not seem too high, but keep in mind:
  - There is lots of nonresponse, and probably mostly the "nice guys" participate.
  - Even with these low numbers we would expect about 200 papers a year containing plagiarism and about 40 papers, that are entirely falsified, at a small University with about 10000 Students.

Thank you for your attention!

References I

- Coutts, E., B. Jann, I. Krumpal, F. Näher (forthcoming). Plagiarism in Student Papers: Prevalence Estimation Using Special Techniques for Sensitive Questions. *Journal of Economics and Statistics*.
- Fox, J. A., P. E. Tracy. 1986. *Randomized response: A method for sensitive surveys.* London: Sage
- Jann, B., J. Jerke und I. Krumpal (forthcoming). Asking Sensitive Questions Using the Crosswise Model: Some Experimental Results. *Public Opinion Quarterly*.
- Locander, W., S. Sudman, N. Bradburn. 1976. An Investigation of Interview Method, Threat and Response Distortion. *Journal of the American Statistical Association* 71: 269-275.
- van der Heijden, P. G. M., G. van Gils, J. Bouts, J. J. Hox. 2000. A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning. Eliciting Sensitive Information in the Context of Welfare and Unemployment Benefit. *Sociological Methods and Research* 28: 505–537.
- Warner, S. L. 1965. Randomized-response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60:63–69.

References II

- Wolter, F. 2010. Sensitive Questions in Surveys and the Randomized Response Technique: Theoretical Explanations of Respondent Behavior and Results from a Validation Study. Paper presented at the workshop on "Rational Choice Sociology: Theory and Empirical Applications" at Venice International University, November 30, 2010.
- Yu, J.-W., G.-L. Tian, M.-L. Tang. 2008. Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika* 67:251-263.