INDIRECT QUESTIONING ABOUT SENSITIVE FEATURES – ON MODIFIED POISSON AND NEGATIVE BINOMIAL ITEM COUNT TECHNIQUES

Barbara Kowalczyk

SGH Warsaw School of Economics, Poland

Robert Wieczorkowski

Statistics Poland

Agenda

- Introduction to indirect methods of questioning about sensitive attributes
  - item count techniques
- Poisson and negative binomial item count techniques, Tian et al. (2017)
- A new proposed model – improved Poisson and negative binomial item count techniques
  - we have increased efficiency of the estimation of the unknown sensitive proportion with the same degree of the privacy protection DPP, and only a slight modification in the questionnaire design

Sensitive questions

- Private, stigmatizing, socially unaccepted attributes, illegal behaviors
- e.g. abortion, corruption, tax frauds, illegal work, black market, using drugs, dangerous or atypical sexual behaviors, politically incorrect views (only in some countries)
- e.g. I cannot ask directly „Have you ever bribed an official?" because the most likely answer will be one of the two:
  - Bribery in our country? It's a fake news!
  - Are you out of your mind? I am the most honest person in the world

Indirect methods of questioning

- *Randomized response techniques*
  - Mirror question design
  - Nonrelated question design
  - Forced question design
- *Non-randomized response techniques*
  - Crosswise model
  - Triangular model
- *Item count techniques*

Randomized response technique RRT

- Unrelated question design Greenberg at al. (1969)

Please take out any banknote that you have in your wallet. But do not show it to me.

If the last number of the serial number on your banknote is 0,1,2,3,4 or 5 please answer the question:

Have you ever bribed an official?

If the last number of the serial number on your banknote is 6, 7, 8 or 9 please answer the question:

Were you born in an even month of the year?

Randomized response technique RRT

- Mirror question design Warner (1965)

With probability p, $0 < p < 1$ respondents answer the question 'Is it true that you have bribed an official?' and with probability $1 - p$ they answer the question 'Is it true that you have never bribed an official?', $p \neq 0.5$

- Forced question design Fox and Tracy (1986)

With probability p, respondents are asked (forced) to write NO, with probability q, respondents are asked (forced) to write YES, with probability $1 - p - q$ they are asked to answer the sensitive question.

Deck 1	
„I belong to A” with probability W	„Go to deck 3” with probability (1-W)

↓

Deck 3		
„I belong to A” with probability P	„Yes” with probability 0.5(1-P)	„No” with probability 0.5(1-P)

Deck 2	
„I belong to A” with probability Q	„Go to deck 4” with probability (1-Q)

↓

Deck 4		
„I belong to A” with probability T	„Yes” with probability 0.5(1-T)	„No” with probability 0.5(1-T)

Source: Abdelfatah S. and Mazloum R. Efficient estimation in a two-stage randomized response model, Mathematical Population Studies, 22: 234-251, 2015

Randomized response techniques

- They need a randomized device
- They require a face to face survey
- They cannot be used in telephone surveys
- They cannot be used in internet surveys
- Respondents view them as tricky (they do not understand the mathematics behind it)
- RRTs have met some serious criticism among applied researchers

Non-randomized response techniques

- Crosswise model, Tan et. al. 2009

Have you ever bribed an official?

Were you born in an even month of the year?

Choose one of the following statements:

- The answers to the two questions are the same (both are answered YES or both are answered NO)
- Answers to the two questions are different (one question is answered YES and one is answered NO)

Non-randomized response techniques

- Triangular model, Yu et. al. 2008

Have you ever bribed an official?

Were you born in an even month of the year?

Choose one of the following:

- Neither is true
- At least one is true

Triangular model

X – neutral non-related variable, $X \sim Bernoulli(r)$

Z – sensitive variable, $Z \sim Bernoulli(\pi)$

X, Z – independent

Y – observed binary response (1 if at least one is true)

$$P(Y = 1) = 1 - (1 - r)(1 - P(Z = 1) = r + (1 - r)P(Z = 1)$$

$$P(Z = 1) = (P(Y = 1) - r) \cdot \frac{1}{1 - r}$$

$$\hat{\pi}_{ML} = \frac{\bar{Y} - r}{1 - r}$$

$$Var(\hat{\pi}_{ML}) = \frac{1}{n}(1 - \pi)\left(\pi + \frac{r}{1 - r}\right)$$

Degree of privacy protection

- Degree of privacy protection $DPP = P(Z = 1|Y)$
- The smaller the probability the more respondent is being protected (for negative – badly seen – sensitive attributes)
- Triangular model:

$$DPP(\pi, r|Y = 0) = P(Z = 1|Y = 0) = 0$$

$$DPP(\pi, r|Y = 1) = P(Z = 1|Y = 1) = \frac{\pi}{\pi + (1 - \pi)r}$$

$$DPP(\pi, r = 0|Y = 1) = 1$$

$$DPP(\pi, r = 1|Y = 1) = \pi$$

Item Count Technique, Miller (1984)

- Survey respondents are randomly assigned to either the control or treatment group, $n = n_C + n_T$
- Respondents in the control group are given a list of *J* neutral questions (or statements) with binary outcomes
- Respondents in the treatment group are given a list of *J+1 questions, J* the same neutral questions as in the control group plus 1 sensitive
- Respondents are asked to report only the total of their Yes (or True) answers. In the control group it can be a number from 0 to *J*, in the treatment group it can be a number from 0 to *J*+1.

Real questionnaire (survey on racism in USA)

(Conrol group) *"Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many)*

- *the federal government increasing the tax on gasoline;*
- *professional athletes getting million-dollar-plus salaries;*
- *large corporations polluting the environment.*

How many, if any, of these things upset you?"

Source: Imai K., (2011), *Multivariate regression analysis for the item count technique*, Journal of American Statistical Assosiasion, 206, p. 407-416.

Real questionnaire (survey on racism in USA)

(Treatment group) *"Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many)*

- *the federal government increasing the tax on gasoline;*
- *professional athletes getting million-dollar-plus salaries;*
- *large corporations polluting the environment;*
- *a black family moving next door to you.*

How many, if any, of these things upset you?"

Source: Imai K., (2011), *Multivariate regression analysis for the item count technique*, Journal of American Statistical Assosiasion, 206, p. 407-416.

Item Count Technique

- Advantages:
  - It is very simple and easy for implementation
  - It does not need any randomize device
  - It can be used in telephone surveys
  - It can be used in internet surveys
  - Respondents know how their privacy is being protected

Item Count Technique
Problem 1

- The method was proposed by Miller (1984) and since that time it has been widely used in practice. For many years applied researchers used only simple MME $\hat{p} = \bar{Y}_T - \bar{Y}_C$, which can result in values < 0

- Proper mathematical background with ML estimation using EM algorithm was given only in 2011
  - Imai K. (2011), Multivariate regression analysis for the item count technique, Journal of American Statistical Assosiasion, Vol. 206, pp. 407-416.

Item Count Technique
Problem 2

- The ceiling effect
  - If all neutral statements (questions) are applicable to the respondent and he or she possesses the sensitive attribute then their privacy is no longer being protected (most dangerous for negative - badly seen – sensitive attributes)

- The floor effect
  - if none of the neutral statements (questions) is applicable to the respondent and he or she does not possess the sensitive attribute then their privacy is no more being protected (most dangerous for positive - well seen – sensitive attributes)

Item Count Technique
Problem 2

- The ceiling effect - solutions
  - *Not respondent-friendly solution*

    Chaudhuri A and Christofides TC. Item Count Technique in estimating the proportion of people with a sensitive feature. J Stat Plann Inference 2007, 137, 589-593
  - *Respondents-friendly solution*

    Tian G-L, Tang M-L, Wu Q, Liu Y. Poisson and negative binomial item count techniques for surveys with sensitive question. Stat Methods Med Res 2017, 26, 931-947.
- The floor effect - solution

  Kowalczyk, Niemiro, Wieczorkowski, Item count technique with a continuous control variable for analyzing sensitive questions in surveys (submitted)

Item Count Technique
Problem 2

- Not respondent-friendly solution

Chaudhuri A and Christofides TC. Item Count Technique in estimating the proportion of people with a sensitive feature. J Stat Plann Inference 2007, 137, 589-593

  - *One of the J+1 statements is of the form*
    - Control group: either *N* or *S*
    - Treatment group: either not *N* or not *S*

- ICT has lost his main advantage – simplicity

Poisson and negative binomial item count techniques, Tian et al. (2017)

- *Control group:*

'(1) How many times did you travel abroad last year? Please report your answer (denoted by X) to this question.'

- *Treatment group:*

'(1) How many times did you travel abroad last year?
(2) Have you ever shoplifted? (1 for 'yes'; and 0 for 'no')
Please report ONLY the sum (denoted by Y=X+Z) of the answers to the two questions.'

Tian G-L, Tang M-L, Wu Q, Liu Y. Poisson and negative binomial item count techniques for surveys with sensitive question. Stat Methods Med Res 2017, 26, 931-947.

Degree of Privacy Protection

- For Poisson and NB ICTs degree of privacy protection $DPP = P(Z = 1|Y) < 1$

- Tian at al. (2017) propose to choose a question with $\lambda = 2$ for Poisson ICT. Example: if $y < 8$ then for Poisson ICT

$$DPP(\pi = 0.2; \lambda = 2|y) = \frac{P(Z=1,Y=y)}{P(Y=y)} = \frac{y}{y+8} < 0.5$$

Poisson and negative binomial ICTs

- The ceiling effect is eliminated
- The simplicity of the method is remained (the questionnaire is even slightly simplified)
- The method is not very efficient
- The aim of the new model:
  - *To increase efficiency of the estimation but with the same DPP and almost the same questionnaire*

Newly proposed model

Group I

- How many times did you use a taxi last month ($X^{(1)}$)?

Your answer is …

- How many times were you at the cinema last month ?
- Have you ever bribed an official ? Assign number 1 if ‘yes’ and number 0 if ‘not’.

Please report the sum ($X^{(2)}$ +Z) of the two numbers ONLY. The sum is …

Newly proposed model

Group II

- How many times were you at the cinema last month ($X^{(2)}$)?

Your answer is…

- How many times did you use a taxi last month?
- Have you ever bribed an official? Assign number 1 if ‘yes’ and number 0 if ‘not’ .

Please report the sum ($X^{(1)}$ +Z) of the two numbers ONLY. The sum is …

Model

$X^{(1)}$ – answer to the first non-sensitive question, $X^{(1)} \in \{0,1,2, \dots\}$

$X^{(2)}$ – answer to the second non sensitive question, $X^{(2)} \in \{0,1,2, \dots\}$

Z – answer to the sensitive question, $Z \in \{0,1\}$,

$\pi = P(Z = 1)$ – unknown sensitive proportion under study

We assume that $X^{(1)}$, $X^{(2)}$, Z are independent

Model

Vector of observable variables:

$$\left(X_1^{(1)},\ldots,X_{n_1}^{(1)},Y_1^{(1)},\ldots,Y_{n_1}^{(1)},X_{n_1+1}^{(2)},\ldots,X_{n_1+n_2}^{(2)},Y_{n_1+1}^{(2)},\ldots,Y_{n_1+n_2}^{(2)}\right)$$

where

$$Y_i^{(1)} = X_i^{(2)} + Z_i \text{ for i} = 1,2,\ldots,n_1$$

$$Y_j^{(2)} = X_j^{(1)} + Z_j \text{ for i} = n_1+1, n_1+2,\ldots,n_1+n_2$$

Z is not directly observable (it is a hidden, latent variable)

Empirical BLUE estimator

- Empirical BLUE estimator:

$$\hat{\pi}^{EBLUE} = \hat{w}^{emp}\left(\bar{Y}^{(2)} - \bar{X}^{(1)}\right) + (1 - \hat{w}^{emp})\left(\bar{Y}^{(1)} - \bar{X}^{(2)}\right)$$

$$w^{emp} = \frac{\frac{1}{n_1}S^2\left(Y^{(1)}\right) + \frac{1}{n_2}S^2\left(X^{(2)}\right)}{\frac{1}{n_2}S^2(Y^{(2)}) + \frac{1}{n_1}S^2(X^{(1)}) + \frac{1}{n_1}S^2(Y^{(1)}) + \frac{1}{n_2}S^2(X^{(2)})}$$

where $S^2\left(X^{(1)}\right)$, $S^2\left(X^{(2)}\right)$, $S^2\left(Y^{(1)}\right)$, $S^2\left(Y^{(2)}\right)$ are sample variances of observed variables $X^{(1)}$, $X^{(2)}$, $Y^{(1)}$, $Y^{(2)}$ respectively.

ML estimation via EM algorithm

- $Z \sim Bernoulli(\pi)$, π – unknown sensitive proportion under study
- $X^{(1)} \sim p_{\theta_1}(x)$ pmf depending on parameter θ_1
- $X^{(2)} \sim p_{\theta_2}(x)$ pmf depending on parameter θ_2

$$L_{com}\left(\pi, \theta_1, \theta_2; \boldsymbol{x^{(1)}}, \boldsymbol{x^{(1)}}, \boldsymbol{x^{(2)}}, \boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}, \boldsymbol{z}\right) =$$

$$\prod_{i=1}^{n_1} p_{\theta_1}(x_i)\left[p_{\theta_2}(y_i - 1)\right]^{z_i}\left[p_{\theta_2}(y_i)\right]^{1-z_i}\pi^{z_i}[1-\pi]^{1-z_i} \cdot$$

$$\prod_{j=n_1+1}^{n_1+n_2} p_{\theta_j}(x_j)\left[p_{\theta_1}(y_j - 1)\right]^{z_j}\left[p_{\theta_1}(y_j)\right]^{1-z_i}\pi^{z_j}[1-\pi]^{1-z_j}$$

ML estimation via EM algorithm

Complete log-lik is:

$$lnL_{com}\left(\pi, \theta_1, \theta_2; \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \boldsymbol{z}\right) =$$
$$\sum_{i=1}^{n_1} lnp_{\theta_1}(x_i) + \sum_{i=n_1+1}^{n_1+n_2} lnp_{\theta_2}(x_j) +$$
$$+ \sum_{i=1}^{n_1} z_i lnp_{\theta_2}(y_i - 1) + \sum_{j=n_1+1}^{n_1+n_2} z_j lnp_{\theta_1}(y_j - 1) +$$
$$+ \sum_{i=1}^{n_1} (1 - z_i) lnp_{\theta_2}(y_i) + \sum_{j=n_1+1}^{n_1+n_2} (1 - z_j) lnp_{\theta_1}(y_j) +$$
$$+ \sum_{j=1}^{n_1+n_2} z_j \ln\pi + \sum_{j=1}^{n_1+n_2} (1 - z_j) \ln(1 - \pi)$$

ML estimation via EM algorithm

Conditional expectation computed in E-step of EM algorithm:

$$E_{\pi_0,\theta_{10},\theta_{2o}}\left[lnL_{com;\pi,\theta_1,\theta_2}(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y})\right] =$$
$$\sum_{i=1}^{n_1} lnp_{\theta_1}(x_i) + \sum_{i=n_1+1}^{n_1+n_2} lnp_{\theta_2}(x_j) +$$
$$+\sum_{i=1}^{n_1} \check{z}_i lnp_{\theta_2}(y_i-1) + \sum_{j=n_1+1}^{n_1+n_2} \check{z}_j lnp_{\theta_1}(y_j-1) +$$
$$+\sum_{i=1}^{n_1} (1-\check{z}_i) lnp_{\theta_2}(y_i) + \sum_{j=n_1+1}^{n_1+n_2} (1-\check{z}_j) lnp_{\theta_1}(y_j) +$$
$$+\sum_{j=1}^{n_1+n_2} \check{z}_j \ln\pi + \sum_{j=1}^{n_1+n_2} (1-\check{z}_j)\ln(1-\pi)$$

where

- $\check{z}_i = E_{\pi_0,\theta_{2o}}\left(Z_i \middle| Y_i^{(1)} = y_i\right) = \frac{p_{\theta_{20}}(y_i-1)\pi_0}{p_{\theta_{20}}(y_i-1)\pi_0 + p_{\theta_{20}}(y_i)(1-\pi_0)}$ for $i = 1, \dots, n_1$
- $\check{z}_j = E_{\pi_0,\theta_{10}}\left(Z_j \middle| Y_j^{(2)} = y_j\right) = \frac{p_{\theta_{10}}(y_j-1)\pi_0}{p_{\theta_{10}}(y_j-1)\pi_0 + p_{\theta_{10}}(y_j)(1-\pi_0)}$ for $j = n_1+1, \dots, n_1+n_2$

ML estimation

We assume that $X^{(1)}$, $X^{(2)}$ can be modelled by either $Poisson(\lambda)$ or negative binomial $NB(r, p)$ distributions

ML estimation via EM algorithm:

- Poisson-Poisson (neutral questions) model
- Poisson-NB (neutral questions) model
- NB-NB (neutral questions) model

Model selection:

- Over-dispertion test
- Chi-squared test
- AIC, BIC

ML estimators via EM iterative algorithm Poisson-Poisson (neutral questions) model

- E Step:

$$E\left(Z_i|Y_i^{(1)}\right) = \frac{y_i{}^{(1)}\pi}{y_i{}^{(1)}\pi+\lambda_2(1-\pi)} \text{ for } i = 1,\dots,n_1$$

$$E\left(Z_j|Y_j^{(2)}\right) = \frac{y_j{}^{(2)}\pi}{y_j{}^{(2)}\pi+\lambda_1(1-\pi)} \text{ for } j = n_1+1,\dots,n_1+n_2$$

- M step:

$$\pi = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} z_i + \sum_{j=n_1+1}^{n_1+n_2} z_j\right)$$

$$\lambda_1 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} x_i^{(1)} + \sum_{j=n_1+1}^{n_1+n_2}\left(y_j{}^{(2)} - z_j\right)\right)$$

$$\lambda_2 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}\left(y_i{}^{(1)} - z_i\right) + \sum_{j=n_1+1}^{n_1+n_2} x_j^{(2)}\right)$$

RMSE of EBLUE and MM estimators

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim Poisson(2)$, $X^{(2)} \sim Poisson(2)$				
$n = 500$	0.090	0.090	0.090	0.092
$n = 1000$	0.063	0.064	0.065	0.065
$n = 2000$	0.045	0.046	0.046	0.046
Original model $X \sim Poisson(2)$				
$n = 500$	0.129	0.128	0.130	0.130
$n = 1000$	0.091	0.091	0.091	0.092
$n = 2000$	0.064	0.064	0.065	0.065

Simulations: RMSE of ML estimators

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim Poisson(2), X^{(2)} \sim Poisson(2)$				
$n = 500$	0.070	0.080	0.086	0.086
$n = 1000$	0.052	0.061	0.063	0.061
$n = 2000$	0.040	0.045	0.044	0.043
Original model $X \sim Poisson(2)$				
$n = 500$	0.097	0.104	0.118	0.121
$n = 1000$	0.070	0.080	0.087	0.086
$n = 2000$	0.053	0.060	0.062	0.060

RMSE of EBLUE and ML estimators

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim Poisson(2)$, $X^{(2)} \sim Poisson(2)$ EBLUE estimators				
$n = 500$	0.090	0.090	0.090	0.092
$n = 1000$	0.063	0.064	0.065	0.065
$n = 2000$	0.045	0.046	0.046	0.046
New model $X^{(1)} \sim Poisson(2)$, $X^{(2)} \sim Poisson(2)$ ML estimators				
$n = 500$	0.070	0.080	0.086	0.086
$n = 1000$	0.052	0.061	0.063	0.061
$n = 2000$	0.040	0.045	0.044	0.043

RMSE of restricted EBLUE and ML estimators

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim Poisson(2)$. $X^{(2)} \sim Poisson(2)$ restr. EBLUE estimators				
$n = 500$	0.070	0.080	0.089	0.092
$n = 1000$	0.052	0.061	0.065	0.065
$n = 2000$	0.040	0.045	0.046	0.046
New model $X^{(1)} \sim Poisson(2)$. $X^{(2)} \sim Poisson(2)$ ML estimators				
$n = 500$	0.070	0.080	0.086	0.086
$n = 1000$	0.052	0.061	0.063	0.061
$n = 2000$	0.040	0.045	0.044	0.043

Pitman closeness (ML, restricted EBLUE)

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim \text{Poisson}(2)$. $X^{(2)} \sim \textit{Poisson}(2)$				
$n = 500$	0.648	0.569	0.545	0.556
$n = 1000$	0.603	0.551	0.543	0.554
$n = 2000$	0.589	0.528	0.546	0.564

Pitman closeness (ML, restricted EBLUE)

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim \mathrm{NB}(\mathrm{r} = 2; \mathrm{p} = 0.4)$. $X^{(2)} \sim NB(\mathrm{r} = 2. \mathrm{p} = 0.4)$				
$n = 500$	0.625	0.589	0.587	0.620
$n = 1000$	0.600	0.559	0.597	0.623
$n = 2000$	0.573	0.562	0.593	0.618

Pitman closeness (ML, restricted EBLUE)

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim Poisson(2)$, $X^{(2)} \sim NB(\mathrm{r} = 2, \mathrm{p} = 0.4)$				
$n = 500$	0.623	0.566	0.549	0.578
$n = 1000$	0.588	0.541	0.562	0.577
$n = 2000$	0.559	0.534	0.551	0.579

Thank you for you attention ☺