Kowalczyk, Barbara; Wieczorkowski, Robert

Article

New improved poisson and negative binomial item count techniques for eliciting truthful answers to sensitive questions

Statistics in Transition new series (SiTns)

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Kowalczyk, Barbara; Wieczorkowski, Robert (2022) : New improved poisson and negative binomial item count techniques for eliciting truthful answers to sensitive questions, Statistics in Transition new series (SiTns), ISSN 2450-0291, Sciendo, Warsaw, Vol. 23, Iss. 1, pp. 75-88,
https://doi.org/10.2478/stattrans-2022-0005

STATISTICS IN TRANSITION new series, March 2022
Vol. 23 No. 1, pp. 75–88, DOI 10.2478/stattrans-2022-0005
Received – 21.12.2020; accepted – 21.11.2021

New improved Poisson and negative binomial item count techniques for eliciting truthful answers to sensitive questions

Barbara Kowalczyk[1], **Robert Wieczorkowski**[2]

ABSTRACT

Item count techniques (ICTs) are indirect survey questioning methods designed to deal with sensitive features. These techniques have gained the support of many applied researchers and undergone further theoretical development. Latterly in the literature, two new item count methods, called Poisson and negative binomial ICTs, have been proposed. However, if the population parameters of the control variable are not provided by the outside source, the methods are not very efficient. Efficiency is an important issue in indirect methods of questioning due to the fact that the protection of respondents' privacy is usually achieved at the expense of the efficiency of the estimation. In the present paper we propose new improved Poisson and negative binomial ICTs, in which two control variables are used in both groups, although in a different manner. In the paper we analyse best linear unbiased and maximum likelihood estimators of the proportion of the sensitive attribute in the population in the introduced new models. The theoretical findings presented in the paper are supported by a comprehensive simulation study. The improved procedure allowed the increase of the efficiency of the estimation compared to the original Poisson and negative binomial ICTs.

Key words: sensitive questions, indirect questioning methods, item count techniques, Poisson ICT, negative binomial ICT, EM algorithm

1. Methodology and questionnaire design

Reliable data on stigmatizing, socially unaccepted or illegal features are very hard to obtain in direct questioning. Many indirect methods of questioning have been developed to help in eliciting honest answers to sensitive questions and to eliminate the social desirability bias. Among them two methods are predominant: randomised response techniques (Warner 1965, Chaudhuri 2011, Imai 2015, Dihidar and

[1] SGH Warsaw School of Economics, Collegium of Economic Analysis, Poland.
E-mail: bkowal@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-5407-3438.
[2] Statistics Poland, Programming and Coordination of Statistical Surveys Department, Poland.
E-mail: R.Wieczorkowski@stat.gov.pl. ORCID: https://orcid.org/0000-0003-2706-4306.

Bhattacharya, 2017) and item count techniques (Miller, 1984, Blair and Imai, 2012, Chaudhuri and Christofides, 2007, Imai, 2011, Holbrook and Krosnick, 2010, Comsa and Postelnicu, 2013, Wolter and Laier, 2014, Kuha and Jackson, 2014, Trappman et al., 2014, Kowalczyk and Wieczorkowski, 2017, Krumpal et al., 2018). Item count techniques have many practical advantages (Tourangeau and Yan, 2007). They are very easy to implement, they do not require the use of any randomize device, and they are very easy to understand so the respondents realize how their privacy is being protected.

Latterly Tian et al. (2017) proposed new item count techniques, called Poisson and negative binomial ICTs. In their method (if the population parameters of the control variable are not given from the outside source) a sample of *n* elements is divided into a control group and a treatment group, of n_1 and n_2 elements respectively. Respondents in the control group are asked one neutral question with possible count outcomes 0,1,2,… An exemplary questionnaire might look like the following:

Q: How many times did you use an Uber last month? Your answer is ….

Respondents in the treatment group are presented with two questions: one exactly the same as in the control group, and one sensitive with possible outcomes 0 or 1. Respondents in the treatment group are asked to report only the sum of the two questions. An exemplary questionnaire might look like the following:

Q: How many times did you use an Uber last month?

S: Have you ever bribed an official? Assign number 1 if 'yes' (YES = 1) and number 0 if 'not' (NOT = 0).

Please report ONLY the sum of the two numbers. The sum is …

To increase efficiency of the estimation we propose a new item count method, which draws on the idea of the Poisson and negative binomial ICTs introduced by Tian et al. (2017) and advances the original method in order to attain greater efficiency of the estimation. Our improved technique incorporates the sensitive question in two groups and combines it with two different neutral questions. Below we describe the newly proposed methodology.

We divide the sample of *n* elements into the first and second treatment groups, of n_1 and n_2 elements respectively. In the first group respondents are asked one neutral question Q_1 with possible count outcomes 0,1,2,… Then respondents are presented with two questions, one neutral Q_2 with possible count outcomes 0,1,2,…, and one sensitive S with possible outcomes 0 or 1. To protect their privacy respondents are asked to report only the sum of their answers to questions Q_2 and S. They are never asked to report their answer to the sensitive question *S*. Below we give an exemplary questionnaire for the first treatment group.

Q_1: How many times did you use a taxi last month? Your answer is ….

Now, we show you two questions. Do not answer them until you read the end.

Q_2: How many times were you at the cinema last month?

S: Have you ever bribed an official? Assign number 1 if 'yes' (YES = 1) and number 0 if 'not' (NOT = 0).

Please report ONLY the sum of the two numbers. The sum is …

In the second treatment group neutral questions are switched. Therefore, an exemplary questionnaire for the second group is given below.

Q_2: How many times were you at the cinema last month? Your answer is ….

Now, we show you two questions. Do not answer them until you read the end.

Q_1: How many times did you use a taxi last month? Remember your number but do not reveal it.

S: Have you ever bribed an official? Assign number 1 if 'yes' (YES = 1) and number 0 if 'not' (NOT = 0).

Please report ONLY the sum of the two numbers. The sum is …

It is very important that the sensitive question is mentioned only once in each group and the respondents are never asked to answer the sensitive question directly. To assure complete privacy the two neutral questions should be unrelated with each other and unrelated with the sensitive question. It also ensures that the privacy protection level in the newly proposed methods is exactly the same as in the original Poisson and negative binomial ICTs.

2. Statistical model and estimation

2.1. Notation

Let $X^{(1)}$, $X^{(2)}$ denote control variables being the answers to the neutral questions Q_1 and Q_2 respectively, and let Z denote a Bernoulli distributed variable being the answer to the sensitive question *S*. To assure complete protection of the privacy we assume that $X^{(1)}, X^{(2)}, Z$ are independent. Let $P(Z = 1) = \pi$ be an unknown sensitive proportion under study. Let $Y^{(1)}$ denote an observed variable indicating the sum of answers to questions Q_2 and *S* in the first treatment group, i.e. $Y^{(1)} = X^{(2)} + Z$. Analogously, let $Y^{(2)}$ be an observed variable indicating the sum of answers to questions Q_1 and *S* in the second treatment group, i.e. $Y^{(2)} = X^{(1)} + Z$. In the first treatment group we have two vectors of observed variables: $\left(X_1^{(1)}, \dots, X_{n_1}^{(1)}\right)$ and $\left(Y_1^{(1)}, \dots, Y_{n_1}^{(1)}\right)$. In the second group vectors of observed variables are $\left(X_{n_1+1}^{(2)}, \dots, X_{n_1+n_2}^{(2)}\right)$ and $\left(Y_{n_1+1}^{(2)}, \dots, Y_{n_1+n_2}^{(2)}\right)$. Sensitive variable under study *Z* is not directly observable in this model. *Z* is a latent variable, which is in line with the principle of privacy protection.

2.2. Best linear unbiased estimator

We consider a linear estimator of the form

$$\hat{\pi} = \sum_{i=1}^{n_1} \alpha_i X_i^{(1)} + \sum_{i=1}^{n_1} \beta_i Y_i^{(1)} + \sum_{j=n_1+1}^{n_1+n_2} \gamma_i X_i^{(2)} + \sum_{i=n_1+1}^{n_1+n_2} \delta_i Y_i^{(2)}, \quad (1)$$

where $\alpha, \beta, \gamma, \delta$ are constants weight factors. We determine $\alpha, \beta, \gamma, \delta$ so as to minimize variance $Var(\hat{\pi})$ of the estimator $\hat{\pi}$ subject to the condition that this estimator is unbiased.

Conditions for unbiasedness are

$$\begin{cases} \sum_{i=1}^{n_1} \alpha_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i = 0 \\ \sum_{i=1}^{n_1} \beta_i + \sum_{j=n_1+1}^{n_1+n_2} \gamma_i = 0 \\ \sum_{i=1}^{n_1} \beta_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i = 1 \end{cases} \quad (2)$$

To achieve the smallest variance, the expression to be minimized is

$$Var(\hat{\pi}) - \lambda_1\left(\sum_{i=1}^{n_1} \alpha_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i\right) - \lambda_2\left(\sum_{i=1}^{n_1} \beta_i + \sum_{j=n_1+1}^{n_1+n_2} \gamma_i\right) - \\ -\lambda_3\left(\sum_{i=1}^{n_1} \beta_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i - 1\right) \quad (3)$$

The minimization leads to the best linear unbiased estimator (BLUE) of the sensitive population proportion π, which can be written in the final form

$$\hat{\pi} = w\left(\bar{Y}^{(2)} - \bar{X}^{(1)}\right) + (1-w)\left(\bar{Y}^{(1)} - \bar{X}^{(2)}\right) \quad (4)$$

where

$$w = \frac{Var\left(\bar{Y}^{(1)}\right) + Var\left(\bar{X}^{(2)}\right)}{Var\left(\bar{Y}^{(2)}\right) + Var\left(\bar{X}^{(1)}\right) + Var\left(\bar{Y}^{(1)}\right) + Var\left(\bar{X}^{(2)}\right)} \quad (5)$$

Variance of the BLUE estimator is

$$Var(\hat{\pi}) = \frac{\frac{1}{n_1 n_2}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right)\left(2Var\left(X^{(2)}\right) + \pi(1-\pi)\right)}{\frac{1}{n_1}\left(2Var\left(X^{(2)}\right) + \pi(1-\pi)\right) + \frac{1}{n_2}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right)} \quad (6)$$

For $n_1 = n_2 = 0.5n$ and for $Var\left(X^{(2)}\right) = Var\left(X^{(1)}\right)$ formula (6) simplifies to the form

$$Var(\hat{\pi}) = \frac{1}{n}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right) \quad (7)$$

For $n_1 = n_2 = 0.5n$ variance of the method of moment estimator in original Tian et al. (2017) Poisson and negative binomial ICTs with one neutral variable $X^{(1)}$ is

$$Var\left(\hat{\pi}^{orig}\right) = \frac{2}{n}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right) \quad (8)$$

From (7) and (8) it can be easily seen that for $n_1 = n_2 = 0.5n$ and for $Var\left(X^{(2)}\right) = Var\left(X^{(1)}\right)$ we get

$$Var(\hat{\pi}) = 0.5Var\left(\hat{\pi}^{orig}\right) \quad (9)$$

and the theoretical BLUE estimator in the improved model is more efficient than the method of moment estimator in the original model. Due to the fact that variances that

appear in formula (5) are not known in advance the theoretical BLUE estimator cannot be used directly. Therefore, we propose to use in practice the empirical BLUE estimator (EBLUE) of the form

$$\hat{\pi}^{emp} = \hat{w}^{emp}\left(\bar{Y}^{(2)} - \bar{X}^{(1)}\right) + \left(1 - \hat{w}^{emp}\right)\left(\bar{Y}^{(1)} - \bar{X}^{(2)}\right) \tag{10}$$

where

$$w^{emp} = \frac{\frac{1}{n_1}s^2\left(Y^{(1)}\right) + \frac{1}{n_2}s^2\left(X^{(2)}\right)}{\frac{1}{n_2}s^2\left(Y^{(2)}\right) + \frac{1}{n_1}s^2\left(X^{(1)}\right) + \frac{1}{n_1}s^2\left(Y^{(1)}\right) + \frac{1}{n_2}s^2\left(X^{(2)}\right)} \tag{11}$$

and $s^2\left(X^{(1)}\right)$, $s^2\left(X^{(2)}\right)$, $s^2\left(Y^{(1)}\right)$, $s^2\left(Y^{(2)}\right)$ are sample variances of observed variables $X^{(1)}$, $X^{(2)}$, $Y^{(1)}$, $Y^{(2)}$ respectively. Properties of the proposed EBLUE estimator of the sensitive proportion π are analyzed in Section 3.

2.3. Maximum likelihood estimation via EM algorithm

In our model, the sensitive variable under study $Z \sim Bernoulli(\pi)$ is not directly observable. In order to obtain maximum likelihood estimators in models with latent variables it is convenient to use expectation maximization (EM) algorithm introduced by Dempster et al. (1977) and further developed by, e.g. McLachlan and Krishnan (2008). EM algorithm has also become a standard tool for determining ML estimators when dealing with item count techniques, see, e.g. Imai (2011), Kuha and Jackson (2014), Tian et al. (2017). Complete log-likelihood function in our model is

$$\begin{aligned} &lnL_{com}\left(\pi, \theta_1, \theta_2; x^{(1)}, x^{(1)}, x^{(2)}, y^{(1)}, y^{(2)}, z\right) = \\ &= \sum_{i=1}^{n_1} lnp_{\theta_1}(x_i) + \sum_{j=n_1+1}^{n_1+n_2} lnp_{\theta_2}(x_j) + \\ &+ \sum_{i=1}^{n_1} z_i lnp_{\theta_2}(y_i - 1) + \sum_{j=n_1+1}^{n_1+n_2} z_j lnp_{\theta_1}(y_j - 1) + \\ &+ \sum_{i=1}^{n_1}(1 - z_i) lnp_{\theta_2}(y_i) + \sum_{j=n_1+1}^{n_1+n_2}(1 - z_j) lnp_{\theta_1}(y_j) + \\ &+ \sum_{j=1}^{n_1+n_2} z_j \ln \pi + \sum_{j=1}^{n_1+n_2}(1 - z_j)\ln(1 - \pi), \end{aligned} \tag{12}$$

where $p_{\theta_1}(x)$ and $p_{\theta_2}(x)$ are probability mass functions of the control variables $X^{(1)}$ and $X^{(2)}$ respectively. Conditional expectation computed in E-step of the EM algorithm is

$$\begin{aligned} &E_{\pi_0, \theta_{10}, \theta_{2o}}[lnL_{com}(\pi, \theta_1, \theta_2; y, Z | Y = y)] = \\ &= \sum_{i=1}^{n_1} lnp_{\theta_1}(x_i) + \sum_{i=n_1+1}^{n_1+n_2} lnp_{\theta_2}(x_j) + \\ &+ \sum_{i=1}^{n_1} \check{z}_i lnp_{\theta_2}(y_i - 1) + \sum_{j=n_1+1}^{n_1+n_2} \check{z}_j lnp_{\theta_1}(y_j - 1) + \\ &+ \sum_{i=1}^{n_1}(1 - \check{z}_i) lnp_{\theta_2}(y_i) + \sum_{j=n_1+1}^{n_1+n_2}(1 - \check{z}_j) lnp_{\theta_1}(y_j) + \\ &+ \sum_{j=1}^{n_1+n_2} \check{z}_j \ln \pi + \sum_{j=1}^{n_1+n_2}(1 - \check{z}_j)\ln(1 - \pi) \end{aligned} \tag{13}$$

where

$$\check{z}_i = E_{\pi_0,\theta_{2o}}\left(Z_i \middle| Y_i^{(1)} = y_i\right) = \frac{p_{\theta_{20}}(y_i-1)\pi_0}{p_{\theta_{20}}(y_i-1)\pi_0+p_{\theta_{20}}(y_i)(1-\pi_0)} \text{ for } i = 1,\dots,n_1 \tag{14}$$

$$\check{z}_j = E_{\pi_0,\theta_{10}}\left(Z_j \middle| Y_j^{(2)} = y_j\right) = \frac{p_{\theta_{10}}(y_j-1)\pi_0}{p_{\theta_{10}}(y_j-1)\pi_0+p_{\theta_{10}}(y_j)(1-\pi_0)} \text{ for } j = n_1+1,\dots,n_1+n_2 \tag{15}$$

To represent distribution of the count data Poisson and negative binomial distributions are commonly used. Therefore we consider three different cases: when both neutral variables follow Poisson distribution, when one neutral variable follows Poisson and the other neutral variable follows negative binomial distribution, and the last case, when both neutral variables follow negative binomial distributions.

Consider the first case where both neutral variables follow Poisson distribution. In this case we have $X^{(1)}\sim Poisson(\lambda_1)$, $X^{(2)}\sim Poisson(\lambda_2)$. Based on (12-15) we derive final iterative formulas for ML estimators via E-step and M-step of the EM algorithm as below.

E Step:

$$\check{z}_i = E\left(Z_i|Y_i^{(1)}\right) = \frac{y_i{}^{(1)}\pi}{y_i{}^{(1)}\pi+\lambda_2(1-\pi)} \text{ for } i = 1,\dots,n_1 \tag{16}$$

$$\check{z}_j = E\left(Z_j|Y_j^{(2)}\right) = \frac{y_j{}^{(2)}\pi}{y_j{}^{(2)}\pi+\lambda_1(1-\pi)} \text{ for } j = n_1+1,\dots,n_1+n_2 \tag{17}$$

M step:

$$\hat{\pi} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}\check{z}_i + \sum_{j=n_1+1}^{n_1+n_2}\check{z}_j\right) \tag{18}$$

$$\hat{\lambda}_1 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}x_i^{(1)} + \sum_{j=n_1+1}^{n_1+n_2}\left(y_j{}^{(2)} - \check{z}_j\right)\right) \tag{19}$$

$$\hat{\lambda}_2 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(y_i{}^{(1)} - \check{z}_i) + \sum_{j=n_1+1}^{n_1+n_2}x_j^{(2)}\right) \tag{20}$$

When one variable follows Poisson distribution and the other one follows negative binomial distribution, say $X^{(1)}\sim Poisson(\lambda)$, $X^{(2)}\sim NB(r,p)$, first we assess parameter r based on the second treatment group

$$\hat{r} = \frac{\left(\bar{x}^{(2)}\right)^2}{S^2(X^{(2)})-\bar{x}^{(2)}} \tag{21}$$

and then we derive iterative formulas for the ML estimators via E-step and M-step of the EM algorithm as below.

E Step:

$$E\left(Z_i|Y_i^{(1)}\right) = \frac{y_i{}^{(1)}\pi}{y_i{}^{(1)}\pi+\left(y_i{}^{(1)}+r-1\right)p(1-\pi)} \text{ for } i = 1,\dots,n_1 \tag{22}$$

$$E\left(Z_j|Y_j^{(2)}\right) = \frac{y_j{}^{(2)}\pi}{y_j{}^{(2)}\pi+\lambda(1-\pi)} \text{ for } j = n_1+1,\dots,n_1+n_2 \tag{23}$$

M step:

$$\hat{\pi} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} z_i + \sum_{j=n_1+1}^{n_1+n_2} z_j\right) \tag{24}$$

$$\hat{\lambda} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} x_i^{(1)} + \sum_{j=n_1+1}^{n_1+n_2}\left(y_j{}^{(2)} - z_j\right)\right) \tag{25}$$

$$\hat{\boldsymbol{p}} = \frac{\left(\sum_{i=1}^{\boldsymbol{n_1}}(y_i{}^{(1)}-\boldsymbol{z_i})+\sum_{j=\boldsymbol{n_1}+1}^{\boldsymbol{n_1}+\boldsymbol{n_2}}\boldsymbol{x_j^{(2)}}\right)}{(\boldsymbol{n_1}+\boldsymbol{n_2})r+\left(\sum_{i=1}^{\boldsymbol{n_1}}(y_i{}^{(1)}-\boldsymbol{z_i})+\sum_{j=\boldsymbol{n_1}+1}^{\boldsymbol{n_1}+\boldsymbol{n_2}}\boldsymbol{x_j^{(2)}}\right)} \tag{26}$$

When both neutral variables follow negative binomial distribution, say $X^{(1)}\sim NB(r_1, p_1)$ and $X^{(2)}\sim NB(r_2, p_2)$, we first assess parameters r_1, r_2 based on the first and second treatment groups respectively by:

$$\hat{r_1} = \frac{\left(\bar{x}^{(1)}\right)^2}{S^2(X^{(1)})-\bar{x}^{(1)}} \tag{27}$$

$$\hat{r_2} = \frac{\left(\bar{x}^{(2)}\right)^2}{S^2(X^{(2)})-\bar{x}^{(2)}} \tag{28}$$

Next we derive formulas necessary to implement the EM algorithm.

E Step:

For $i = 1, \dots, n_1$:

$$E\left(Z_i|Y_i{}^{(1)}\right) = \frac{y_i{}^{(1)}\pi}{y_i{}^{(1)}\pi+\left(y_i{}^{(1)}+r_2-1\right)p_2(1-\pi)} \tag{29}$$

For $j = n_1 + 1, \dots, n_1 + n_2$:

$$E\left(Z_j|Y_j{}^{(2)}\right) = \frac{y_j{}^{(2)}\pi}{y_j{}^{(2)}\pi+\left(y_j{}^{(2)}+r_1-1\right)p_1(1-\pi)} \tag{30}$$

M step:

$$\hat{\pi} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} z_i + \sum_{j=n_1+1}^{n_1+n_2} z_j\right) \tag{31}$$

$$\hat{p}_1 = \frac{\left(\sum_{i=1}^{n_1} x_i^{(1)}+\sum_{j=n_1+1}^{n_1+n_2}(y_j{}^{(2)}-z_j)\right)}{(n_1+n_2)r_1+\left(\sum_{i=1}^{n_1} x_i^{(1)}+\sum_{j=n_1+1}^{n_1+n_2}(y_j{}^{(2)}-z_j)\right)} \tag{32}$$

$$\hat{p}_2 = \frac{\left(\sum_{i=1}^{n_1}(y_i{}^{(1)}-z_i)+\sum_{j=n_1+1}^{n_1+n_2} x_j^{(2)}\right)}{(n_1+n_2)r_2+\left(\sum_{i=1}^{n_1}(y_i{}^{(1)}-z_i)+\sum_{j=n_1+1}^{n_1+n_2} x_j^{(2)}\right)} \tag{33}$$

3. Simulation studies

To examine properties of the proposed improved Poisson and negative binomial ICTs and compare them with original Tian et al. (2017) design we conduct a comprehensive simulation study. For each set of model parameters separately, namely for $n = 500, 1000, 2000$ and for $\pi = 0.05, 0.1, 0.2, 0.3$, we generate n independent

variables $Z_1, Z_2, \dots, Z_n$ from *Bernoulli*(π) distribution. We use these once generated variables for all models considered in this section. Next, for each set of model parameters we generate $0.5n$ independent variables $X_1^{(1)}, \dots, X_{0.5n}^{(1)}$ from $Poisson(\lambda_1)$ distribution. These variables are used for both improved and original Poisson ICTs. For the improved Poisson ICT (Poisson-Poisson model) we additionally generate $0.5n$ independent variables $X_{0.5n+1}^{(2)}, \dots, X_n^{(2)}$ from $Poisson(\lambda_2)$ distribution. Next, we generate $0.5n$ independent variables $X_1^{(1)}, \dots, X_{0.5n}^{(1)}$ from $NB(r_1, p_1)$ distribution. We use these variables for both improved and original negative binomial ICTs. For the improved negative binomial ICT (NB-NB model) we additionally generate $0.5n$ independent variables $X_{0.5n+1}^{(2)}, \dots, X_n^{(2)}$ from $NB(r_2, p_2)$ distribution. Last but not least we generate $0.5n$ independent variables $X_1^{(1)}, \dots, X_{0.5n}^{(1)}$ from $Poisson(\lambda_1)$ and $0.5n$ independent variables $X_{0.5n+1}^{(2)}, \dots, X_n^{(2)}$ from $NB(r, p)$ distribution for the improved Poisson-NB model.

Based on the generated values we obtained n realizations of the two dimensional observable variables (X, Y) in the new models

$$(X_j, Y_j) = \begin{cases} \left(X_j^{(1)}, X_j^{(2)} + Z_j\right) & for\ j = 1, \dots, 0.5n \\ \left(X_j^{(2)}, X_j^{(1)} + Z_j\right) & for\ j = 0.5n + 1, \dots, n \end{cases},$$

and in the original Tian et al. (2017) models

$$Y_j = \begin{cases} X_j^{(1)} & for\ j = 1, \dots, 0.5n \\ X_j^{(1)} + Z_j & for\ j = 0.5n + 1, \dots, n \end{cases}.$$

Finally, we calculated EBLUE and ML estimators via EM algorithm according to formulas obtained in Section 2 and analogous MM and ML estimators according to formulas given in Tian et al. (2017). This process was replicated for each set of model parameters independently 10 000 times. In the simulation study we consider values $\pi \leq 0.3$. This corresponds to applications as the proportion of individuals possessing the sensitive feature is usually not very high in the general population.

The R codes used in our simulations are available at

https://github.com/rwieczor/ICT_Poisson_Negativebinomial.

In Table 1 root mean square error and bias of empirical best linear unbiased estimator is presented for different overall sample sizes, different sensitive proportions, and different models. It should be noted that obtained values of the RMSE of the EBLUE estimators are very close to the theoretical values $\sqrt{Var(\hat{\pi})}$, where $Var(\hat{\pi})$ is the variance of the theoretical BLUE estimator given in formula (7). RMSE and bias of maximum likelihood (ML) estimator is presented in Table 2. Naturally efficiency of the estimation increases (RMSE decreases) with the increase of the sample size. By

comparing Tables 1 and 2 it can be easily seen that ML estimators are more efficient than the corresponding EBLUE estimators. Advantage of ML over EBLUE estimators in terms of efficiency is especially highly visible for the small sample sizes and small values of π. Bias of the EBLUE estimators is very small. For ML estimators bias is visible for small values of π and small values of n.

Table 1. RMSE and BIAS (in parenthesis) of the EBLUE for different model parameters in the new model

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X^{(1)} \sim Poisson(2), X^{(2)} \sim Poisson(2)$				
n = 500	0.089 (0.001)	0.090 (-0.001)	0.091 (-0.001)	0.091 (0.000)
n = 1000	0.063 (0.001)	0.065 (0.001)	0.064 (0.001)	0.064 (0.000)
n = 2000	0.045 (0.000)	0.045 (0.000)	0.046 (0.000)	0.046 (0.001)
$X^{(1)} \sim Poisson(2), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.092 (0.000)	0.093 (0.000)	0.093 (-0.001)	0.093 (0.001)
n = 1000	0.065 (0.000)	0.065 (0.000)	0.067 (0.000)	0.066 (0.000)
n = 2000	0.046 (0.000)	0.046 (0.000)	0.046 (0.000)	0.046 (0.000)
$X^{(1)} \sim NB(r = 2, p = 0.4), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.096 (0.001)	0.095 (0.001)	0.097 (0.001)	0.096 (-.002)
n = 1000	0.068 (0.000)	0.067 (-0.001)	0.068 (0.001)	0.068 (0.001)
n = 2000	0.047 (0.000)	0.048 (-0.001)	0.048 (0.000)	0.048 (0.000)

Table 2. RMSE and BIAS (in parenthesis) of the ML estimators in the new model

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X^{(1)} \sim Poisson(2), X^{(2)} \sim Poisson(2)$				
n = 500	0.070 (0.017)	0.079 (0.006)	0.087 (-0.001)	0.086 (0.000)
n = 1000	0.052 (0.008)	0.061 (0.002)	0.062 (0.001)	0.060 (0.000)
n = 2000	0.040 (0.003)	0.044 (0.000)	0.044 (0.000)	0.043 (0.001)
$X^{(1)} \sim Poisson(2), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.067 (0.015)	0.076 (0.006)	0.080 (-0.002)	0.077 (-.002)
n = 1000	0.052 (0.007)	0.057 (0.002)	0.058 (-0.001)	0.054 (-.002)
n = 2000	0.039 (0.003)	0.043 (0.000)	0.041 (0.000)	0.038 (-.001)
$X^{(1)} \sim NB(r = 2, p = 0.4), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.062 (0.013)	0.071 (0.004)	0.074 (-0.003)	0.070 (-.005)
n = 1000	0.049 (0.007)	0.054 (0.001)	0.054 (-0.001)	0.050 (-.002)
n = 2000	0.037 (0.003)	0.040 (-0.001)	0.038 (-0.001)	0.035 (-.001)

In Tables 3-4 we present RMSE of moments and ML estimators in original Tian et. al (2017) Poisson and negative-binomial ICTs. It should be noted that obtained values of the RMSE of moments estimators are very close to the theoretical values $\sqrt{Var(\hat{\pi}^{orig})}$, where $Var(\hat{\pi}^{orig})$ is given in formula (8). By determining sample sizes and the privacy protection level at the same level we can see that the new proposed models are more efficient. Gain in efficiency is achieved for all sample sizes and all values of π when comparing ML estimators in original and improved techniques and also when comparing MM with EBLUE estimators. It has to be emphasized that the new models resulted also in smaller bias when ML estimators are concerned.

Table 3. RMSE and BIAS (in parenthesis) of moments estimators in original Tian et al. (2017) Poisson and negative-binomial ICTs

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X \sim Poisson(2)$				
n = 500	0.128 (-.001)	0.128 (0.000)	0.128 (-0.002)	0.129 (-.001)
n = 1000	0.091 (0.001)	0.092 (0.000)	0.090 (0.001)	0.091 (0.000)
n = 2000	0.064 (0.000)	0.064 (0.000)	0.065 (0.000)	0.065 (0.001)
$X \sim NB(r = 2, p = 0.4)$				
n = 500	0.136 (0.003)	0.135 (0.002)	0.137 (-0.001)	0.136 (-.002)
n = 1000	0.094 (0.000)	0.095 (0.000)	0.096 (0.001)	0.096 (0.001)
n = 2000	0.068 (-0.001)	0.069 (-0.002)	0.069 (0.001)	0.068 (-.001)

Table 4. RMSE and BIAS (in parenthesis) of ML estimators in original Tian et. al (2017) Poisson and negative-binomial ICTs

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X \sim Poisson(2)$				
n = 500	0.095 (0.030)	0.105 (0.016)	0.117 (0.000)	0.120 (-.002)
n = 1000	0.071 (0.018)	0.081 (0.006)	0.086 (0.001)	0.086 (-.001)
n = 2000	0.052 (0.008)	0.06 (0.001)	0.063 (-0.001)	0.060 (0.001)
$X \sim NB(r = 2, p = 0.4)$				
n = 500	0.083 (0.024)	0.091 (0.008)	0.101 (-0.006)	0.098 (-.008)
n = 1000	0.063 (0.014)	0.071 (0.002)	0.075 (-0.002)	0.070 (-.004)
n = 2000	0.048 (0.005)	0.055 (-0.001)	0.055 (-0.001)	0.050 (-.002)

For further investigation let us consider succeeding model parameters and compare ML estimators in the improved and original Tian et al. (2017) Poisson ICT. Results of the simulation studies are given in Tables 5 and 6. In all cases both RMSE and BIAS of the ML estimators are visibly smaller when using newly proposed models.

Table 5. RMSE and BIAS (in parenthesis) of the ML estimators in the new model and original Poisson ICT

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim Poisson(1)$, $X^{(2)} \sim Poisson(1)$				
n = 500	0.053 (0.009)	0.06 (0.002)	0.063 (-0.001)	0.060 (0.000)
n = 1000	0.040 (0003)	0.045 (0.000)	0.044 (-0.001)	0.042 (-.001)
n = 2000	0.030 (0.001)	0.032 (-0.001)	0.031 (0.000)	0.030 (0.000)
Original model $X \sim Poisson(1)$				
n = 500	0.070 (0.017)	0.079 (0.006)	0.087 (-0.001)	0.084 (0.000)
n = 1000	0.053 (0.007)	0.060 (0.001)	0.062 (-0.001)	0.060 (-.001)
n = 2000	0.040 (0.003)	0.044 (0.000)	0.044 (-0.001)	0.042 (-.0010
New model $X^{(1)} \sim Poisson(3)$, $X^{(2)} \sim Poisson(3)$				
n = 500	0.083 (0.024)	0.093 (0.011)	0.106 (0.001)	0.106 (-.001)
n = 1000	0.062 (0.013)	0.071 (0.004)	0.076 (-0.002)	0.076 (-.001)
n = 2000	0.047 (0.005)	0.053 (0.001)	0.054 (0.000)	0.054 (0.000)
Original model $X \sim Poisson(3)$				
n = 500	0.116 (0.043)	0.124 (0.026)	0.141 (0.007)	0.146 (0.000)
n = 1000	0.084 (0.024)	0.093 (0.010)	0.104 (0.000)	0.108 (-.001)
n = 2000	0.061 (0.012)	0.070 (0.004)	0.076 (0.000)	0.075 (0.000)

Table 6. RMSE and BIAS (in parenthesis) of the ML estimators in the new model and original negative-binomial ICT

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim NB(r = 2, p = 0.6)$, $X^{(2)} \sim NB(r = 2, p = 0.6)$				
n = 500	0.094 (0.029)	0.102 (0.012)	0.115 (-0.002)	0.112 (-.007)
n = 1000	0.072 (0.017)	0.080 (0.004)	0.087 (-0.003)	0.081 (-.005)
n = 2000	0.055 (0.009)	0.062 (0.000)	0.062 (-0.002)	0.057 (-.002)
Original model $X \sim NB(r = 2, p = 0.6)$				
n = 500	0.124 (0.046)	0.131 (0.026)	0.147 (0.004)	0.153 (-.010)
n = 1000	0.096 (0.031)	0.103 (0.013)	0.116 (-0.003)	0.114 (-.009)
n = 2000	0.071 (0.016)	0.080 (0.003)	0.086 (-0.002)	0.082 (-.003)
New model $X^{(1)} \sim NB(r = 3, p = 0.5)$, $X^{(2)} \sim NB(r = 3, p = 0.5)$				
n = 500	0.098 (0.033)	0.108 (0.018)	0.118 (-0.001)	0.118 (-.004)
n = 1000	0.074 (0.019)	0.083 (0.008)	0.089 (-0.001)	0.085 (-.002)
n = 2000	0.056 (0.010)	0.064 (0.002)	0.064 (-0.001)	0.060 (-.001)
Original model $X \sim NB(r = 3, p = 0.5)$				
n = 500	0.133 (0.052)	0.139 (0.034)	0.153 (0.005)	0.160 (-.004)
n = 1000	0.101 (0.033)	0.108 (0.017)	0.118 (0.000)	0.120 (-.003)
n = 2000	0.073 (0.018)	0.083 (0.007)	0.089 (-0.001)	0.085 (-.003)

It is worth mentioning that in all comparisons we have set the overall sample size and privacy protection at the same level. Privacy protection is usually measured by the probability that the respondent possesses the sensitive attribute conditional on his or her answer. This probability was set to be the same in both compared methods by attaching the identical parameters to the control neutral variable associated with the sensitive one. In some surveys, however, asking a sensitive question – even indirectly – can be slightly more costly than asking the neutral one. In the new methods an indirect question about the sensitive variable is asked in the two groups and also two neutral questions are asked. Therefore, the newly proposed techniques can be slightly more costly in some situations, which also should be mentioned. However, this does not seem to apply to all surveys. Nevertheless, evident advantages of the newly proposed techniques in terms of efficiency and privacy protection should initiate its further development and application.

4. Conclusions

Item count techniques have attracted much attention among applied researchers. Methodology and theory of this method is still being developed, with a significant contribution by Tian et al. (2017), who introduced Poisson and negative binomial item count techniques. The two techniques allow for eliciting honest answers to sensitive questions, simplify the questionnaire design and theory. But this effect is achieved at the expense of the efficiency of the estimation, which is not high in the proposed techniques. In the paper three new models are proposed: Poisson-Poisson neutral questions ICT, Poisson-negative binomial neutral questions ICT, and negative binomial-negative binomial neutral questions ICT. Newly proposed methods maintain privacy of respondents at the same level regarding the sensitive question. At the same time the three newly proposed techniques increase efficiency of the estimation, which is very important in indirect methods of questioning.

References

Blair, G., Imai, K., (2012). Statistical Analysis of List Experiments. Polit Anal 20, pp. 47–77.

Chaudhuri, A., (2011). Randomized response and indirect questioning techniques in surveys, CRC Press, Boca Raton, FL.

Chaudhuri, A., Christofides, T. C., (2007). Item Count Technique in estimating the proportion of people with a sensitive feature. J Stat Plann Inference 137, pp. 589–593.

Comsa, M., Postelnicu C., (2013). Measuring Social Desirability Effects on Self-Reported Turnout Using the Item-Count Technique. Int J Public Opin Res 25, pp. 153–172.

Dempster, A.P., Laird, N. M., Rubin, D. B., (1977). Maximum-likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society Series B, Vol. 39, pp. 1–37.

DIHIDAR, K., BHATTACHARYA, M., (2017). Estimating sensitive population proportion using a combination of binomial and hypergeometric randomized responses by direct and inverse mechanism, Statistics in Transition new series, Vol. 18, No. 2, pp. 193–210.

Holbrook, A. L., Krosnick, J. A., (2010). Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique. Public Opin Quart 74, pp. 37–67.

Imai, K., (2011). Multivariate regression analysis for the item count technique, Journal of the American Statistical Association, Vol. 206, pp. 407–416.

Imai, K., (2015). Design and Analysis of the Randomized response Technique, Journal of the American Statistical Association, Vol. 110, No. 511, pp. 1304–1319.

Kowalczyk, B., Wieczorkowski, R., (2017). Comparing Proportions of sensitive Items in Two Populations when Using Poisson and Negative Binomial Item Count Techniques, Quantitative Methods in Economics, Vol. 18, pp. 68–77.

Krumpal, I., Jann, B., Korndörfer, M., Schmukle, S., (2018). Item Sum Double-List Technique: An Enhanced Design for Asking Quantitative Sensitive Questions, Survey Research Methods, Vol. 12, pp. 91–102.

Kuha, J., Jackson, J., (2014). The item count method for sensitive survey questions: modeling criminal behavior, Journal of the Royal Statistical Society Series C, Vol. 63, pp. 321–341.

Mclachlan, G. J., Krishnan, T., (2008). EM Algorithm and Extensions, Wiley Series in Probability and Statistics.

Miller Jd. (1994). A new survey technique for studying deviant behavior. PhD Thesis, The George Washington University, USA, 1984.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL https://www.R-project.org/.

Tian, G-L., Tang, M-L., Wu., Q, Liu Y., (2017). Poisson and negative binomial item count techniques for surveys with sensitive question, Statistical Methods in Medical Research, Vol. 26, pp. 931–947.

Tourangeau, R., Yan, T., (2007). Sensitive questions in surveys. Psychol Bull 133, pp. 859–883.

Trappman, M., Krumpal, I., Kirchner, A., Jann, B., (2014). Item Sum: A New Technique for Asking Quantitative Sensitive Questions, Journal of Survey Statistics and Methodology, Vol. 2, pp. 58–77.

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

Wolter, F., Laier, B., (2014). The Effectiveness of the Item Count Technique in Eliciting Valid Answers to Sensitive Questions. An Evaluation in the Context of Self-Reported Delinquency, Survey Research Methods, Vol. 8, pp. 153–168.