

Метрологическая аттестация алгоритма оценивания несогласованных данных ключевых сличений национальных эталонов

Наталья Александровна Бурмистрова^{1✉}, Иван Викторович Виктор², Анна Гурьевна Чуновкина³

^{1, 2, 3} Всероссийский научно-исследовательский институт метрологии им. Д. И. Менделеева, Санкт-Петербург, Россия

¹N.A.Burmistrova@vniim.ru✉

²I.V.Viktorov@vniim.ru, <https://orcid.org/0000-0002-9702-1224>

³A.G.Chunovkina@vniim.ru, <https://orcid.org/0000-0002-6222-5884>

Аннотация. Рассмотрен важный этап подготовки отчёта по ключевым сличениям национальных эталонов – обработка данных с использованием различных алгоритмов. Исследованы свойства оценок алгоритма обработки несогласованных данных сличений, основанного на модели случайных эффектов (random effects models, REM). Данная модель использована для описания так называемой dark uncertainty (невывявленной неопределённости). При оценке свойств алгоритма применена методология метрологической аттестации алгоритмов, в основе которой лежит исследование алгоритма на типовых моделях данных, отвечающих практическим ситуациям. Методом статистического моделирования получены следующие характеристики оценок dark uncertainty: условная плотность распределения оценки, среднее квадратическое отклонение и смещение. По результатам исследований сформулированы практические рекомендации по применению алгоритма REM.

Ключевые слова: несогласованные данные, метрологическая совместимость, неопределённость измерения, алгоритм REM, dark uncertainty

Для цитирования: Бурмистрова Н. А., Виктор И. В., Чуновкина А. Г. Метрологическая аттестация алгоритма оценивания несогласованных данных ключевых сличений национальных эталонов // Измерительная техника. 2022. № 7. С. 43–48. <https://doi.org/10.32446/0368-1025it.2022-7-43-48>

Metrological certification of the algorithm for estimating inconsistent data of key comparisons of national standards

Natalia A. Burmistrova^{1✉}, Ivan V. Viktorov², Anna G. Chunovkina³

^{1, 2, 3} D. I. Mendeleyev Institute for Metrology (VNIIM), St. Petersburg, Russia

¹N.A.Burmistrova@vniim.ru✉

²I.V.Viktorov@vniim.ru, <https://orcid.org/0000-0002-9702-1224>

³A.G.Chunovkina@vniim.ru, <https://orcid.org/0000-0002-6222-5884>

Abstract. Key comparisons of national measurement standard are used for establishing equivalence of measurement standards and confirmation of calibration capabilities declared by National metrology institutes. Key comparison data evaluation is an important part of comparison report. The Random Effects Model (REM) algorithm using for evaluation of inconsistent data obtained in interlaboratory comparison is investigated. Methodology of metrological attestation/validation of algorithms is applied, which implies testing the algorithms on typical data models corresponding to some practical cases. Estimates of the following performance characteristics of REM algorithms are obtained by statistical simulation: conditional probability function of the estimator of dark uncertainty as well as its bias and standard deviation. Some practical application of the algorithm, based on results obtained is discussed.

Keywords: inconsistent data, metrological compatibility, measurement uncertainty, REM algorithm, dark uncertainty

Введение. Целью ключевых сличений национальных эталонов является установление степеней эквивалентности эталонов и подтверждение заявленных участниками сличений неопределённостей измерений, что является необходимым условием признания калибровочных и измерительных возможностей (СМС) национальных метрологических институтов. Степень эквивалентности эталона

определяется как отклонение результата участника от опорного значения ключевого сличения и неопределённость этого отклонения, поэтому алгоритм определения опорного значения оказывает существенное влияние как на установление эквивалентности эталонов, так и на подтверждение заявленных неопределённостей. При обсуждении результатов сличений и при разработке алгоритмов

оценивания данных акцент делается именно на оценивание опорного значения.

Существует большое число алгоритмов обработки несогласованных данных сличений эталонов, при этом отсутствуют строгие критерии выбора того или иного алгоритма. В настоящей статье исследуется алгоритм обработки, основанный на применении модели случайных эффектов (REM). Случайные эффекты описывают некоторый дополнительный источник неопределённости (dark uncertainty), не учтённый участниками сличений и оцениваемый на основе представленных результатов измерений. В целом добавление dark uncertainty позволяет добиться согласованности данных сличений, но ставит вопросы обоснованности расширения неопределённостей всем участникам и подтверждения неопределённостей измерений.

Цель статьи – исследовать характеристики оценок dark uncertainty с применением методологии метрологической аттестации алгоритмов обработки данных и на основе полученных оценок дать рекомендации по применению алгоритма REM.

Оценивание данных ключевых сличений. В основе оценивания данных сличений лежит базовая модель $X_i = X$, где X – измеряемая величина; X_i – измеряемая величина в i -й лаборатории, $i = 1, \dots, n$; n – количество лабораторий, участвующих в сличениях. При этом предполагается, что все участники измеряют одну и ту же величину и представляют измеренное значение и соответствующую (стандартную) неопределённость (x_i, u_i). Средневзвешенное значение x_w результатов участников сличений с весами, обратно пропорциональными квадратам соответствующих стандартных неопределённостей, и неопределённость $u(x_w)$ этого значения вычисляются по формулам

$$\left. \begin{aligned} x_w &= \sum [x_i u^{-2}(x_i)] / \sum u^{-2}(x_i); \\ u(x_w) &= \sqrt{\sum u^{-2}(x_i)}, \end{aligned} \right\} \quad (1)$$

где x_i – результаты участников сличений; $u(x_i)$ – стандартные неопределённости результатов.

Оценку (1) можно использовать в качестве опорного значения, если результаты всех участников согласуются с принятой моделью, т. е. принадлежат одному и тому же распределению.

В математической статистике разработано большое количество критериев совместимости/согласованности данных. При ключевых сличениях традиционно используют критерий χ^2 . Вычисляется наблюдаемое значение статистики χ^2 :

$$\chi_n^2 = \sum_{i=1}^n (x_i - x_w)^2 u^{-2}(x_i).$$

Если вероятность $P\{\chi^2(n-1) > \chi_{n-1}^2\} < 0,05$, то согласованность данных не подтверждается. Несогласованность результатов сличений означает, что результаты участников отличаются друг от друга существенно, чем это может быть объяснено заявленными неопределённостями измерений. Причины несогласованности результатов измерений могут

быть разными: нестабильность транспортируемого эталона, занижение неопределённости измерений отдельными участниками, существенное отклонение одного или нескольких результатов, обусловленное невыявленной систематической погрешностью измерения, и др. Часто простой графический анализ представленных данных позволяет выявить так называемые «выбросы», т. е. данные, которые значительно отличаются от большинства. Хотя существуют статистические критерии выявления выбросов, но с учётом важности и ответственности анализа данных ключевых сличений при принятии решений об исключении того или иного результата требуется установить возможные причины появления выброса.

Если не удаётся добиться согласованности данных путём предметного анализа бюджета неопределённостей участников, свойств транспортируемого эталона и протокола измерений, а также другой доступной информации, то необходимо применять «искусственные» алгоритмы обработки таких (несогласованных) данных. Все указанные алгоритмы основаны либо на расширении заявленных неопределённостей участников сличений, либо на исключении результатов отдельных участников. Но при расширении неопределённостей участников сличений возникает много вопросов: насколько можно считать сличения удачными, если они показали, что исходные неопределённости участников не подтверждаются, согласны ли все участники с расширением неопределённостей, есть ли предел в допустимом расширении неопределённостей участников, какие неопределённости указывать в калибровочных и измерительных возможностях и др. При использовании алгоритмов, связанных с расширением неопределённостей измерений участников сличений, ответы на эти вопросы должны содержаться в отчёте о сличениях.

Модель случайных эффектов. За более чем два десятилетия, прошедших с момента подписания CIPM MRA [1], для несогласованных данных предложены разнообразные алгоритмы оценивания опорного значения сличений [2–10], разработано разветвлённое программное обеспечение [11]. Однако анализ алгоритмов, встречающихся в отчётах о результатах ключевых сличений, показывает, что наиболее часто используются простые, «прозрачные» алгоритмы обработки данных, например, алгоритм наибольшего согласованного подмножества (largest consistent subset, LCS) [2].

Алгоритм REM пока не получил широкого практического применения, но активно обсуждается в публикациях и имеет много модификаций [3–7], поэтому представляется важным рассмотреть возможности и ограничения его использования на практике. Этот алгоритм основан на статистической модели (модели случайных эффектов) $X_i = X + B$, где B – случайный эффект (dark uncertainty), который описывает неучтённые компоненты неопределённости измерений. Дополнительно предполагается, что B принадлежит нормальному распределению N и имеет нулевое математическое ожидание: $B \in N(0, \sigma)$, причём дисперсия σ оценивается численными методами по экспериментальным данным, представленным участниками. Соответственно, неопределённость измерений увеличивается одинаково для всех участников: $\hat{u}_i^2 = u_i^2 + \sigma^2$.

Оценки опорного значения μ и дисперсии σ можно получить путём максимизации функции правдоподобия $L(\mu, \sigma^2)$ [3]:

$$\max_{\mu, \sigma^2} L(\mu, \sigma^2) = \max_{\mu, \sigma^2} \prod_{i=1}^n [2\pi(u_i^2 + \sigma^2)]^{-1/2} e^{-\frac{1}{2} \frac{(x_i - \mu)^2}{(u_i^2 + \sigma^2)}}.$$

Оценка REM имеет вид $x_{\text{REM}} = \frac{\sum_{i=1}^n x_i (u_i^2 + \hat{\sigma}^2)^{-1}}{\sum_{i=1}^n (u_i^2 + \hat{\sigma}^2)^{-1}}$, где $\hat{\sigma}$ – неотрицательная оценка параметра, определяемая численно из уравнения

$$\sum_{j=1}^n \left[x_j - \frac{\sum_{i=1}^n x_i (u_i^2 + \hat{\sigma}^2)^{-1}}{\sum_{i=1}^n (u_i^2 + \hat{\sigma}^2)^{-1}} \right]^2 (u_j^2 + \hat{\sigma}^2)^{-2} = \sum_{j=1}^n (u_j^2 + \hat{\sigma}^2)^{-1}. \quad (2)$$

При отсутствии вещественных решений (2) $\hat{\sigma} = 0$.

Свойства алгоритмов и критериев в настоящей работе исследованы методами статистического моделирования с применением методологии метрологической аттестации алгоритмов обработки данных [12]: на вход алгоритма подаются смоделированные статистические данные с известными параметрами моделирования; данные обрабатываются в соответствии с исследуемым алгоритмом; результат применения алгоритма сопоставляется с заданным параметром моделирования, что позволяет оценить определённые характеристики алгоритма. Оценивали следующие характеристики алгоритма REM:

условную функцию распределения оценки dark uncertainty при заданном значении σ : $P(\hat{\sigma} < x | \sigma)$;

относительное смещение оценки dark uncertainty $[E(\hat{\sigma}) - \sigma] / \sigma$;

относительное стандартное отклонение оценки dark uncertainty $S(\hat{\sigma}) / \sigma$.

Рассмотрим статистическую модель для моделирования исходных данных сличений: предположим, что все результаты измерений нормально распределены с нулевым математическим ожиданием и дисперсией $u_i^2 + \sigma^2$, т. е.

$$x_i = N(0, u_i^2 + \sigma^2). \quad (3)$$

Также предположим, что участников сличений $n=10$ и они заявляют близкие неопределённости измерений $u_i \approx u$, кроме специально оговоренных случаев. Для удобства при моделировании зададим параметр σ , характеризующий dark uncertainty, пропорционально неопределённости измерений в лабораториях: $0 \leq \sigma \leq 2u$. Неопределённость u при расчётах зададим равной 1, что не является ограничением, так как оцениваемые далее характеристики алгоритмов зависят от отношения σ/u . Далее для модели (3) при фиксированном значении σ из приведённого интервала с шагом дискретизации $0,05u$ с использованием генератора случайных величин зададим 10^5 наборов результатов сличений. На основе полученных данных вычислим вероятностные характеристики алгоритма REM.

Оценка согласованности данных ключевых сличений.

Так как применение критерия χ^2 для проверки согласованности

данных является важной «развилкой» при обосновании выбора алгоритма оценивания опорного значения ключевых сличений, то оценим свойства этого критерия для различного количества участников сличений.

Определим величину, «обратную» мощности критерия χ^2 , т. е. вероятность того, что критерий подтверждает согласованность данных при условии, что в результатах участников присутствует dark uncertainty.

На рис. 1 приведена вероятность выполнения критерия χ^2 для $n=5, 10, 20$ и 30 участников сличений. С увеличением dark uncertainty вероятность выполнения критерия χ^2 логично уменьшается. Чем меньше количество участников сличения, тем с меньшей вероятностью критерий выявляет избыточную дисперсию (dark uncertainty). При этом для $\sigma < 0,2u$ вероятности выполнения критерия χ^2 для всех n мало изменяются и мало различаются. Таким образом, вероятность обнаружения dark uncertainty существенна при $\sigma > 0,2u$, а учитывая цели сличения, рассматривать значения $\sigma > u$ не имеет смысла. Полученный результат говорит о том, что, возможно, алгоритм REM следует непосредственно применять к анализу данных без предварительной проверки их согласованности.

Алгоритм REM имеет разные модификации, отличающиеся в том числе методами оценки избыточной дисперсии. Кроме приведённой выше сложной в реализации оценки, которая основана на максимуме функции правдоподобия и обеспечивает минимально смещённые оценки, можно перечислить оценки Менделя и Пауля (Mandel and Paule), Вангеля и Рухина (Vangel and Rukhin), Дер-Симонян и Лейрда (DerSimonian and Laird). Последняя из указанных оценок наиболее проста в применении и задаёт оценку избыточной дисперсии в явном виде: $\hat{\sigma}_{D-L} = \sqrt{\hat{\sigma}_{D-L}^2}$, где

$$\hat{\sigma}_{D-L}^2 = \left[\sum_{i=1}^n \frac{(x_i - \bar{x}_w)^2}{u_i^2} - (n-1) \right] \left[\sum_{i=1}^n \frac{1}{u_i^2} - \frac{\sum_{i=1}^n u_i^{-4}}{\sum_{i=1}^n u_i^{-2}} \right]^{-1}.$$

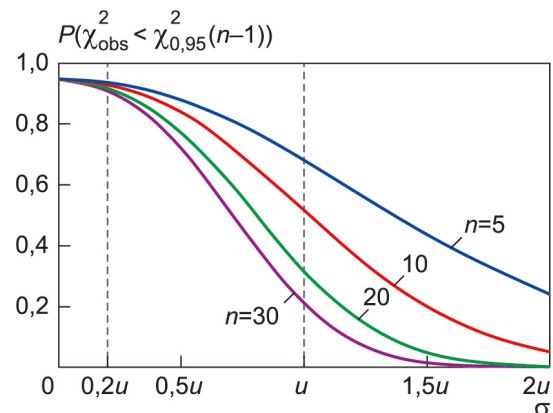


Рис. 1. Вероятность выполнения критерия χ^2 для $n=5, 10, 20$ и 30 участников сличений

Fig. 1. Probability of meeting the criterion χ^2 for $n=5, 10, 20$ and 30 participants of comparisons

На рис. 2 показаны вероятности получения нулевой оценки dark uncertainty при сличении десяти лабораторий в зависимости от значений σ (2) и оценки по Дер-Симонян и Лейрду (3). С одной стороны, даже при отсутствии dark uncertainty вероятность получения её ненулевой оценки при использовании алгоритма REM составляет более 35 % (а для оценки по Дер-Симонян и Лейрду ещё больше). С другой стороны, даже при избыточной дисперсии, равной заявленной неопределённости участников сличений ($\sigma=u$) существует довольно высокая вероятность получить

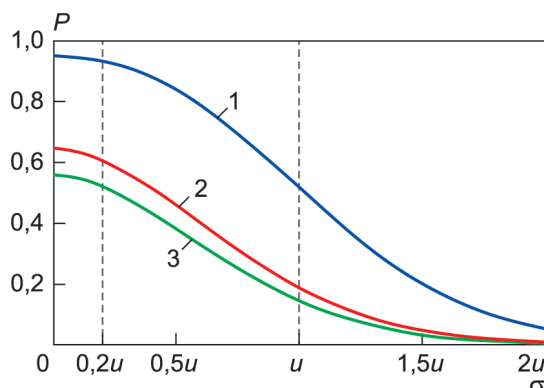


Рис. 2. Вероятности получения нулевой оценки dark uncertainty для сличения десяти лабораторий в зависимости от значений σ (2) и оценки по Дер-Симонян и Лейрду (3) и вероятность выполнения критерия χ^2 (1)

Fig. 2. Probability of equality to zero of the dark uncertainty estimate to compare ten laboratories depending on the values of σ (2) and the DerSimonyan and Laird estimates (3) and probability of meeting the criterion χ^2 (1)

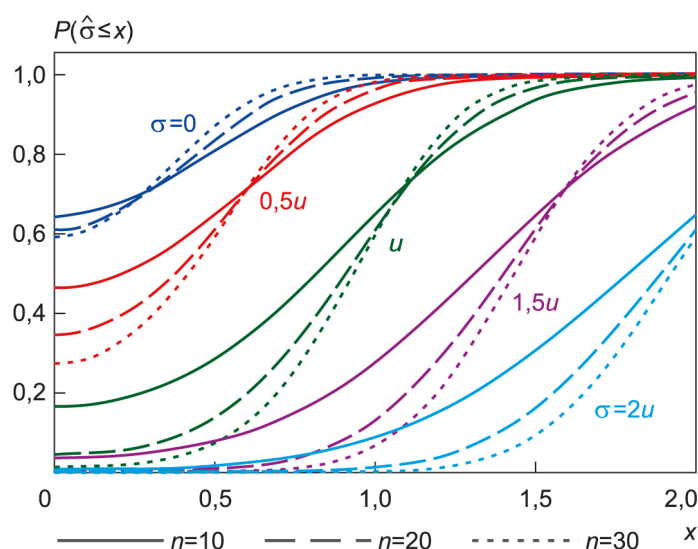


Рис. 3. Функция распределения вероятности $P(\hat{\sigma} < x|\sigma)$ для разного количества n участников сличений и различной дисперсии σ

Fig. 3. Probability distribution function $P(\hat{\sigma} < x|\sigma)$ for different n and σ

оценку $\sigma=0$. Также на рис. 2 (кривая 1) приведена вероятность выполнения критерия χ^2 .

Функция распределения вероятности оценки dark uncertainty. Применяемый в настоящей работе метод статистического моделирования позволяет построить функцию распределения вероятности для оценки избыточной дисперсии при некоторых фиксированных значениях σ и проанализировать влияние различных факторов модели. На рис. 3 показано влияние количества участников сличений для пяти возможных значений dark uncertainty, а на рис. 4 – влияние выбора оценки избыточной дисперсии для этих пяти значений (сплошная линия – оценка из максимума функции правдоподобия; пунктирная – оценка по Дер-Симонян и Лейрду). С увеличением числа участников сличений уменьшается влияние выбора оценки.

Относительное смещение оценки dark uncertainty. Методология метрологической аттестации алгоритмов позволяет определить характеристики оценок REM. На рис. 5, а приведено относительное смещение оценки избыточной дисперсии $[E(\hat{\sigma})-\sigma]/\sigma$ для разного количества участников сличений. Независимо от количества лабораторий систематическая погрешность меньше всего при значениях σ , близких к $0,2u$. При $\sigma < 0,2u$ оценка σ завышена, а при $\sigma > 0,2u$ занижена, что особенно заметно при малом числе лабораторий. Таким образом, при $\sigma < 0,2u$ существенное положительное смещение оценки σ в итоге приводит к необоснованному завышению заявленных неопределённостей участников сличений при использовании алгоритма REM.

Относительное СКО оценки dark uncertainty. На рис. 5, б представлено относительное стандартное отклонение dark

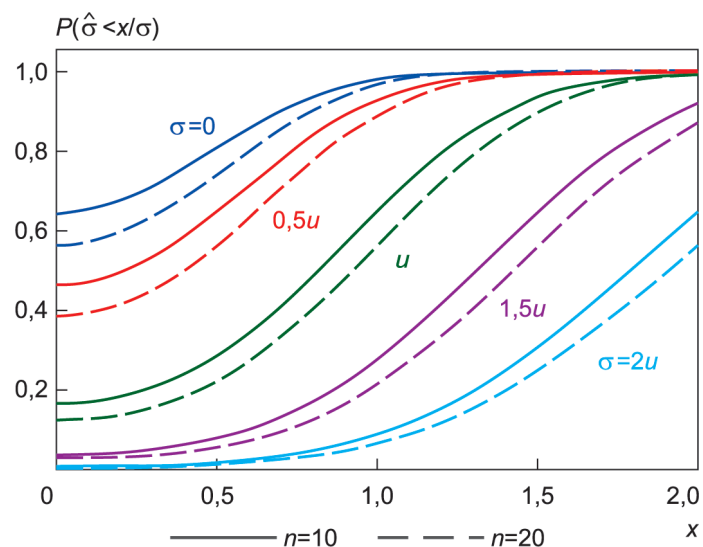


Рис. 4. Сравнение функций распределения вероятности $P(\hat{\sigma} < x|\sigma)$ для разного количества n участников сличений и различной дисперсии σ

Fig. 4. Comparison of probability distribution functions $P(\hat{\sigma} < x|\sigma)$ depending on the choice of estimates for n participants of comparisons and for different σ

uncertainty $S(\hat{\sigma})/\sigma$ для разного количества участников сличений. При $\sigma < 0,2u$ относительное стандартное отклонение значительно, а при $\sigma > 0,2u$ оно стремится к нулю.

Заключение. При метрологической аттестации алгоритма REM, в котором для обработки несогласованных данных ключевых сличений используется dark uncertainty, получены следующие результаты:

при малых значениях $\sigma < 0,2u$ вероятность обнаружения dark uncertainty с использованием критерия χ^2 мала (менее 10 %) для количества n участников сличений меньше 30. В то же время оценка σ завышена и имеет значительное стандартное отклонение, что на практике может приводить к необоснованному расширению заявленных неопределённостей участниками сличения;

при значениях $\sigma > 0,2u$ точность оценки σ увеличивается, но необходимо убедиться в практической обоснованности применения алгоритма, поскольку результатом является увеличение неопределённостей участников сличений в среднем на 20 % (и более).

Обнаружение dark uncertainty является основанием для предметного анализа результатов сличений с целью выяснения причин несогласованности данных, представленных участниками сличений. На взгляд авторов настоящей статьи, основными причинами несогласованности данных могут быть: занижение неопределённости измерений отдельными участниками; присутствие в результатах участников невыявленных (неучтённых в бюджете неопределённости) систематических погрешностей; нестабильность транспортируемого эталона. Именно незначительная нестабильность транспортируемого эталона может приводить к эффекту, который проявляется в измерениях всех участников и может быть описан с применением алгоритма REM. Для первых двух причин, которые проявляются только для отдельных результатов сличений, использование dark uncertainty может приводить к необоснованному увеличению неопределённостей всех участников.

Таким образом, важно ещё раз обратиться к целям сличений, а именно установлению эквивалентности эталонов и подтверждению заявленных неопределённостей. Эквивалентность эталонов устанавливается для конкретной группы эталонов, степень эквивалентности определена внутри этой группы. Строго говоря, если изменяется состав этой группы, то изменяются и опорное значение, и степень эквивалентности. В этом смысле применение dark uncertainty по сути является констатацией того уровня согласованности, который существует внутри конкретной группы эталонов. В то же время заявленные неопределённости участников сличений являются основой для калибровочных и измерительных возможностей, которые, в свою очередь, являются исключительной характеристикой соответствующего национального метрологического института. Поэтому вопрос подтверждения заявленных неопределённостей измерений в случае применения dark uncertainty требует обсуждения и отражения в отчёте по ключевым сличениям.

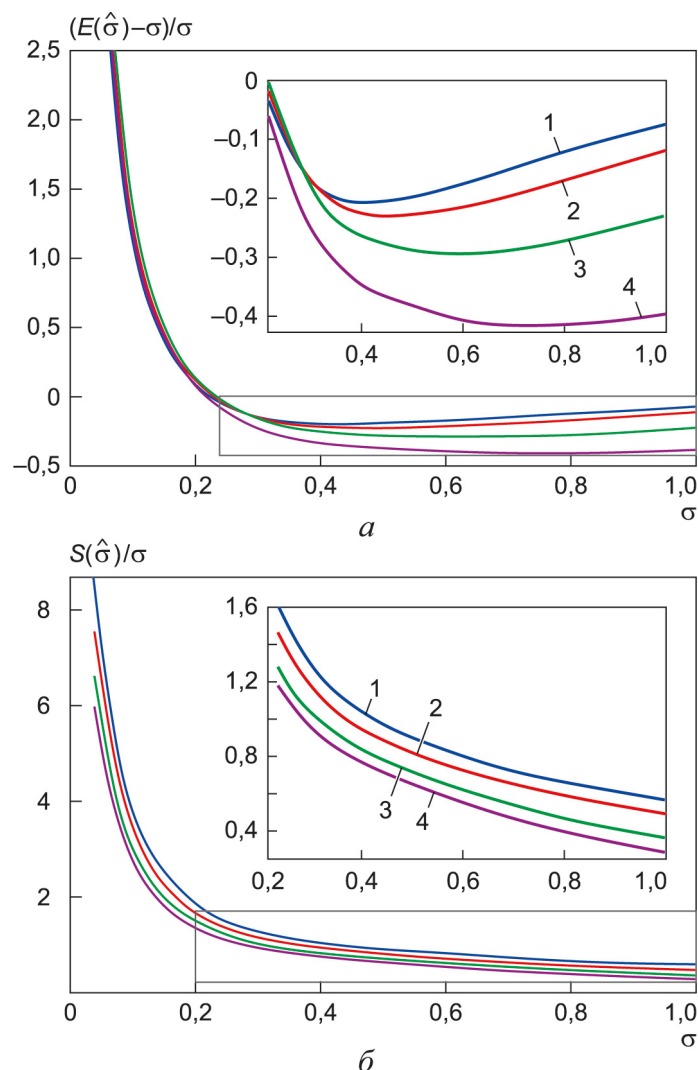


Рис. 5. Относительные смещение $[E(\hat{\sigma}) - \sigma]/\sigma$ (а) и стандартное отклонение $S(\hat{\sigma})/\sigma$ (б) оценки dark uncertainty при количестве участников сличений $n = 5, 10, 20, 30$ – кривые 1–4 соответственно

Fig. 5. Relative bias of the dark uncertainty estimate and relative standard deviation of the dark uncertainty standard deviation estimate for $n = 5, 10, 20$ and 30 participants of comparisons – curves 1–4, respectively

Список источников/References

1. CIPM MRA, Bureau International des Poids et Mesures, 1999.
2. Cox M. G., *Metrologia*, 2007, vol. 44, no. 3, pp. 187–200. <https://doi.org/10.1088/0026-1394/44/3/005>
3. Willink R., *Metrologia*, 2002, vol. 39, no. 4, pp. 343–354. <https://doi.org/10.1088/0026-1394/39/4/3>
4. Birge R. T., *Phys. Rev.*, 1932, vol. 40, pp. 207–227. <https://doi.org/10.1103/PhysRev.40.207>
5. DerSimonian R., Laird N., *Controlled Clinical Trials*, 1986, no. 7, pp. 177–188. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)
6. John Mandel and Robert C. Paule, *Anal. Chemistry*, 1970, vol. 42, no. 11, pp. 1194–1197. <https://doi.org/10.1021/ac60293a019>

7. Robert C. Paul, John Mandel, *J. Res. Natl. Bur. Stand.*, 1982, vol. 87, no. 5, pp. 377–385.

<https://doi.org/10.6028/jres.087.022>

8. Бурмистрова Н. А. Разработка и исследования алгоритмов обработки несогласованных данных ключевых сличений эталонов // Измерительная техника. 2014. № 10. С. 7–12 [Burmistrova N., *Measurement Techniques*, 2015, vol. 57, no. 10, pp. 1103–1112.

<https://doi.org/10.1007/s11018-015-0587-4>.

9. Burmistrova N., Viktorov I., Chunovkina A., *Proceedings of the International Conference IEEE 2019 XXIX International Scientific Symposium “Metrology and Metrology Assurance” (MMA)*, Sozopol, Bulgaria, September 06, 2019, pp. 14–18.

10. Rukhin A. L., Possolo A., *Computational Statistics and Data Analysis*, vol. 55, iss. 4, pp. 1815–1827.

<https://doi.org/10.1016/j.csda.2010.11.016>

11. Koepe A., Lafarge T., Toman B., Possolo A., *NIST Consensus Builder – User’s Manual*, National Institute of Standards and Technology, Gaithersburg, MD, 2020, available at: <https://consensus.nist.gov/> (accessed: 01.06.2022).

12. Тарбеев Ю. В., Челпанов И. Б., Сирая Т. Н. Аттестация алгоритмов обработки данных при измерениях // Измерения, контроль, автоматизация. 1991. №. 3. С. 3–13 [Tarbeev Ju. V., Chelpanov I. B., Siraja T. N., *Attestaciya algoritmov obrabotki dannyh pri izmereniyah, Izmerenija, kontrol’, avtomatizacija*, 1991, no. 3, pp. 3–13. (In Russ.)].

Статья поступила в редакцию 02.06.2022; одобрена после рецензирования 15.06.2022; принята к публикации 16.06.2022.
The article was submitted 02.06.2022; approved after reviewing 15.06.2022; accepted for publication 16.06.2022.