WORKING PAPER 6/2021 (STATISTICS)

Bayesian Model Selection for Small Datasets of Measurement Results

Olha Bodnar

ISSN 1403-0586

Örebro University School of Business
SE-701 82 Örebro, Sweden

Bayesian Model Selection for Small Datasets of Measurement Results

O. Bodnar

Unit of Statistics, School of Business, Örebro University, Fakultetsgatan 1, SE-70182 Örebro, Sweden

olha.bodnar@oru.se

Abstract

In the Cochrane Database of Systematic Reviews (CDSR) 75% of reported meta-analyses contain five or fewer studies. For a small dataset a reasonable goodness-of-fit test on a statistical model cannot be performed since either it requires a large sample size for the validity of asymptotic approximation or it might be not powerful enough to detect a deviation from the target model.

Random effects model under the assumption of normality is commonly used in many fields of science. It also appears to be a classical approach for data reduction in interlaboratory studies in metrology and in meta-analysis in medicine. However, the assumption of normality might not be fulfilled in many practical applications. If a data set is small, then no statistical test on distribution will perform well.

The intrinsic Bayes factor is used for selecting an appropriate probability model among several competitors, which not necessarily have to be nested. We apply the proposed methodology to the measurement results used to determine the Newtonian constant of gravitation and the Planck constant.

Keywords: random effects model; t-distribution; Bayesian model selection; intrinsic Bayes factor; Newtonian constant of gravitation; Planck constant.

JEL Classification: C11; C18; C02

1. Introduction

Random effects model is an established tool to perform the interlaboratory comparison study in metrology [1] and meta-analysis in medicine [2, 3]. It is also widely used to determine the values of the physical constants [4]. The assumption of normality is imposed in many applications of the random effects model without verifying its validity. While assuming normality might be appropriate for some datasets, it might deviate considerably from reality in other situations. The impact of the distributional assumption used in the random effects model was studied in two empirical illustrations in [5], which showed that the resulting values of the overall mean and of the between-study variance might be strongly influenced by the assumed distribution.

Most of meta-analyses and interlaboratory comparison studies are based on data that consist of five or fewer observations [6]. As a result, a goodness-of-fit test cannot be carried out, since it is asymptotic in nature, or it is not powerful enough to detect deviations from the distributional model specified under the null hypothesis. For this reason, we opt for Bayesian approach. The parameters of the model are endowed with the Berger and Bernardo reference prior, which is a non-informative prior. Since a non-informative prior is usually improper, the conversional Bayes model selection based on the Bayes factor cannot be used. We employ the intrinsic Bayes factor to select the most suitable model among several competing models that do not necessarily have to be nested.

The suggested approach is applied to data consisting of measurement results used in the determination of the Newtonian constant of gravitation and the Planck constant. While the assumption of normality is found to be appropriate in the case of the Newtonian constant, the data used in the calculation of the Planck constant appear to be heavy-tailed and the random effects model based on t-distribution with three degrees of freedom provides a better fit to data than the one based on the assumption of normality.

2. Bayesian model selection based on the intrinsic Bayes factor

Let $\mathbf{x} = (x_1, \dots, x_n)$ denote the measurement results and let $\mathbf{U} = \left(u_{kq}\right)_{k,q\in 1,\dots,n}$ be the covariance matrix provided together with the measurement results by participating laboratories. The generalized random effects model assumes that the density of $\mathbf{x}$ is given by (see, [7])

$$p_i(\mathbf{x}|\mu,\tau) = \frac{1}{\sqrt{\det(\mathbf{U}+\tau^2\mathbf{I})}} f_i\left((\boldsymbol{x}-\mu\mathbf{1})^T(\mathbf{U}+\tau^2\mathbf{I})^{-1}(\boldsymbol{x}-\mu\mathbf{1})\right), \tag{1}$$

where μ is the common mean and τ is the between-study standard deviation, also known as the heterogeneity parameter "dark uncertainty"; $\mathbf{1}$ denotes the vector of ones and $\mathbf{I}$ is the identity matrix. The function $f_i(.)$ determines the specific class of the random effects model. If $f_i(u) = (2\pi)^{-n/2}\exp(-u/2)$, then (1) is the normal random effects model, while the t-distributed random effects model with d degrees of freedom is obtained from (1) with

$$f_i(u) = (\pi d)^{-n/2}\frac{\Gamma((n+d)/2)}{\Gamma(d/2)}(1+u/d)^{-(n+d)/2}. \tag{2}$$

Bayes factor is widely used for model selection in Bayesian statistics. It is defined by

$$BF_{ji}(\mathbf{x}) = \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})} \text{ with } m_i(\mathbf{x}) = \int_0^{+\infty}\int_{-\infty}^{+\infty} p_i(\mathbf{x}|\mu,\tau)\,\pi_i(\mu,\tau)d\mu d\tau, \tag{3}$$

where $\pi_i(\mu,\tau)$ stands for a prior assigned to the parameters of the model M_i. If $BF_{ji}(\mathbf{x}) > 1$, then one concludes that the model M_j is preferable to M_i, otherwise one prefers the model M_i to M_j.

If $\pi_i(\mu,\tau)$ is improper as in the case of the Berger-Bernardo reference prior whose expressions derived for the normal random effects model and for the t-distributed random effects model are given in [7], then the Bayes factor in (3) cannot be computed since the marginal distribution of data is improper as well. As a solution to the problem, the intrinsic Bayes factor (IBF) is defined in [8]. The idea behind the approach is to use a part of observations, the so-called training sample to transform the improper prior to the proper posterior, which is then used in the computation of the IBF. The recommendation is to use the smallest possible number of observations as a training sample, in order to have more observations to draw a decision about the preferable model. In the case of the random effects model, the size of the minimal training sample is two independently of $f_i(.)$ following [4].

Let $\mathbf{x}_l$ denote the minimal training sample and let $\mathbf{x}_{(l)} = \mathbf{x} - \mathbf{x}_l$ denote the rest of the sample when the elements $\mathbf{x}_l$ are excluded. Then the IBF for model M_j to M_i is defined by

$$IBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) = \frac{m_j\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right)}{m_i\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right)} \text{ with } m_i\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) = \int_0^{+\infty}\int_{-\infty}^{+\infty} p_i\left(\mathbf{x}_{(l)}|\mu,\tau,\mathbf{x}_l\right)\pi_i(\mu,\tau|\mathbf{x}_l)d\mu d\tau, \tag{3}$$

where $\pi_i(\mu,\tau|\mathbf{x}_l)$ is the posterior for the parameters of M_i given the observations in the minimal training sample $\mathbf{x}_l$.

The training sample $\mathbf{x}_l$ is not uniquely chosen. When the minimal training sample consists of two elements as in the case of the random effects model (1), then one has $L = n(n-1)/2$ possible choices of two elements out of n measurement results. In such a situation one considers all possible sets of two measurement results as a training sample, while the rest of data is used for the model selection. As a result, one obtains $n(n-1)/2$ IBF values which are aggregated into a single value. Following [9] the following three aggregation approaches are used:

1) Average logarithmic IBF: $aIBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) = \frac{1}{L}\sum_l \log\left(IBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right)\right)$,

2) Median logarithmic IBF: $mIBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) = median\left(\log\left(IBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right)\right)\right)$,

3) Empirical probability IBF: $epIBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) = \frac{1}{L}\sum_l 1_{(0,+\infty)}\left(\log\left(IBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right)\right)\right)$,

where $1_{(0,+\infty)}(.)$ denotes the indicator function of set $(0,+\infty)$. If $aIBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) > 0$, then the model M_j is preferable to M_i. Similarly, the inequality $aIBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) > 0$ indicates that the model M_j should be selected, while $epIBF_{ji}\left(\mathbf{x}_{(l)}|\mathbf{x}_l\right) > 0.5$ means that the model M_j is better.

Using the IBF and three aggregation methods we compare the ability of the random effects model (1) based on the assumption of the t_3-distribution, t_5-distribution, t_{10}-distribution, and normal distribution to fit the data used in the determination of the Newtonian constant of gravitation (Section 3) and the Planck constant (Section 4).

3. Model specification for measurement results in the case of the Newtonian constant

In this section we apply the Bayes model selection approach based on the IBF to the measurement results used in the computation of the Newtonian constant of gravitation (see, [9]).

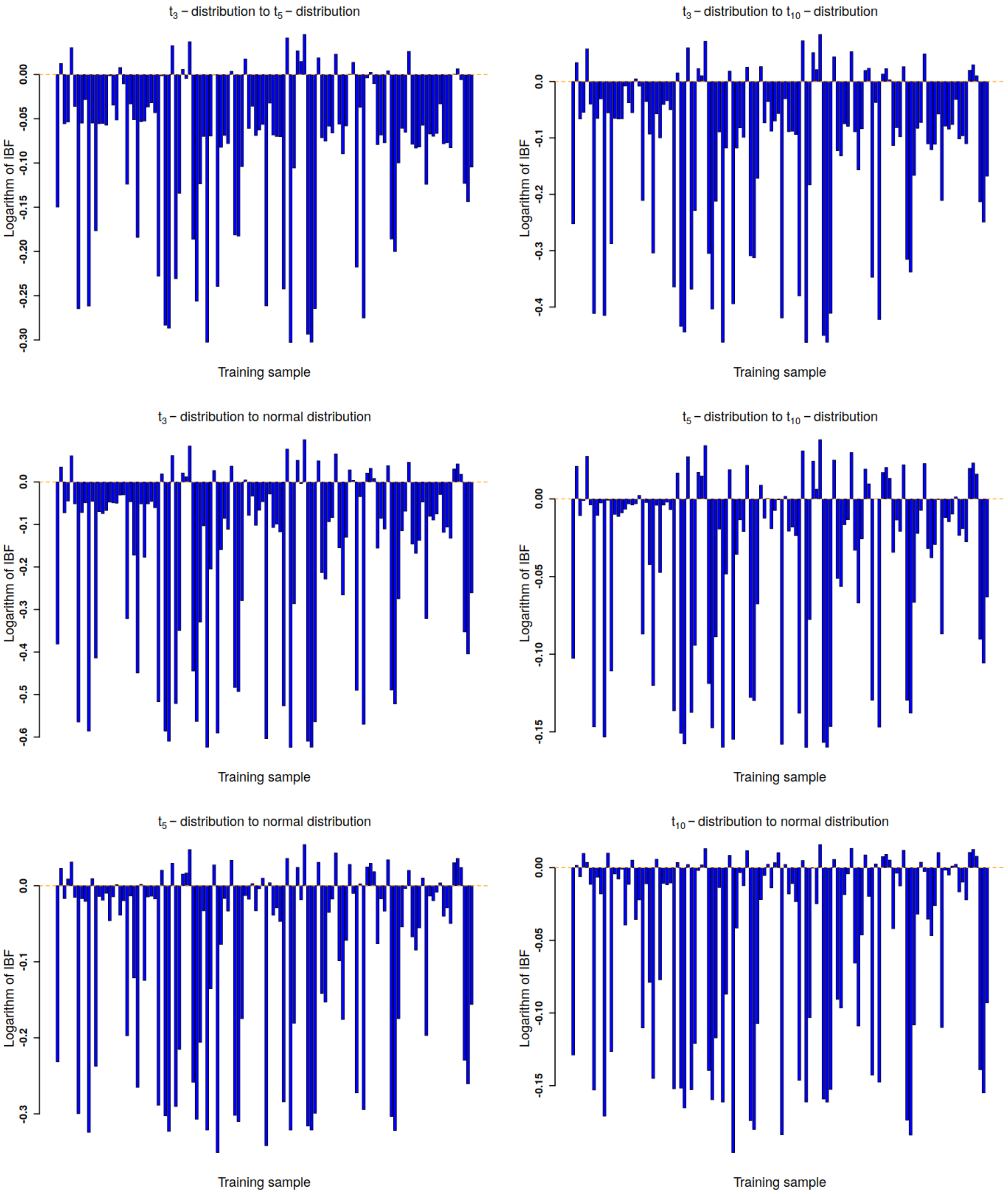

Fig.1. Logarithm of intrinsic Bayes factors for the comparison between the random effects model based on t_3-distribution, t_5-distribution, t_{10}-distribution, and normal distribution. Data: Measurement results used in the computation of the Newtonian constant for gravitation (see,[9]).

Fig. 1 depicts the values of the logarithmic IBF computed for all possible subsets consisting of measurement results used in the computation of the Newtonian constant of gravitation. The plots show that the random effects model based on the normal distribution provides a better fit to the data than the one based on the assumption of a t –distribution. Also, a t –distribution with a large number of degrees of freedom are preferable to the one with small degrees of freedom.

In Table 1 the aggregated values of the logarithms of the IBF are presented for the pairwise model comparisons between the considered t –distributions and the normal distribution. The results in the table are in line with the findings of Fig. 1 and they indicate that the normal random effects model should be chosen.

Table 1: Average logarithmic IBF, median logarithmic IBF, and empirical probability logarithmic IBF computed for the measurement results used in determination of the Newtonian constant for gravitation.

Models	t_3 to t_5	t_3 to t_{10}	t_3 to normal	t_5 to t_{10}	t_5 to normal	t_{10} to normal
a$IBF_{ji}\left(\mathbf{x}_{(l)}\vert\mathbf{x}_l\right)$	-0.0849	-0.1248	-0.1769	-0.0399	-0.092	-0.0521
m$IBF_{ji}\left(\mathbf{x}_{(l)}\vert\mathbf{x}_l\right)$	-0.0656	-0.0826	-0.0915	-0.0156	-0.031	-0.0174
$epIBF_{ji}\left(\mathbf{x}_{(l)}\vert\mathbf{x}_l\right)$	0.175	0.225	0.2083	0.25	0.275	0.275

4. Model specification for measurement results in the case of the Planck constant

The aggregated values of the logarithmic IBF are provided in Table 2. In the case of the comparison of any t –distributed random effects model to the normal one, the computed values are considerably larger than one. These finding clearly indicate the presence of heavy tails in the measurement data that cannot be captured by the normal distribution. Moreover, we conclude that the random effects model based on the t_3-distribution provides the best fit to the data used in the computation of the Planck constant.

Table 2: Average logarithmic IBF, median logarithmic IBF, and empirical probability logarithmic IBF computed for the measurement results used in determination of the Newtonian constant for gravitation.

Models	t_3 to t_5	t_3 to t_{10}	t_3 to normal	t_5 to t_{10}	t_5 to normal	t_{10} to normal
$aIBF_{ji}\left(\mathbf{x}_{(l)}\vert\mathbf{x}_l\right)$	0.4269	0.8796	1.5418	0.4527	1.1148	0.6622
$mIBF_{ji}\left(\mathbf{x}_{(l)}\vert\mathbf{x}_l\right)$	0.4169	0.8591	1.5017	0.4397	1.0892	0.656
$epIBF_{ji}\left(\mathbf{x}_{(l)}\vert\mathbf{x}_l\right)$	1	1	1	1	1	1

Fig.2 presents the values of the logarithmic IBF computed for the data used in the determination of the Planck constant (see, e.g. [10]). All values in the plots are considerably larger than zero showing that the assumption of normal distribution is not recommendable. Furthermore, we observe that the random effects model based on the t –distribution with three degrees of freedom should be selected.

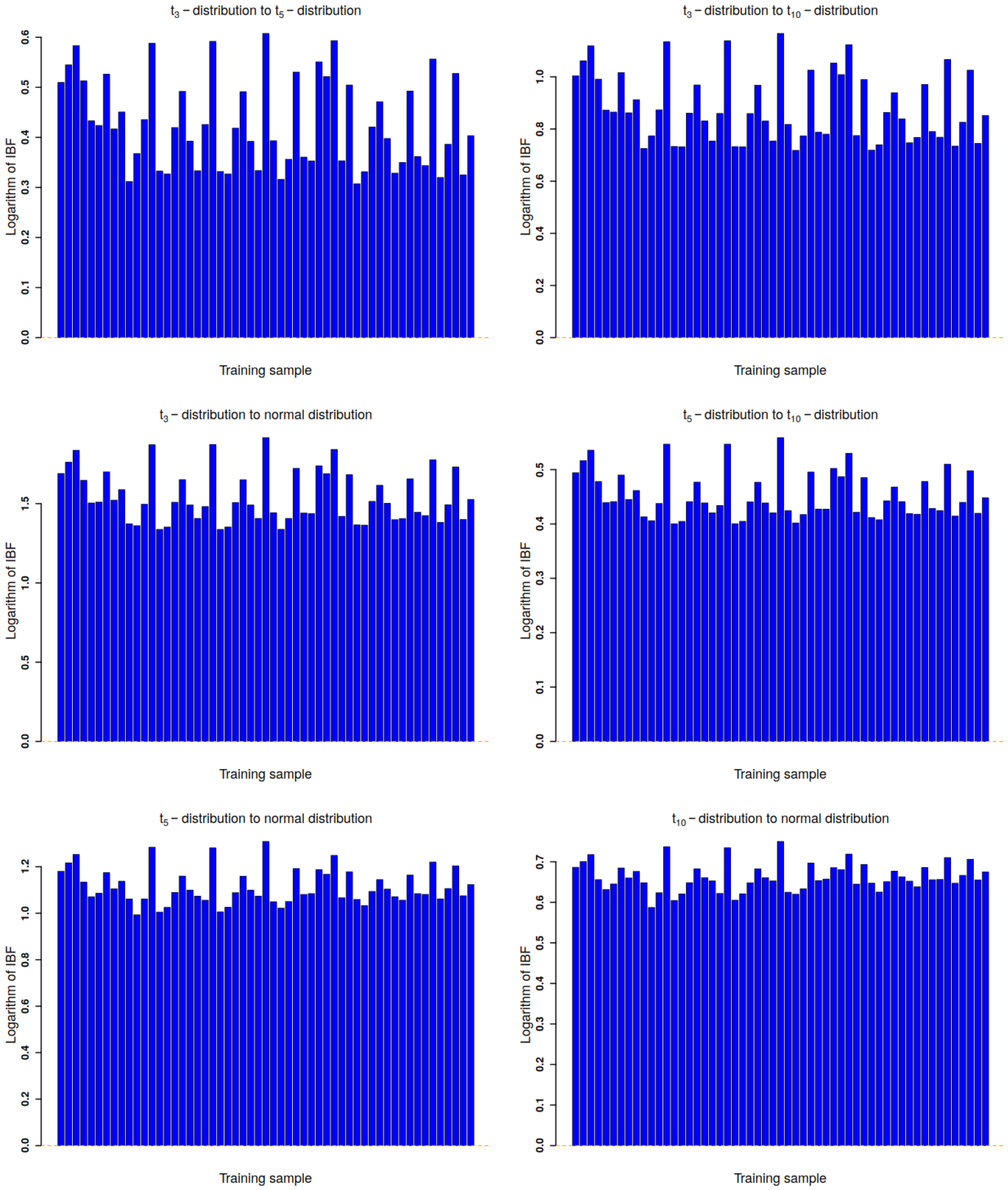

Fig. 2. Logarithm of intrinsic Bayes factors for the comparison between the random effects model based on t_3-distribution, t_5-distribution, t_{10}-distribution, and normal distribution. Data: Measurement results used in the computation of the Planck constant (see, e.g., [10]).

4. Conclusion

The model choice is a very challenging task when the sample consists only of several values. It is remarkable that most of the interlaboratory comparison studies are performed by using a few measurement results. A similar situation is also present in medicine when a meta-analysis is carried out as well as in the case of the determination of physical constants, like the Newtonian constant of gravitation and the Planck constant.

In the paper we apply the Bayesian model selection approach based on the intrinsic Bayes factor to compare the ability of the normal distribution and the t-distribution to fit measurement data. While we find that the measurement data used in the computation of the Newtonian constant of gravitation can be modeled by the normal random effects model, it is not longer a case with the data used in the determination of the Planck constant, the random effects model based on the t –distribution with three degrees of freedom should be used instead.

Баєсівський метод вибору моделі для малої кількості результатів вимірювань

О. Боднар
Unit of Statistics, School of Business, Örebro University, Fakultetsgatan 1, SE-70182 Örebro, Sweden
olha.bodnar@oru.se

Анотація

У Кокранівській базі даних систематичних оглядів (CDSR) 75% наданих мета-аналізів містять п’ять або менше досліджень. Для невеликого набору даних неможливо виконати прийнятний тест на придатність статистичної моделі, оскільки або він вимагає великого обсягу вибірки для обґрунтованості асимптотичного наближення, або він може бути недостатньо потужним для виявлення відхилення від цільової моделі.

Модель випадкових ефектів за припущення розподілу Гауса зазвичай використовується в багатьох галузях науки. Ця модель являється також найбільш поширеною для аналізу даних у міжлабораторних звіреннях у метрології та для мета-аналізу в медицині. Однак припущення нормального розподілу може не виконуватися у багатьох практичних застосуваннях. Якщо набір даних невеликий, жоден статистичний тест на розподіл не буде добре працювати.

Ми застосовуємо внутрішній коефіцієнт Баєса, запропонований у випадку, коли класичний коефіцієнт Баєса не існує, для вибору найбільш придатної ймовірнісної моделі серед кількох моделей конкурентів, які не обов'язково повинні бути вкладеними. Ми застосовуємо запропоновану методологію до результатів вимірювань, що використовуються для визначення Гравітаційної сталої та сталої Планка.

Ключові слова: Модель випадкових ефектів; t-розподіл; Баєсівський метод вибору моделей; внутрішній коефіцієнт Байєса; Гравітаційна стала; стала Планка.

References

1. Koepke A., Lafarge T., Possolo A., Toman B. Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia*, 2017, vol. 54, pp. S34–S62.
2. Bodnar O., Link A., Arendacká B., Possolo A., Elster C. Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine*, 2017, vol. 36, pp. 378–399.
3. Röver C. Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software*, 2020, vol. 93 (6).
4. Bodnar O., Link A., Elster C. Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Analysis*, 2016, vol. 11. pp. 25–45.
5. Bodnar O., Muhumuza R.N., Possolo A. Bayesian inference for heterogeneity in meta-analysis. *Metrologia*, 2020, vol. 57(6): 064004.
6. Davey J., Turner R.M., Clarke M.J., Higgins J. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, 2011, vol. 11(1):160.
7. Bodnar O. Non-informative Bayesian inference for heterogeneity in a generalized marginal random effects meta-analysis. *Theory of Probability and Mathematical Statistics*, 2019, vol. 100, pp. 7-23.
8. Berger J.O. and Pericchi L.R. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 1996, vol. 91., pp.109–122.
9. Bodnar O., Eriksson V. Bayesian model selection: Application to adjustment of fundamental physical constants. arXiv preprint arXiv:2104.01977. 2021.
10. Mohr P.J., Newell D.B., Taylor B. N. CODATA recommended values of the fundamental physical constants: 2014. *Reviews of Modern Physics*, 2016, vol. 88(3): 035009.