ITEM COUNT TECHNIQUES UNDER SOME ASSUMPTION VIOLATIONS

Barbara Kowalczyk
Szkoła Główna Handlowa w Warszawie
Robert Wieczorkowski
Główny Urząd Statystyczny

1-3 lipca 2025 r.
V Kongres Statystyki Polskiej
The 5th Congress of Polish Statistics

Agenda

- Introduction
  - Direct and indirect methods of questioning for eliciting truthful answers to sensitive questions
- Selected item count models
  - Poisson and negative binomial item count technique
  - Item count technique with a continuous or count control variable
- Assumption violations and robustness of models
- Monte Carlo simulations results
- Conclusions

Sensitive questions in surveys

Questions about

- private
- stigmatizing
- socially unaccepted
- illegal

behaviors, features and attributes.

Direct methods of questioning

- In direct methods of questioning the focus is on reducing measurement error (Tourangeau et al., 2000, Tourangeau and Jan, 2007, Groves et al., 2009).
- In direct questioning respondents tend to under-report undesirable sensitive features and over-report the desirable ones (Blair and Imai 2012, Krumpal 2013)

Indirect methods of questioning

- In indirect methods of questioning we do not ask the sensitive question directly
- The aim is to increase degree of privacy protection in order to elicit truthful answers to sensitive questions
- This is usually done at the cost of the more complicated questionnaire and lower efficiency of the estimation, i.e. larger sample sizes are needed
- Sensitive variable under study is not directly observable, i.e. it is a latent (hidden) variable

Indirect methods of questioning

- Randomized Response Techniques (Warner, 1965)
- Non-randomized Response Techniques (Yu et al., 2008)
- Item Count Technique (Miller, 1984)
    - Imai (2011) intoduced ML estimators via EM algorithm

- Other individual methods:
    - Three card method (Droitcour and Larson, 2002)
    - Negative Question Method (Esponda and Guerrero, 2009)
    - Bean Method (Lau et al., 2011)

(Classical) Item Count Technique

- Respondents are randomly divided into a control group and a treatment group
- Respondents in the control group are given a list of several neutral questions with binary outcomes
- Respondents in the treatment group are given a list of the same neutral questions as in the control group plus 1 sensitive question
- Respondents are asked to report only the total of their Yes answers.

Drawbacks of classical ICT

- The ceiling effect
  - If the respondent answers YES to all neutral questions and possesses the sensitive attribute then their privacy is no longer being protected
  - Most dangerous for negative - badly seen – sensitive attributes
- The floor effect
  - If the respondent answers NO to all neutral questions and does not possess the sensitive attribute then their privacy is no longer being protected
  - Most dangerous for positive - well seen – sensitive attributes

Selected alternative Item Count Methods

- Item Sum Technique (Trappman et al., 2014)
- Poisson and Negative Binomial Item Count Techniques (Tian et al., 2017)
- Item Sum Double-List Technique (Krumpal et al., 2018),
- Poisson–Poisson item count techniques (Liu et al., 2019)
- Item count technique with a continuous or count control variable (Kowalczyk et al., 2023)

Poisson and negative binomial ICTs, Tian et al. (2017)

$$Y = \begin{cases} X & in\ the\ control\ group \\ X + Z & in\ the\ treatment\ group \end{cases}$$

- X – answer to the non-sensitive question (observed in a control and hidden in a treatment group)

  $$X \in \{0,1,2,3, \dots\}$$

  $X \sim Poisson(\lambda)$ or $X \sim NB(p, r)$
- Z – answer to the sensitive question (hidden, not directly observable)

  $$Z \in \{0,1\}$$

  $Z \sim Bernoulli(\pi)$, *where* X *and* Z *are independent*
- $\hat{\pi}_{MM} = \bar{Y} - \bar{X}$

Poisson ICT, Tian et al. (2017) ML via EM algorithm

- E step (iteration t+1):

$$z_j^{(t+1)} = E\left(Z_j | Y_{obs}, \pi^{(t)}, \lambda^{(t)}\right) = \frac{y_j \hat{\pi}^{(t)}}{y_j \hat{\pi}^{(t)} + \hat{\lambda}^{(t)}\left(1 - \hat{\pi}^{(t)}\right)}, j = 1, \ldots, n_2$$

- M step (iteration t+1)

$$\hat{\pi}^{(t+1)} = \frac{1}{n_2} \sum_{j=1}^{n_2} z_j^{(t)},$$

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

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

Negative-binomial ICT, Tian et al. (2017) ML via EM algorithm

- $r_{MM,control} = \frac{(\bar{x})^2}{s^2 - \bar{x}}$
- E step (iteration t+1):

$$z_j^{(t+1)} = E\left(Z_j | Y_{obs}, \pi^{(t)}, p^{(t)}, r\right)$$

$$= \frac{y_j!\,\Gamma(y_j - 1 + r)\hat{\pi}^{(t)}}{y_j!\,\Gamma\left(y_j - 1 + r\right)\hat{\pi}^{(t)} + (y_j - 1)!\,\Gamma(y_j + r)p(1 - \hat{\pi}^{(t)})}$$

- M step (iteration t+1)

$$\hat{\pi}^{(t+1)} = \frac{1}{n_2}\sum_{j=1}^{n_2} z_j^{(t)},$$

$$\hat{p}^{(t+1)} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2}\left(y_j - z_j^{(t)}\right)}{(n_1 + n_2)r + \sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2}\left(y_j - z_j^{(t)}\right)}$$

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

ICT with a continuous or count control variable… Kowalczyk et al. (2023)

$$Y = \begin{cases} X - aZ & in\ the\ 1st\ treatment\ group \\ X + aZ & in\ the\ 2nd\ treatment\ group \end{cases}$$

- Y observed
- X continuous or count control variable (hidden)
- Z sensitive variable under study (hidden; $\text{Bernoulli}(\pi)$ distributed)
- π unknown sensitive proportion under study, $\pi = P(Z = 1)$
- $\text{Both } X \text{ and } Z$ are latent variables and are not directly observable
- $n_1(n_2) -$ number of elements in the first (second) treatment group
- $\hat{\pi}_{MM} = \frac{1}{2a}\left(\bar{Y}^{(2)} - \bar{Y}^{(1)}\right)$

ICT with a continuous or count control variable..., Kowalczyk et al. (2023)
ML via EM algorithm, $X \sim N(\mu, \sigma)$

- E step (iteration t+1):

$$\tilde{z}_i^{(t+1)} = E\left(Z_j | Y_{obs}, \pi^{(t)}, \mu^{(t)}, \sigma^{2(t)}\right)_{\pi^{(t)}}$$

$$= \frac{}{\pi^{(t)} + (1-\pi^{(t)})\exp\left(\frac{-1}{2(\sigma^2)^{(t)}}\{(y_i - \mu^{(t)})^2 - (y_i \pm 1 - \mu^{(t)})^2\}\right)}$$

- M step (iteration t+1)

$$\hat{\pi}^{(t+1)} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} \tilde{z}_i{}^{(t)},$$

$$\hat{\mu}^{(t+1)} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}\left(y_i \pm a\tilde{z}_i{}^{(t)}\right),$$

$$\widehat{\sigma^2}^{(t+1)} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}\left\{\tilde{z}_i{}^{(t)}\left(y_i \pm a - \mu^{(t)}\right)^2 + \left(1-\tilde{z}_i{}^{(t)}\right)\left(y_i - \mu^{(t)}\right)^2\right\}.$$

Source: Kowalczyk, B., Niemiro, W., & Wieczorkowski R. (2023). Item count technique with a continuous or count control variable for analyzing sensitive questions in surveys. Journal of Survey Statistics and Methodology, 11(4), 919-941. https://doi.org/10.1093/jssam/smab043

Degree of privacy protection versus efficiency in Item Count Models

- An important question arises what type of a control/neutral question should be asked, i.e. what control variable *X* should be used.
- A variable with large variance increases degree of privacy protection but at the same time decreases efficiency of the estimation.
- A variable with small variance, conversely, increases efficiency of the estimation but at the same time decreases degree of privacy protection.
- In all item count models some compromise between efficiency of the estimation and degree of privacy protection should be sought.
- Degree of privacy protection: $\mathrm{DPP} = P(Z = 1|Y = y)$
- DPP for Kowalczyk et al. (2023) model

$$P(Z = 1|Y = y) = \frac{f_\psi(y \pm a)\pi}{f_\psi(y \pm a)\pi + f_\psi(y)(1-\pi)}$$

Assumption violations and robustness of models

- In real-life surveys answer *X* to the non-sensitive question can be modeled by a theoretical distribution that best fits the observed data, which is not the same as theoretical idealized assumption that *X* follows this distribution
- MM estimator of the sensitive proportion does not depend on the distribution of the control variable *X*
- The important question arises how robust are ML (via EM) estimators to slight departures from the idealized theoretical assumption about the distribution of the control variable

Assumption violations and robustness of models

- We introduce some perturbation to the distribution of the control variable:

$$(1-\alpha)\cdot theoretical\_distribution + \alpha \cdot perturbation$$

- Perturbation by definition should be small, say $\alpha \leq 0.25$
- We conduct a Monte Carlo simulation study with 10000 replications for each set of model parameters and compare estimates obtained by MM and ML via EM formulas

Figure 1: Relative root mean square error of various estimators in Poisson ICT with perturbation being negative binomial distribution with two times higher variance

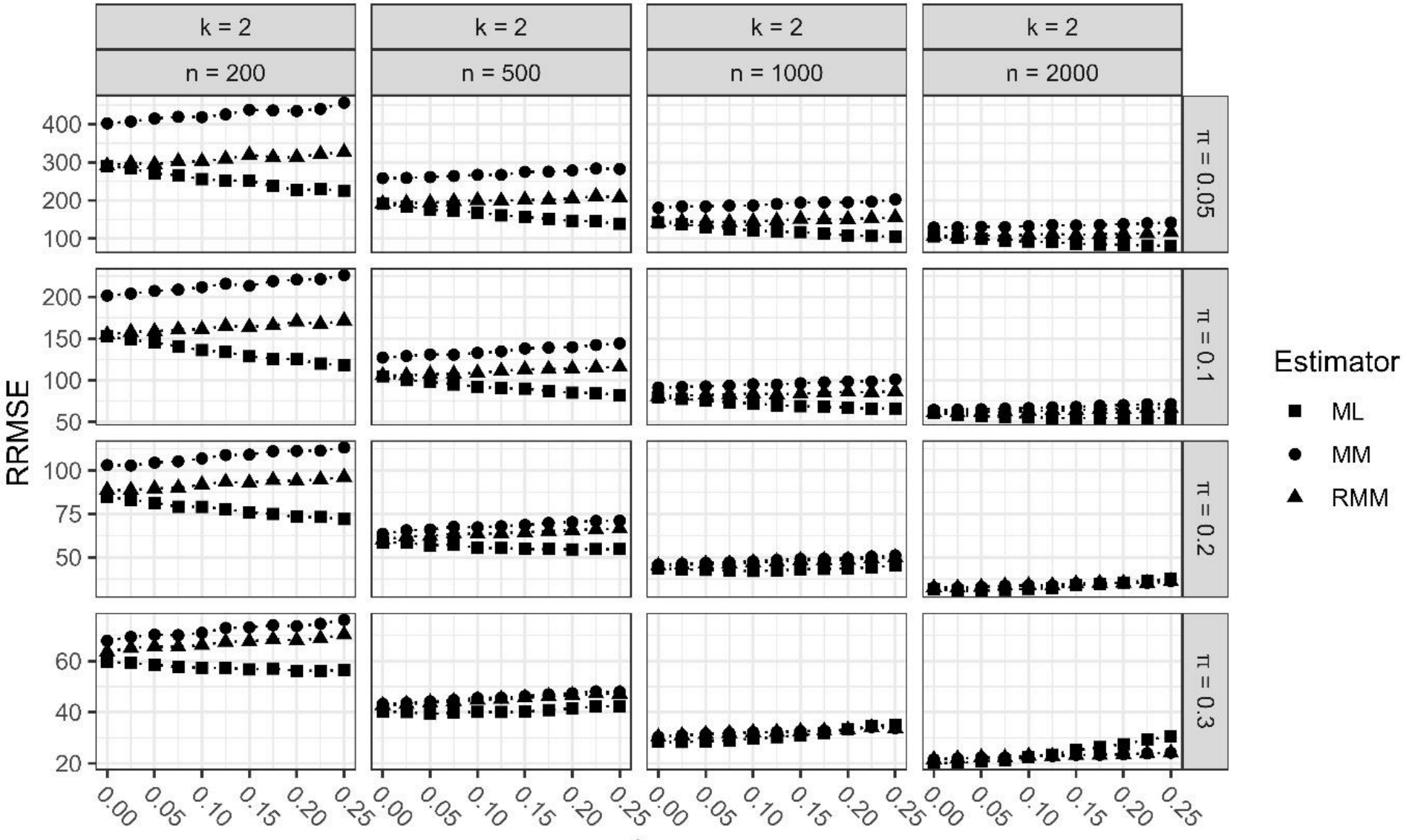


Figure 2: Relative root mean square error of various estimators in normal ICT with perturbation being log-normal distribution with two times higher variance

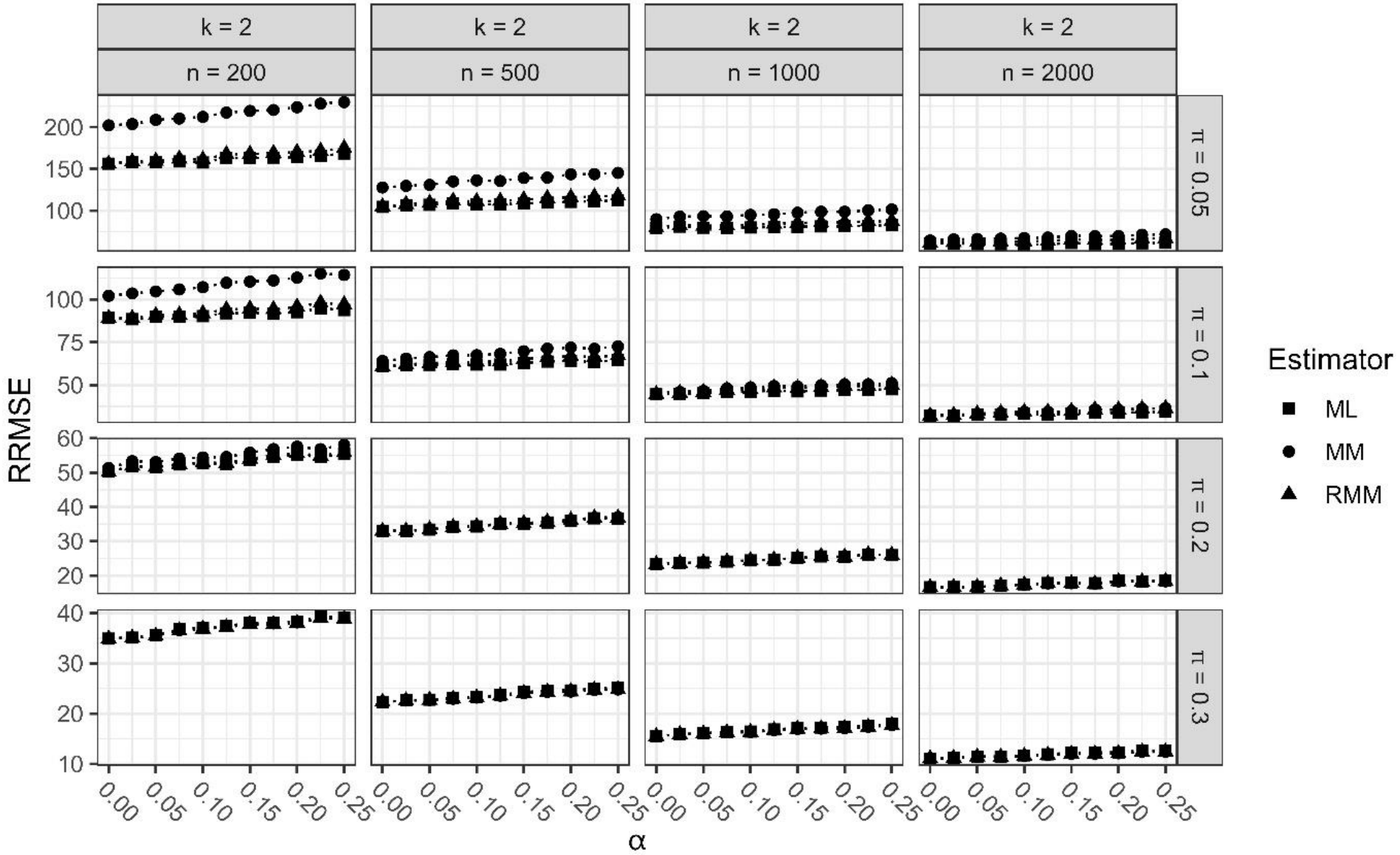
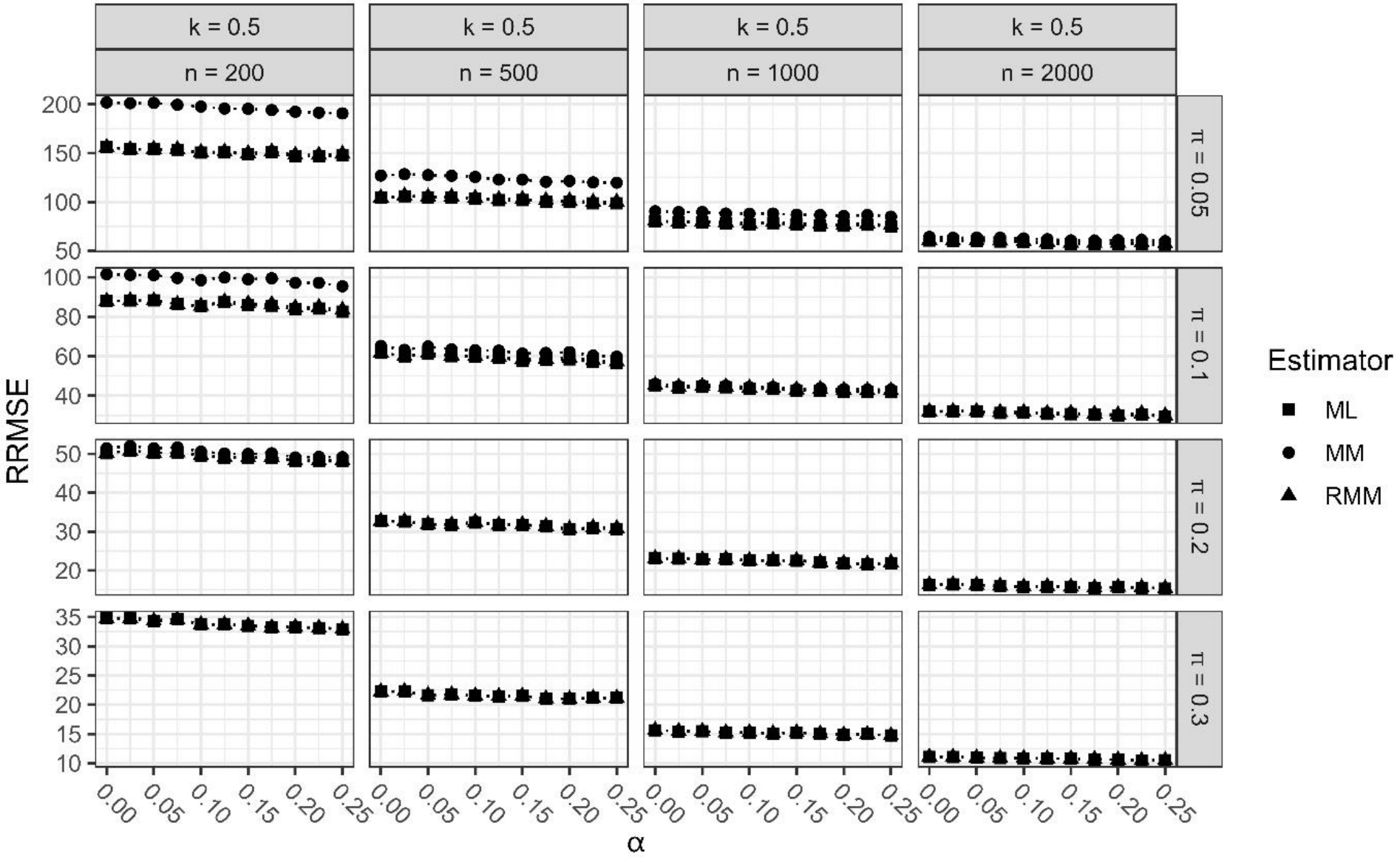


Figure 3: Relative root mean square error of various estimators in normal ICT with perturbation being log-normal distribution with two times smaller variance



Conclusions

- For small departures from the idealized theoretical distribution of the control variable, i.e. for small probability of the perturbation $\alpha \leq 0.10$ estimators obtained by ML formulas via the EM numerical algorithm are still either more efficient or equally efficient as MM estimators in all considered cases, despite the fact that MM estimators do not depend on the distribution of the control variable in item count models.
- Visible gain in efficiency is especially seen for relatively small sensitive population proportions and relatively small sample sizes.
- For moderate departures from the idealized theoretical distribution of the control variable, i.e. for probability of perturbation $0.15 \leq \alpha \leq 0.25$ results are in most cases similar, but there are several exceptions.

Conclusions

- The parametric approach broadly used in ICTs to address the latent variable has many advantages in terms of estimation. However, it also introduces some problems regarding theoretical assumptions about the distribution of the control variable.
- ML estimators via EM numerical algorithm are quite robust to the analyzed departures from the theoretical distributions of the control variable. However, there are individual exceptions. Therefore, caution is still required, and further simulations are advisable.
- In all item count models one should always look for a compromise between privacy protection, efficiency of the estimation and simplicity of the questionnaire.

References

- Blair, G., and Imai, K. (2012), Statistical Analysis of List Experiments, Political Analysis, 20, 47–77
- Droitcour J., Larson E.M. (2002), An innovative technique for asking sensitive questions: the three-card method, Bull. Méthodol. Sociol., 75, pp. 5-23
- Esponda F., Guerrero V.M. (2009), Surveys with negative questions for sensitive items, Stat. Probab Lett, 79, pp. 2456-2461
- Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009), Survey Methodology, Hoboken, NJ: John Wiley & Sons.
- Imai, K. (2011), Multivariate Regression Analysis for the Item Count Technique, Journal of the American Statistical Association, 106, 407–416.
- Kowalczyk, B., Niemiro, W., & Wieczorkowski R. (2023). Item count technique with a continuous or count control variable for analyzing sensitive questions in surveys. *Journal of Survey Statistics and Methodology*, 11(4), 919-941.
- Krumpal I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review, Quality & Quantity 47, 2025-2047.
- Krumpal, I., B. Jann, M. Kornd orfer, and S. Schmukle (2018), Item Sum Double-List Technique: An Enhanced Design for Asking Quantitative Sensitive Questions, Survey Research Methods, 12, 91–102.

References

- Lau J.T.F. , Yeung N.C.Y., Mui L.W.H. , Tsui H.Y., Gu J.(2011), A simple new method to triangulate self-reported risk behavior data – the bean method, Sex. Transm. Dis., 38 , pp. 788-792
- Liu, Y., Tian, G.-L., Wu, Q., and Tang, M.-L.. (2019). Poisson–Poisson item count techniques for surveys with sensitive discrete quantitative data, Statistical Papers, 60, 1763-1791.
- Miller, J. D. (1984), “A New Survey Technique for Studying Deviant Behavior,” PhD thesis, The George Washington University, USA.
- Tian, G.-L., M.-L. Tang, Q. Wu, and Y. Liu (2017), Poisson and Negative Binomial Item Count Techniques for Surveys with Sensitive Question, Statistical Methods in Medical Research, 26, 931–947.
- Tourangeau, R., and T. Yan (2007), Sensitive Questions in Surveys, Psychological Bulletin, 133, 859–883.
- Tourangeau R, Rips LJ., Rasinski K. (Eds.) (2000). The Psychology of Survey Response. Cambridge University Press
- Trappman, M., I. Krumpal, A. Kirchner, and B. Jann (2014), Item Sum: A New Technique for Asking Quantitative Sensitive Questions, Journal of Survey Statistics and Methodology, 2, 58–77.
- Warner SL. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J Am Stat Assoc* 60, 63-69.
- Yu J-W, Tian G-L, Tang M-L (2008) Two new models for sampling with sensitive characteristic: design and analysis. Metrika 67:251–263