**Design and analysis of the improved Poisson and negative binomial item count techniques**

Barbara Kowalczyk, SGH Warsaw School of Economics, bkowal@sgh.waw.pl

Robert Wieczorkowski, Statistics Poland, r.wieczorkowski@stat.gov.pl

**Abstract**

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. Item count techniques (ICTs), pioneered by Miller (1984), are an example of indirect survey techniques designed to deal with sensitive features. These techniques have many practical advantages and have gained the support of many applied researchers. Recently Tian et al. (2017) proposed new item count methods called Poisson and negative binomial ICTs. The new methods give many opportunities for further theoretical and practical developments. However, if the population parameters of the control variable are not given from the outside source, the methods are not very efficient. Efficiency is an important issue in indirect methods of questioning due to the fact that 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 each of the two subsamples serve both as a control and a treatment group. This procedure allows to increase efficiency of the estimation as compared to the original Poisson and negative binomial ICTs and maintain respondents’ privacy at the same level. In the paper we introduce methodology of the proposed improved method and its statistical theory. We analyze best linear unbiased and maximum likelihood estimators of the population proportion of the sensitive attribute. We compare the improved techniques with the original Poisson and negative binomial ICTs. Improvement is obtained in terms of efficiency. Gain in efficiency is achieved without affecting the privacy of respondents. Theoretical results presented in the paper are illustrated by a comprehensive simulation study.

**Keywords:** Sensitive questions, indirect questioning, item count techniques, EM algorithm

**1. Methodology and Questionnaire Design**

Item count techniques are statistical methods of indirect questioning about sensitive feature. They have many practical advantages and gained the support by applied researchers (see e.g. Kuklinski et al., 1997, LaBrie and Earleywine, 2000, Kingsbury et al. 2003, Imai, 2011, Kuha and Jackson, 2014). Recently Tian et al (2017) proposed new ICTs, 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 is divided into control and treatment group. Respondents in the control group are asked one neutral question with possible outcomes 0,1,2, … 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. To fully protect respondents, they are asked to report only the sum of the two questions.

In our modified method, to increase efficiency of the estimation we propose to divide the sample of *n* elements into first and second treatment groups, of *n*1 and *n*2 elements respectively. In the first group respondents are first asked one neutral question with possible outcomes 0,1,2,… and then are presented with two questions, one neutral with possible outcomes 0,1,2,… , and one sensitive with possible outcomes 0 or 1. To protect their privacy they are only asked to report the sum of their answers into questions and . Exemplary questionnaire for the first treatment group is given below:

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

*Now we show you two questions.*

1. *How many times were you at the cinema last month? Remember your number but do not reveal it.*
2. *Have you ever bribed an official? Assign number 1 if ‘yes’ (YES = 1) and number 0 if ‘not’ (NOT = 0). Remember your number but do not reveal it.*

*We do not to violet your privacy. Please report ONLY the sum of the two numbers. To your answer to the question A add your answer to the question B. The sum is ………*

In the second group respondents are first asked one neutral question and then are presented with two questions, one neutral , and one sensitive . To protect their privacy they are only asked to report the sum of their answers to questions and . Analogous questionnaire for the second group is the following:

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

*Now we show you two questions:*

1. *How many times did you use a taxi last month? Remember your number but do not reveal it.*
2. *Have you ever bribed an official? Assign number 1 if ‘yes’ (YES = 1) and number 0 if ‘not’ (NOT = 0). Remember your number but do not reveal it.*

*We do not to violet your privacy. Please report ONLY the sum of the two numbers. To your answer to the question A add your answer to the question B. The sum is ………*

It is very important that the sensitive question is mentioned only once in each group.

**2. Model and Estimation**

Notation: – discrete control variable being the answer to the question ; – discrete control variable being the answer to the question , – Bernoulli distributed variable being the answer to the sensitive question. Let , where – unknown sensitive proportion under study. To assure protection we assume that , , are independent. In the first group we observe , where for . In the second group we observe , where for .

*2.1. Best Linear Unbiased Estimator*

Best linear unbiased estimator (BLUE) of the sensitive population proportion in our new model is of the form:

where

Empirical BLUE estimator is:

where

and , , , are sample variances of observed variables , , , respectively.

*2.2. Maximum Likelihood Estimation*

If both neutral variables follow Poisson distribution, i.e. , , by using EM algorithm (see: Dempster et al. 1977) we derive iterative formulas for ML estimator via E-step and M-step as below:

E Step:

for

for

M step:

If one variable follows Poisson distribution and the other one follows negative binomial distribution, say , , first we assess parameter based on group II:

and then derive iterative formulas for the EM algorithm:

E Step:

for

for

M step:

If both neutral variable follow negative binomial distribution, say and , we first we assess parameters based on groups I and II respectively by:

Next we derive formulas necessary to implement EM:

E Step:

For :

For :

M step:

**3. Simulation Study**

To examine properties of the proposed modified Poisson and negative binomial ICTs, in which each group has a role of both control and treatment group, we conduct a comprehensive simulation study. To examine Poisson-Poisson model, for each set of model parameters separately, namely for *n=*500, 1000 and we generate *n* independent variables from *Bernoulli* distribution, *n* independent variables from *Poisson* distribution. On the basis of generated values we construct for and for . Next we derive empirical BLUE and ML estimators. For each set of model parameters we repeat the process independently 10000. To examine Poisson – negative binomial model we act in the same way, except that we generate *n*/2 variables from *Poisson* distribution and from *NB*(2*;*0.4) distribution. To examine negative binomial – negative binomial model we *n* independent variables from *Bernoulli* distribution, *n* independent variables from *NB* distribution. Then we construct variables for and for and calculate empirical BLUE and ML estimators. Last but not least, to be able to compare the new models with classical Poisson and negative binomial ICTs we generate generate *n*/2 independent variables from *Bernoulli* distribution, and *n* independent variables from *Poisson* distribution. On the basis of generated values we construct for and derive MM and ML estimators according to formulas given in Tian et al. (2017). We act in the same way for negative-binomial ICT, except now we generate from NBdistribution. Number of replications in each simulation study, i.e. for each set of model parameters corresponding to one box in tables 1-4, is 10 000.

In table 1 root mean square error (RMSE) of empirical best linear unbiased (EBLUE) is presented for different overall sample sizes, different sensitive proportions, and different models. Analogous RMSE for maximum likelihood (ML) estimator is presented in table 2. As we can see efficiency of the estimation increases fast (RMSE decreases) with the increase of the sample size. Although absolute RMSE increases slightly as sensitive proportion increases, relative RMSE is definitely highest for the smallest value of the sensitive proportion . For example in Poisson-Poisson model and EBLUE estimator, for , values of the RMSE 0.063, 0.064, 0.065, 0.065 expressed as a percentage of the corresponding sensitive proportions 0.05, 0.10, 0.20, 0.30 are 126%, 64%, 32.5%, 21.7%. Comparing tables 1 and 2 it can be easily seen that ML estimators are more efficient than corresponding EBLUE estimators. Advantage of ML estimators is the greatest for small sample sizes.

**Table 1. RMSE of the EBLUE estimators for different model parameters in the new model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample size |  |  |  |  |
| , | | | | |
|  | 0.090 | 0.090 | 0.090 | 0.092 |
|  | 0.063 | 0.064 | 0.065 | 0.065 |
|  | 0.045 | 0.046 | 0.046 | 0.046 |
| , | | | | |
|  | 0.092 | 0.093 | 0.093 | 0.093 |
|  | 0.065 | 0.065 | 0.067 | 0.066 |
|  | 0.046 | 0.046 | 0.046 | 0.046 |
| *,* | | | | |
|  | 0.095 | 0.095 | 0.096 | 0.096 |
|  | 0.068 | 0.066 | 0.068 | 0.068 |
|  | 0.048 | 0.047 | 0.048 | 0.049 |

Source: Own calculations

**Table 2. RMSE of the ML estimators in the new model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample size |  |  |  |  |
| , | | | | |
|  | 0.070 | 0.080 | 0.086 | 0.086 |
|  | 0.052 | 0.061 | 0.063 | 0.061 |
|  | 0.040 | 0.045 | 0.044 | 0.043 |
| , | | | | |
|  | 0.067 | 0.076 | 0.080 | 0.077 |
|  | 0.052 | 0.057 | 0.058 | 0.054 |
|  | 0.039 | 0.043 | 0.041 | 0.038 |
| *,* | | | | |
|  | 0.063 | 0.071 | 0.074 | 0.069 |
|  | 0.049 | 0.055 | 0.053 | 0.049 |
|  | 0.037 | 0.040 | 0.037 | 0.035 |

Source: Own calculations

**Table 3. RMSE of moments estimators in original Poisson and negative-binomial ICTs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample size |  |  |  |  |
|  | | | | |
|  | 0.129 | 0.128 | 0.130 | 0.130 |
|  | 0.091 | 0.091 | 0.091 | 0.092 |
|  | 0.064 | 0.064 | 0.065 | 0.065 |
|  | | | | |
|  | 0.136 | 0.133 | 0.137 | 0.138 |
|  | 0.095 | 0.095 | 0.096 | 0.096 |
|  | 0.067 | 0.068 | 0.068 | 0.068 |

Source: Own calculations on the basis of Tian et al. (2017) formulas

**Table 4. RMSE of ML estimators in original Poisson and negative-binomial ICTs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lorem** |  |  |  |  |
|  | | | | |
|  | 0.097 | 0.104 | 0.118 | 0.121 |
|  | 0.070 | 0.080 | 0.087 | 0.086 |
|  | 0.053 | 0.060 | 0.062 | 0.060 |
|  | | | | |
|  | 0.082 | 0.091 | 0.100 | 0.099 |
|  | 0.061 | 0.070 | 0.073 | 0.070 |
|  | 0.047 | 0.054 | 0.052 | 0.049 |

Source: Own calculations on the basis of Tian et al. (2017) formulas

In tables 3-4 we present RMSE of moments and ML estimators in original Poisson and negative-binomial ICTs based on traditional control and treatment groups. By determining overall sample sizes at the same level we can see that new proposed models, in which each group serves both as a control and treatment group, are more efficient.

**4. Conclusion**

Item count techniques has 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. In the paper three new models are proposed: Poisson-Poisson neutral questions design, Poisson-negative binomial neutral questions design, and negative binomial-negative binomial neutral questions design. Proposed in the present paper modified methods maintain privacy of respondents at the same level regarding the sensitive question and at the same time increase efficiency of the estimation.

**5. References**

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, 39, pp. 1-37.

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

Kingsbury, N., Droitcour, J., Larson, E. (2003), Estimating Illegal Immigrants in a Census or Survey: the Three-Card Method – an Innovative Technique, United States General Accounting Office, Working Paper No. 3

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

Kuklinski, J.H., Cobb, M.D., Gilens, M. (1997), Racial Attitudes and the “New South”. Journal of Politics, 59, pp. 323-349.

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

LaBrie, J.W. and Earleywine, M. (2000), Sexual Risk behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched –Count Technique. J Sex Res 2000; 37:, pp. 321-326.

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, 26, pp. 931-947.