EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

ASKING ABOUT PRIVATE AND SENSITIVE ATTRIBUTES USING ITEM COUNT TECHNIQUES – METHODOLOGICAL AND THEORETICAL CHALLENGES

Barbara Kowalczyk

SGH Warsaw School of Economics, Poland

bkowal@sgh.waw.pl

Robert Wieczorkowski

Statistics Poland, Poland

R.Wieczorkowski@stat.gov.pl

Sensitive questions in surveys

Questions about:

- private
- socially unaccepted
- stigmatizing
- illegal

behaviors, features and attributes.

Sensitive questions in surveys

corruption,
tax frauds,
drug use,
atypical sexual behaviors,
abortion,
Illegal work,
black market,
beating children,
politically incorrect views,
vote buying,
criminal behaviors
and so on …

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 (to obtain truthful answers to sensitive questions).
- This is usually done at the cost of the more complicated questionnaire and lower efficiency of the estimation .

Item Count Techniques (ICTs)

- Introduced by Miller (1984) Miller, J. D. (1984). A New Survey Technique for Studying Deviant Behavior. *PhD thesis*, The George Washington University, USA.

- Advanced mathematical background with maximum likelihood (ML) estimation using expectation maximization (EM) algorithm was given by Imai (2011) Imai, K. (2011). Multivariate Regression Analysis for the Item Count Technique. *Journal of the American Statistical Association*, 106, 407–416.

Classic ICT - Exemplary questionnaire

<u>Control group</u>

Below you have three questions. How many of them will you answer yes to?

- *Do you like going to the cinema?*
- *Do you like fishing?*
- *Do you like gardening?*

<u>Treatment group</u>

Below you have four questions. How many of them will you answer yes to?

- *Do you like going to the cinema?*
- *Do you like fishing?*
- *Do you like gardening?*
- *Did you cheat on your taxes last year?*

Classic Item Count Technique

The ceiling effect

If respondent answers YES to all neutral questions and possesses the sensitive attribute, then he or she is no longer being protected.

The floor effect

If respondent answers NO to all neutral questions and does not possess the sensitive attribute, then he or she is no longer being protected.

Selected New Item Count Techniques

- **Item Sum Technique** 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*, 2, 58–77.
- **Poisson and Negative Binomial Item Count Techniques** Tian, G.-L., M.-L. Tang, Q. Wu, & Y. Liu (2017). Poisson and Negative Binomial Item Count Techniques for Surveys with Sensitive Question. *Statistical Methods in Medical Research*, 26, 931–947.
- **Item Sum Double-List Technique** Krumpal, I., Jann, B., Korndorfer, M., & Schmukle, S. (2018). Item Sum Double-List Technique: An Enhanced Design for Asking Quantitative Sensitive Questions. *Survey Research Methods*, 12, 91–102.
- **Poisson–Poisson item count techniques** Liu, Y., Tian, G.-L., Wu, Q., & Tang, M.-L. (2019). Poisson–Poisson item count techniques for surveys with sensitive discrete quantitative data. *Statistical Papers*, 60, 1763-1791.
- **Item count technique with a continuous or count control variable** 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

ICT with a continuous or count control variable

First treatment group:

> ‘How many hours did you sleep in total in the last two days? Include also halves and quarters.
>
> ‘Did you cheat on your taxes last year? Assign number 1 if ’yes’, and 0 if ’no’.

Please report only the difference between your answers. From your answer to the first question subtract your answer to the second question. The difference is...

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

ICT with a continuous or count control variable

Second treatment group:

> ‘How many hours did you sleep in total in the last two days? Include also halves and quarters.
>
> ‘Did you cheat on your taxes last year? Assign number 1 if ’yes’, and 0 if ’no’.

Please report only the sum of your answers. To your answer to the first question add your answer to the second question. The sum is...

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

ICT with a continuous or count control variable

Both ceiling and floor effects are eliminated

Observable variable:

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

X – answer to the neutral control question (distanced from zero)

Z – answer to the sensitive question (with binary outcomes)

Both *X* and *Z* are latent (hidden) variables and are not directly observable

Method of Moment (MM) estimator of the unknown sensitive population proportion:

$$\hat{\pi}_{MM} = \frac{1}{2a}(\bar{Y}^2 - \bar{Y}^1)$$

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

ICT with a continuous or count control variable

ML estimation via EM algorithm for normal distribution of *X*

E step (iteration *t*+1):

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

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

M step (iteration *t*+1)

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

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

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

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

Discrepancy between theoretical models and their real-life counterparts

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

Research question: How robust are ML estimators via EM algorithm to slight departures from the idealized theoretical assumption about the distribution of the control variable?

Assumptions violation in theoretical models

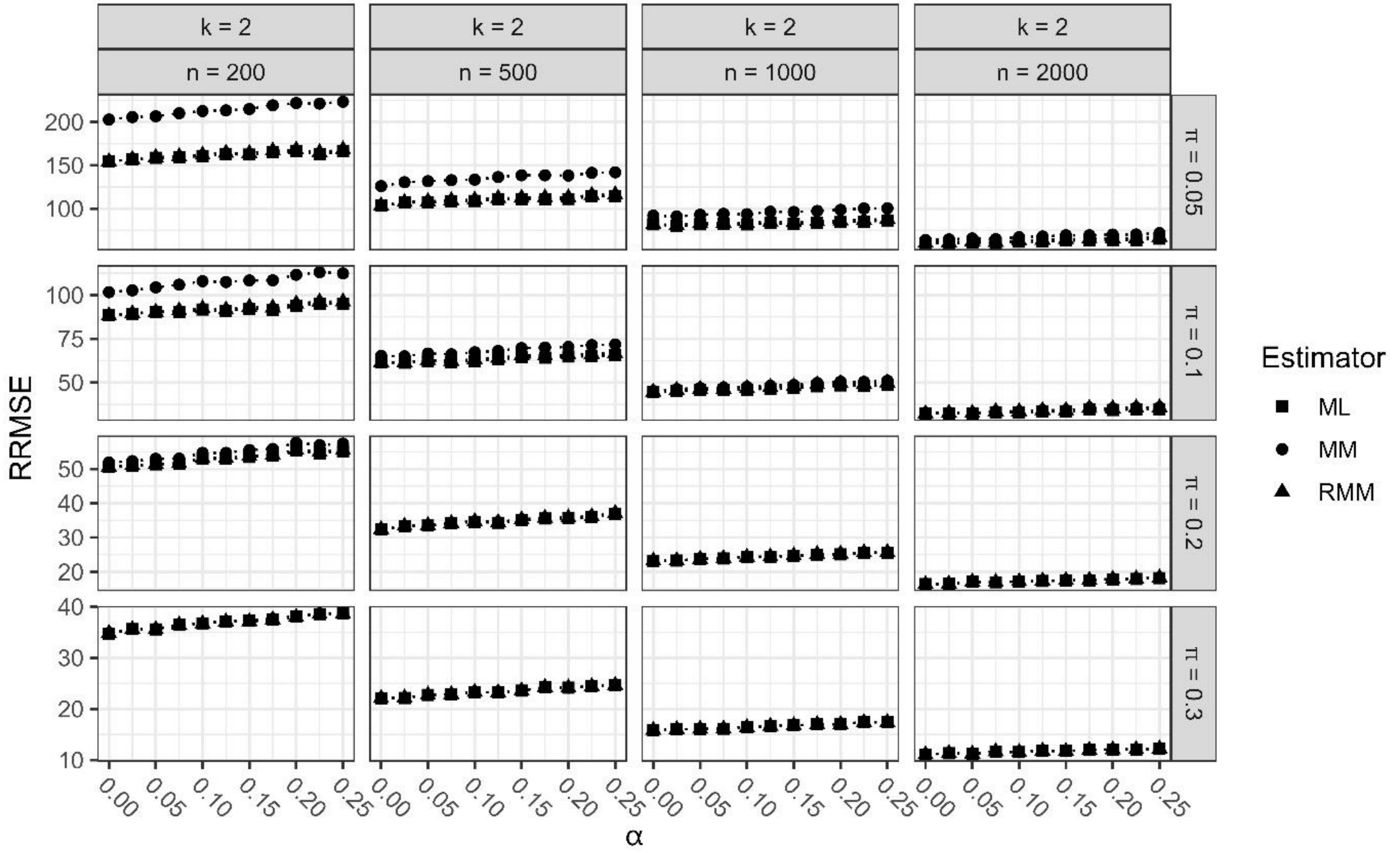
We introduce some perturbation to the distribution of the control variable

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

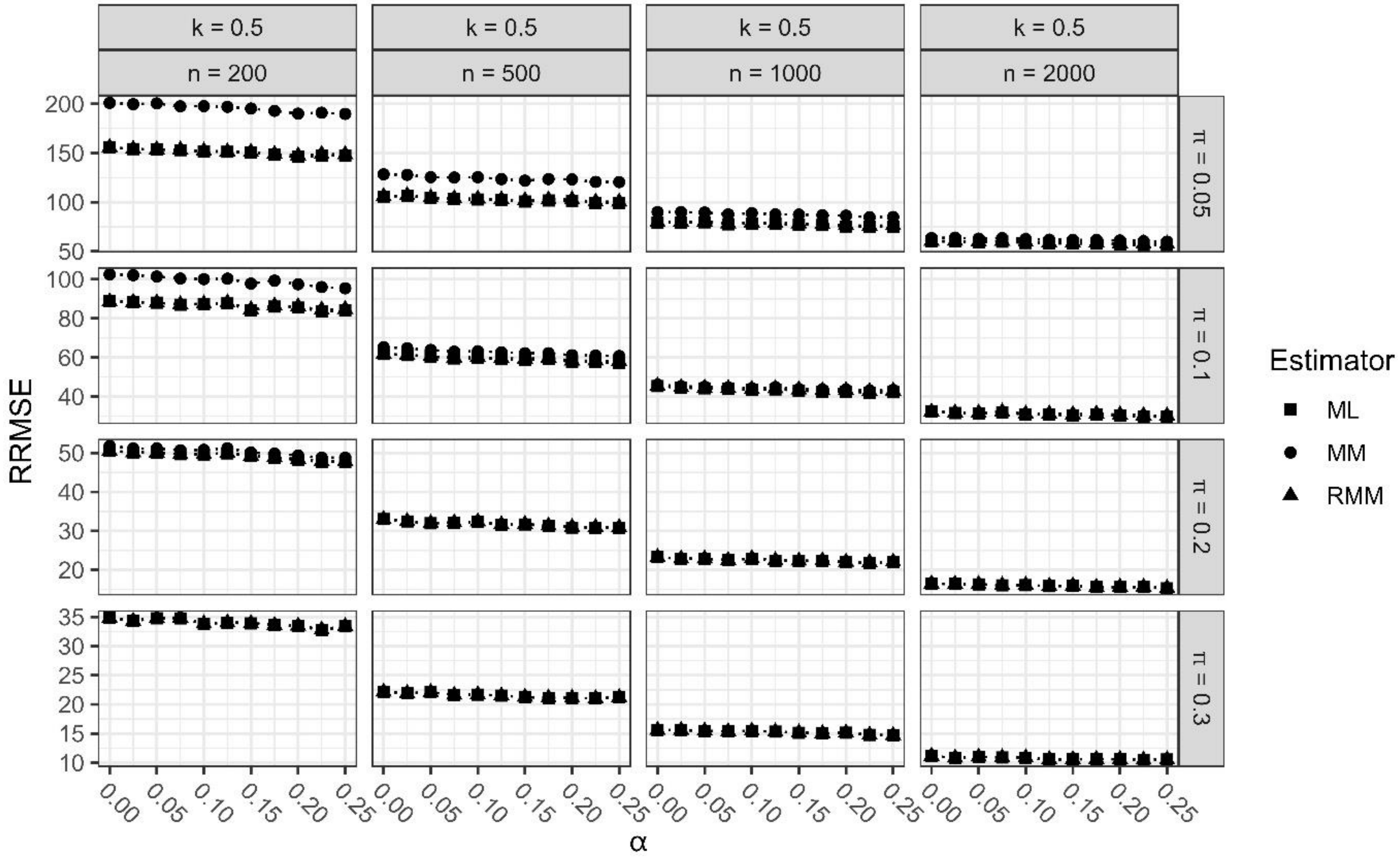
α should be small, say $\alpha < 0.25$

Monte Carlo simulation study with 10 000 replications for each set of model parameters.

Theoretical: normal, perturbation: normal with two times higher variance

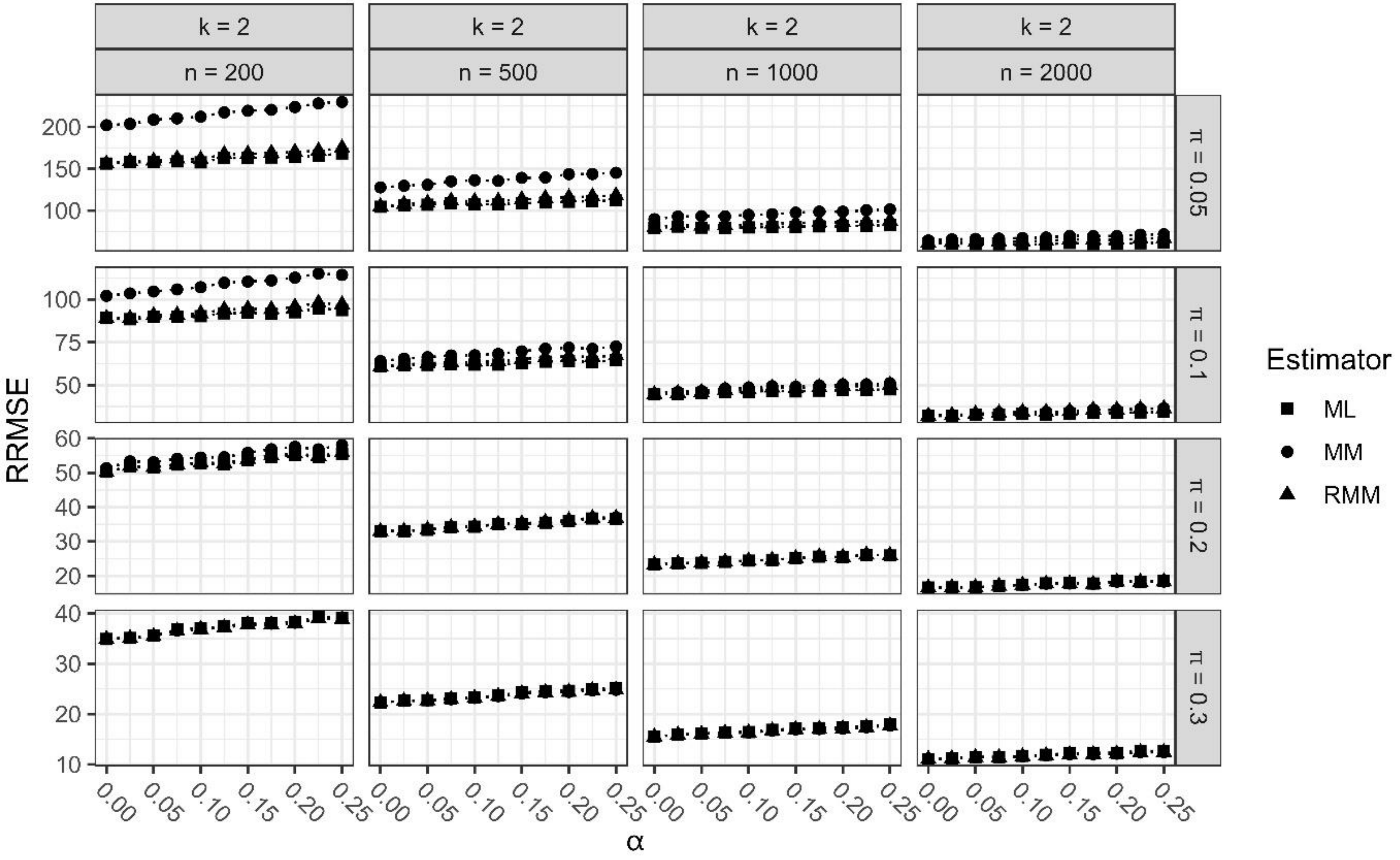


Theoretical: normal, perturbation: normal with two times smaller variance

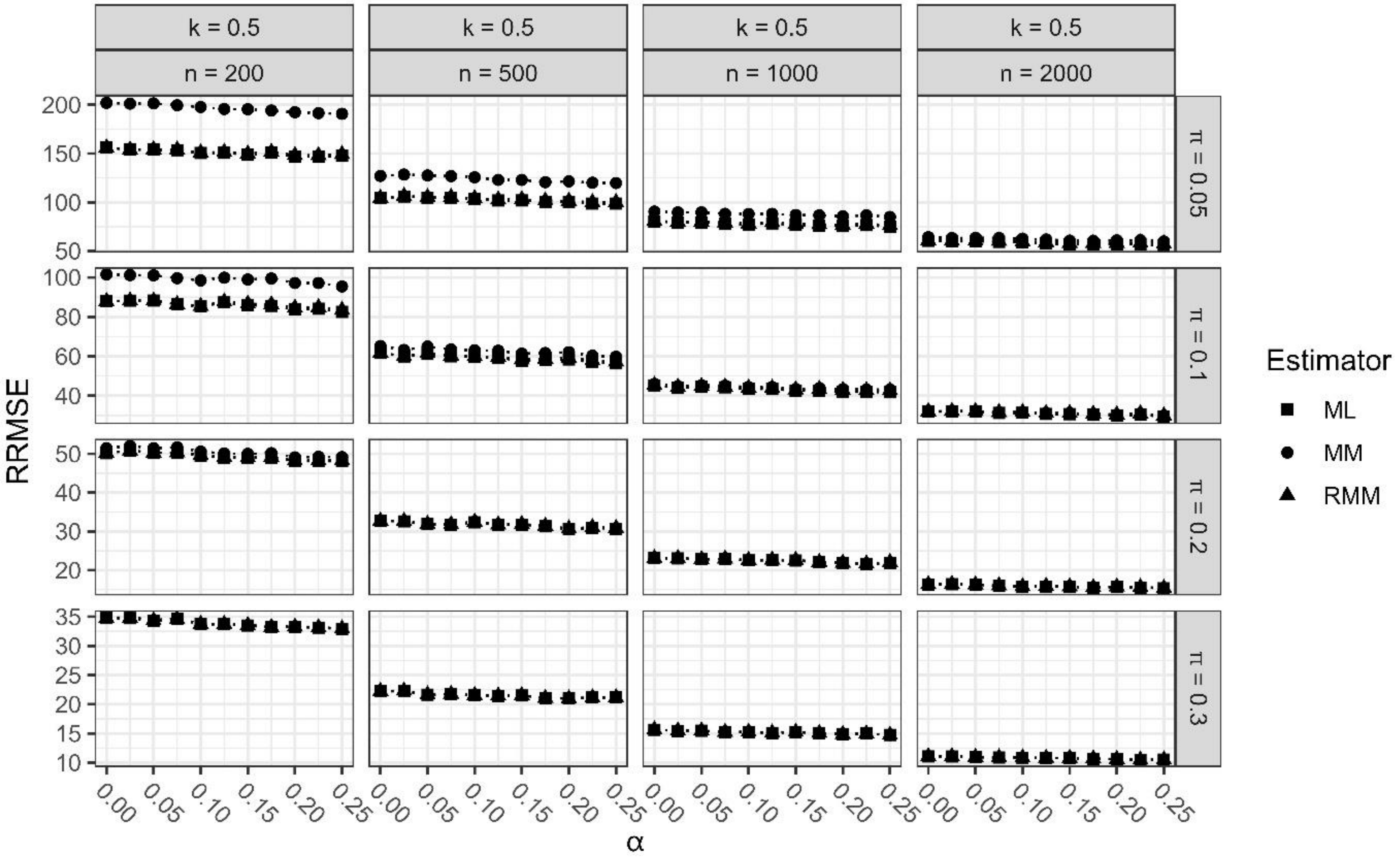


EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

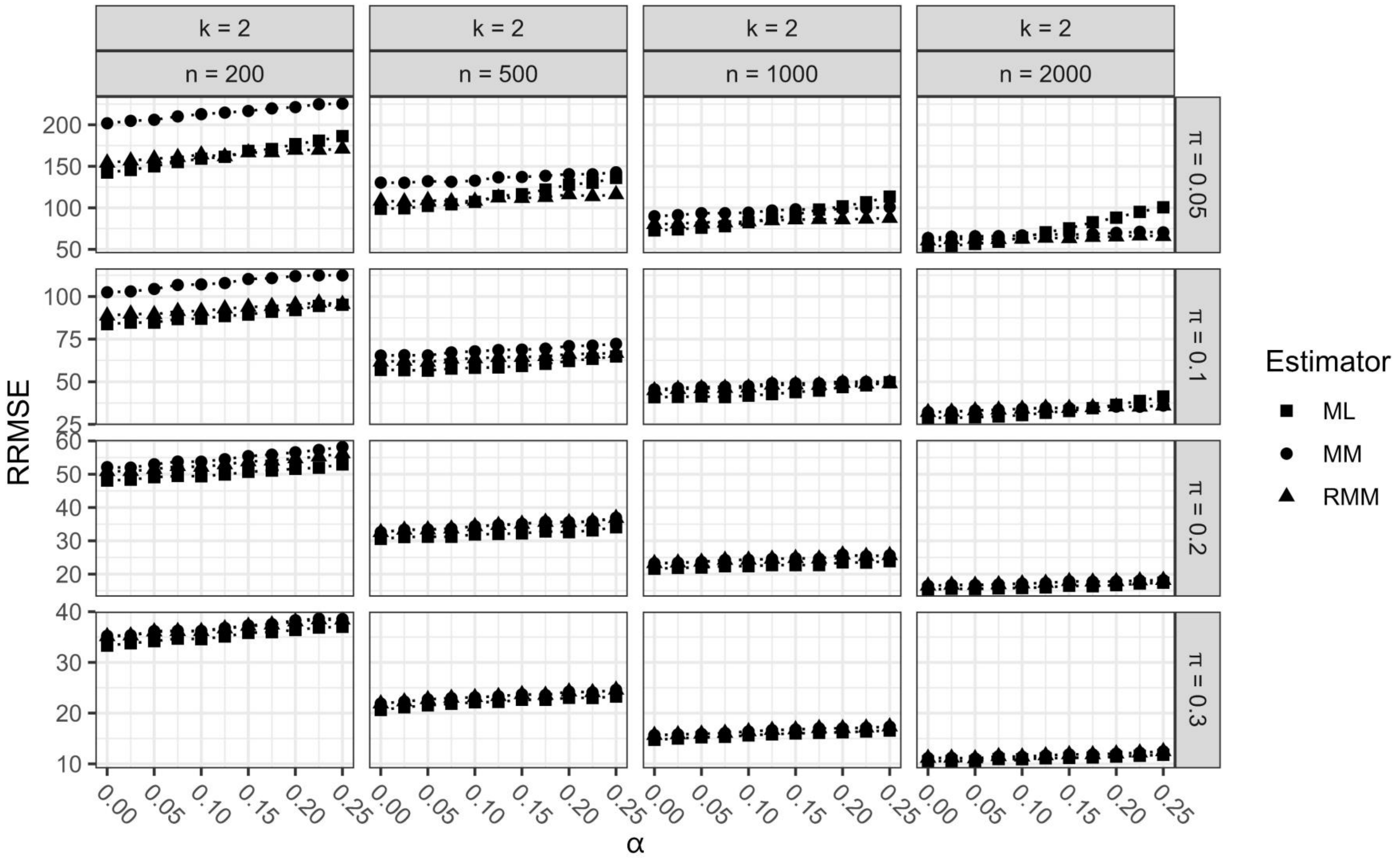
Theoretical distribution: normal, perturbation: log-normal (k=2)



Theoretical distribution: normal, perturbation: log-normal (k=0.5)

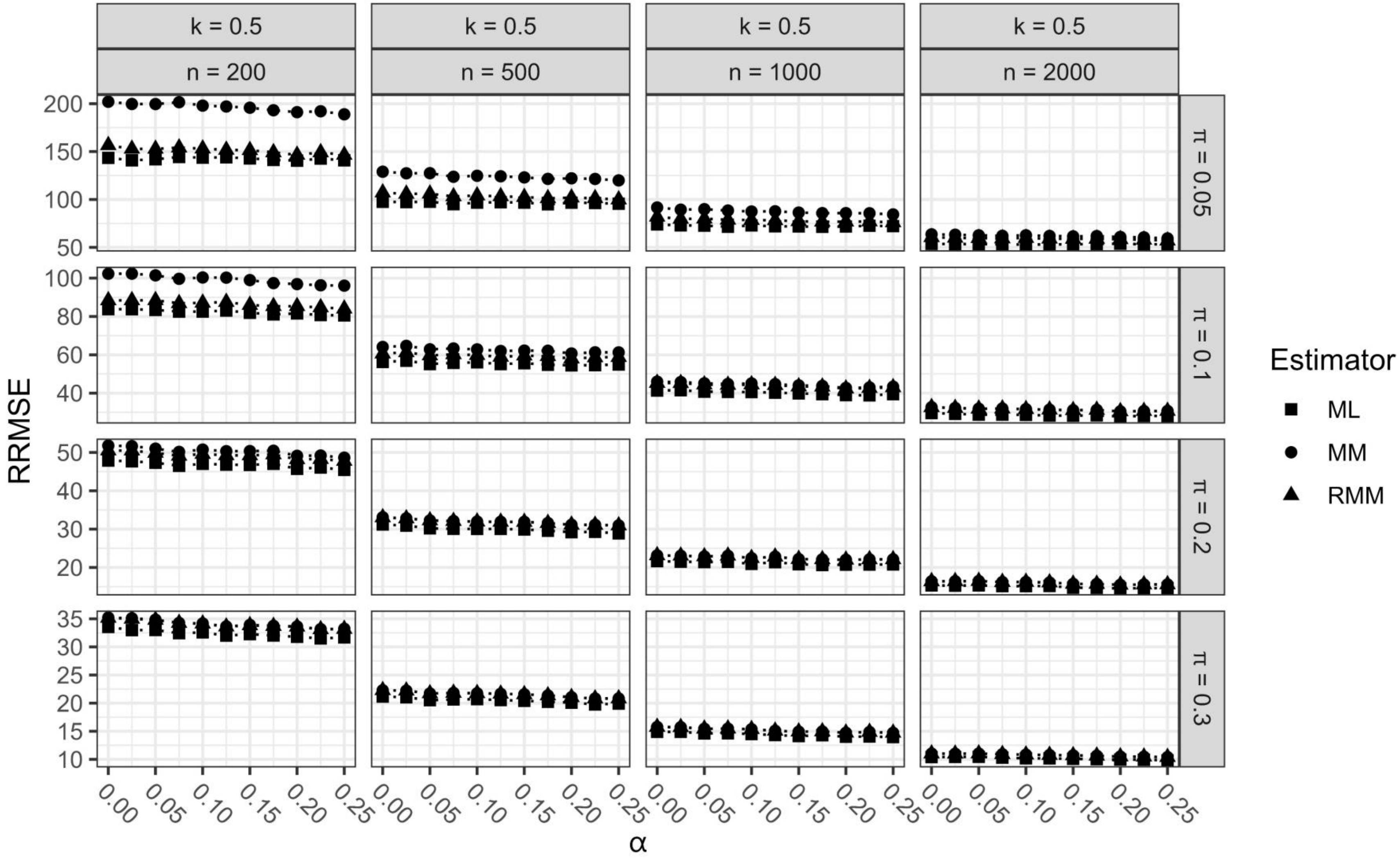


Theoretical: log-normal, perturbation: normal with two times higher variance



EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Theoretical: log-normal, perturbation: normal with two times smaller variance



Conclusions

- In all item count models one should always look for a compromise (balance) between privacy protection, efficiency of the estimation and simplicity of the questionnaire.
- Due to the need to use a control masking variable/variables larger sample sizes are needed to obtain a satisfactory level of the efficiency of the estimation.
- Estimators obtained by numerical formulas for ML via EM algorithm are quite robust to the introduction of slight violation of assumptions in the theoretical model.

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union