Chemometrics and Intelligent Laboratory Systems 219 (2021) 104429

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Robust bootstrapped Mandel's *h* and *k* statistics for outlier detection in interlaboratory studies

Miguel Flores [a], Génesis Moreno [b], Cristian Solórzano [b], Salvador Naya [c], Javier Tarrío-Saavedra [c,*]

[a] Grupo MODES, SIGTI, Departamento de Matemática, Facultad de Ciencias, Escuela Politécnica Nacional. Quito, Ecuador
[b] Departamento de Matemática, Facultad de Ciencias, Escuela Politécnica Nacional. Quito, Ecuador
[c] Grupo MODES, CITIC, ITMATI, Departamento de Matemáticas, Escola Politécnica Superior, Universidade da Coruña, Ferrol, Spain

ARTICLE INFO

Keywords:
ILS
Outlier detection
Bootstrap
Nonparametric statistics

ABSTRACT

This study proposes the use of the bootstrap methodology to estimate the distribution of the Mandel's *h* and *k* statistics, commonly applied to identify laboratories that supply inconsistent results, in the framework of Interlaboratory Studies (ILS). These statistics are usually used to detect those outlier laboratories by testing the hypothesis of reproducibility and repeatability (R&R). The statistical tests involved in the ILS have been currently developed assuming that the measured variables are Gaussian distributed. Thus, the results of the application of these statistical techniques, such is the case of Mandel's *h* and *k* statistics, will be more valid the more plausible is the Gaussian distribution hypothesis. If the variable measured by the laboratories is far from be assumed normal distributed, the application of nonparametric techniques based on bootstrap procedures could be very useful to estimate more accurately the distribution of these statistics and, consequently, the critical values of the tests. Thus, in this case, the laboratories that provide inconsistent results should be identified in a more reliable way.

For the validation of the proposed algorithm, a simulation study has been proposed, where normal, skew normal, and Laplace distributions were simulated, assuming different sample sizes and number of laboratories. The most scenarios the bootstrap approach for the *h* and *k* tests provides better results than those obtained using the parametric classical methodology. Additionally, the proposed bootstrap procedure has been applied to real case studies, such as the ILS corresponding to hematic biometry, on the one hand, and the measure of serum glucose, on the other hand.

1. Introduction

The application of quality control systems to analytical testing laboratories is essential to achieve excellence of the results, i.e. of the analytical information provided. These systems have to be evaluated internally, by the laboratory itself, as well as externally, through the so-called Interlaboratory Studies (ILS), which are designed, organized and managed by independent institutions [1–3].

ILS are collaborative studies, performed by a variable number of testing laboratories, and organized by a separate institution. The latter is in charge of distributing the common material to be tested, as well as analyzing and discussing the results provided by the set of laboratories. The objectives of an ILS are usually the evaluation of the performance of an analytical method or a new test standard, the study of the suitability of a specific laboratory for certification (proficiency testing), or the certification of materials [1,3]. The organizers of these types of studies are often institutions such as the American Society for Testing and Materials (ASTM), which develop and publish standards for the analysis of the chemical and physical properties of substances [4]. Specifically, this institution is in charge of tasks such as ILS design, sample identification, contacting and contracting suppliers, soliciting volunteer laboratories, reviewing the instructions given to the laboratories, collecting and statistically analyzing the results provided by the laboratories, preparing a document indicating the precision of the analysis method and the research report, as well as recognizing the participating laboratories. To achieve the objectives of an ILS study, it is essential for the organizer to provide sufficient information on the material to be analyzed, to draw up a clear work plan, including a detailed description of the experimental work to perform. In addition, it is crucial to ensure a correct and complete statistical analysis of the results, through appropriate protocols that allow

* Corresponding author.
E-mail address: javier.tarrio@udc.es (J. Tarrío-Saavedra).

https://doi.org/10.1016/j.chemolab.2021.104429
Received 17 May 2021; Received in revised form 13 August 2021; Accepted 27 September 2021
Available online 6 October 2021

comparison between laboratories. Moreover, the participants in an ILS study are testing laboratories, whose number ranges from a few to a few hundred, and they must have prior internal quality control procedures, assuming that their results are under statistical control. Participation in an ILS may become mandatory for laboratories, as in the case of accreditation or certification processes.

Regarding the statistical techniques applied in ILS, we highlight those related to the estimation of variability, i.e. with the measurement of the precision of the analysis method and the evaluated laboratories. This variability may be due to natural factors, such as changes in ambient temperature, as well as to assignable causes such as the operators that take the measurements, the use of poorly calibrated instruments, or the incorrect application of the analysis method, among others. Statistical studies of reproducibility and repeatability (R&R) are used for estimating this variability. These studies are developed for evaluating the measurement error of equipment Kenett et al. [5]. In this regard, we should mention some relevant recent works such as that of Browne et al. [6], where two-stage leveraged method has been proposed for assessing the variation of a measurement system in two steps: in the first one, called baseline, a number of parts are measured once, whereas in the second step, a few extreme parts are chosen and remeasured. In addition, the work of Stevens et al. [7] proposes a standard plan for R&R that incorporates a baseline data (produced by the measurement system) and searches for good standard plans with a fixed total number of measurements, assuming the available baseline data, and providing and improvement in terms of precision. It is also worthy to mention the study of Stevens et al. [8], that demonstrates that the use of a proper augmented plan can improve the efficiency to estimate the gauge R&R. Specifically, in the framework of interlaboratory studies, the works of Hund et al. [9] and Vander Heyden and Smeyers–Verbeke [3] provide complete information on the measurement of the accuracy and precision of laboratories and experimental methods evaluated in an ILS, while the ISO standards regulate the application of this type of studies [10].

Among the tasks to be performed in an ILS study to adequately measure the precision and accuracy of methods and laboratories, outlier detection is essential [11]. The motivation for this set of techniques, as part of an ILS, is to detect those laboratories that provide results that are significantly different from others, either in terms of position or dispersion. Discarding laboratories that provide inconsistent results allows us to more accurately estimate the precision and accuracy of methods and laboratories. As for the methods currently used, they are basically univariate, including those based on the measurement of differences between means, such as the Grubbs, Graf and Henning tests, and those that focus on differences in variance, including the Cochran and F tests [2]. There are also robust alternatives for the detection of atypical laboratories, among which we highlight the tests based on the median [12] and on the calculation of robust means [12,13].

It is important to note that, of all the methods for detecting outliers in the framework of an ILS, the use of graphical methods stands out for their ease of interpretation. Among all the alternatives of this type, those based on Mandel's h and k statistics are the most widely used. Mandel's h statistic measures inter-laboratory variability (from the difference between means), while the k statistic estimates intra-laboratory variability. It is hypothesized that the variables measured by the laboratories follow a normal distribution, resulting in the distribution of the h statistic being related to a Student's t, while the distribution of k is related to a Snedecor's F. The graphical methods associated with these statistics consist of the representation of bar charts, 1 bar defining the value of the statistic for each laboratory. These values are compared with respect to critical levels that define the limit above which a laboratory is considered an outlier. The critical values of the h and k statistics are obtained from the t and F distributions, respectively. Currently, there are several computational tools that provide an automatic way to calculate the h and k statistics as well their graphical outputs, as is the case of the `ILS` package of the `R` statistical software [14].

The result of the h and k tests will be more reliable the closer the measurements taken by the laboratories are to a normal distribution. In order to be able to apply the h and k statistics independently of the distribution of the measured variable, we propose to estimate the distribution of the statistics by bootstrap methods. These methods provide estimates of the critical values of the statistics without having to assume a given probability distribution. Nowadays, the use of methods based on resampling techniques has been increasing to the point of being popular not only in the field of statistics, but also in various areas of science, such as biology, chemistry, medicine, geosciences and engineering, among many others.

The bootstrap provides a simple and direct way to carry out statistical inference despite the presence of the outlier. Nevertheless, in the framework of design of experiments, the use of bootstrap resampling were not as used as in other branches of statistics. In this regard, the work of Kenett et al. [5] is a seminal work that has encouraged the proper use of bootstrapping techniques in the design of experiment. Kenett et al. [5] stated that the bootstrapping implementation must emulate the structure of the experiment, taking into account the real number of replicate outcomes, at each one of the experimental conditions. They also pointed out that it is very important to note that the main requirement in order to apply bootstrap resampling is that the real experiment had been replicated at all experimental levels. Thus, the present approach could not be applied to those case studies defined by just one observation or replicate by treatment. For instance, that is the case of the results corresponding to destructive tests. There are alternative methods to deal with these types of experiments, such as the fractional random-weight bootstrap procedure proposed by Xu et al. [15]. Specifically, this methodology can be applied in those cases in which there are a significant proportion of the resamples where estimating all of the parameters in the model is not possible. These cases include situations where the data are greatly censored (response is a rare event), in which there is no sufficient mixing of successes and failures, and experiments in which the number of observations is similar to the number of parameters to be estimated.

In ILS, bootstrap resampling methods have been used to estimate the distribution of functional approximations of the h and k statistics. In fact, when the starting data are curves, smooth functions with respect to time or frequency, it is necessary to adapt the h and k statistics so that, in their calculation, all the information of the curve or experimental data is taken into account. Since there is no starting assumption regarding the distribution of the data, bootstrap methods are essential for estimating the critical values of the tests [2,16]. Also popular are outlier detection methods (in this case outlier curves, not outlier laboratories) based on data depth, whose distribution is also estimated by bootstrap procedures [17]. The robustness of bootstrap methods is one of their main goals and main reasons of their popularity and use. In fact, inference performed by bootstrap methods is not as conditioned by features like heteroscedasticity as parametric methods [5]. In any case, the performance of the bootstrap methodology should be analyzed and compared with the performance of the corresponding parametric alternatives, as indicated in Kenett et al. [5].

The main feature of resampling-based methods is that they provide estimates of the probability distribution of a statistic from only the sample information [18]. The Bootstrap distribution is extremely useful because we can fully characterize the statistic with information about its position, variability and shape. Therefore, it could become a useful tool for determining the empirical distribution of the univariate h and k statistics. The estimation of the distribution of the h and k statistics by bootstrap procedures provides an alternative for their application even in those cases where the starting hypothesis of normality of the data is not satisfied. In fact, the main objective of this work is to provide a robust alternative to the detection of laboratory outliers in an interlaboratory study by extending the use of the h and k statistics.

2. Mandel's h and k statistics

In this section, the main features of scalar Mandel's h and k statistics

will be succinctly defined in the framework of reproducibility and repeatability studies.

2.1. Precision, reproducibility and repeatability

Numerous and different organizations perform interlaboratory studies in order, for instance, to verify compliance with certain quality requirements of a specific material. Therefore, it is very important to have procedures that allow us to evaluate that the measurements obtained are not influenced by the laboratories that perform them. The results must be accurate and precise, thus, it is absolutely necessary to check the reproducibility and repeatability hypotheses. The meeting of these hypotheses can be checked by the application of Mandel's h and k statistics. In this section, we describe the main starting concepts in this study, such as precision and reproducibility and repeatability (R&R) studies. A brief formal description of the Mandel's h and k statistics is also provided.

Studies on the precision of a given analytical method are an important part of ILS, as shown in the ISO 5725-2 [10]. The precision of a measurement system or experimental procedure can be defined as the statistical variability in the measurements we take of the same magnitude on the same sample. Precision is measured in terms of variance or standard deviation of the measurements taken of a given magnitude [19]. This variability will be related to the concepts of reproducibility or repeatability of measurements. In fact, in ILS, reproducibility and repeatability studies are carried out in order to evaluate if the measurements depend on the laboratory that performs them (high reproducibility variability) or if the variability in the measurements performed by each laboratory is too high (repeatability).

In the framework of an interlaboratory study, repeatability is defined as the precision under repeatability [20] conditions, i.e. those conditions that imply that independent measurements are obtained with the same method, on identical materials, in the same laboratory, by the same operator, using the same measuring instrument in short time intervals between measurements, according to ASTM international [20]. On the other hand, reproducibility is defined as the precision under reproducibility conditions [20], that is, conditions where measurements are obtained with the same method applied to the same materials, in different laboratories, with different operators, using different measuring instruments [21].

As pointed out by Kenett and Shmueli [22] the concepts of reproducibility, repeatability and replicability have been referred in scientific literature with different meanings depending on the scientific domain where they are used. For this reason, it is recommended that the researches state the type of generalization intended in each specific case [22]. Two types of generalization can be defined [22]: statistical generalizability, which deals with inferring a population from a sample, and scientific generalizability that deals with extending the application of a model that has been based on a specific population to other populations. Assuming these definitions, the meaning of repeatability in this work is similar to the repeatability meaning in the Gauge R&R studies, i.e. the repeatability studies are applied to estimate the measurement error of a laboratory using a specific experimental procedure from a sample of measurements, in the framework of an ILS. Thus, statistical generalization is performed. In the case of the studies of reproducibility in ILS, the measurement error due to the change of laboratory is estimated from a sample of measurements taken by different laboratories. Therefore, a generalization to future use of a specific testing procedure by different laboratories is done. The change of laboratory often includes changes in experimental instruments (although of the same type), laboratory technicians and may be ambient conditions. Therefore, this type of generalization is closer to scientific generalization. It is also worthy to mention that the concept of replicability is not usually utilized in the framework of ILS.

Once reproducibility and repeatability concepts are defined in the framework of ILS studies, with the corresponding generalization, it is worthy to discuss about the differences of reproducibility and replicability meanings, in order to clarify the scope of the present study. In fact, to state differences between the two concepts is now relatively difficult due to many different domains of science, out of metrology and statistics, as computer science, need to validate the research results, developing methodologies, standards, and corresponding specific nomenclature, in parallel. Thus, in some cases, this leads to establish different meanings for the same terms depending of the scientific field. That is the case for the replicability and reproducibility words. Reproducibility, in the scope of the ILS or collaborative studies, accounts for errors corresponding to different laboratories and equipment, nevertheless using always the same experimental procedure. As pointed out by Miller and Miller [23] and Plesser [24], this is a more restricted definition of reproducibility than that commonly used in metrology. If we focus on other scientific domains such as computer science, different meanings for reproducibility are defined, in addition to distinguish the latter with respect to the concept of replicability. In fact, the Association for Computing Machinery makes differences between reproducibility and replicability by assigning to reproducibility that variability of measurements due to be performed by different team, but using the same experimental setup, whereas the replicability is that variability of measurements due to the changes in the team in addition to changes in the experimental setup [24]. These definitions are in accordance with those proposed by Patil et al. [25], Nichols et al. [26], and the National Academies of Sciences, Engineering, and Medicine [27].

2.2. Definition of univariate Mandel's h and k statistics

The purpose of Mandel's h and k statistics is to detect laboratories that provide atypical or inconsistent results with respect to other laboratories, based on the study of the variability present in the measurement processes. The h and k statistics are used to test the hypotheses of reproducibility and repeatability of the measurements obtained by the laboratories. The results of the tests are usually presented graphically, intuitively and easily understood by non-statisticians, which makes Mandel's statistics one of the most widely used tools to evaluate the quality of the laboratories in an ILS.

2.2.1. Mandel's h

The h statistic is used to estimate the consistency of intra-laboratory results, comparing the internal variability of an individual laboratory with the remaining laboratories. In other words, Mandel's h statistic is used to test the reproducibility hypothesis.

For the case where $l = 1, 2, \ldots, L$ test laboratories are available, performing n replicates of a given experiment on the same material, Mandel's h statistic for the l-th laboratory is defined by the expression

$$h_l = \frac{d_l}{S_x} = \frac{x_l - m}{\sqrt{\sum_{l=1}^{L} \frac{(x_l - m)^2}{L-1}}},$$

where x_l is the mean of the magnitude measured from all replicates performed on the same material (the x_l value is known as the cell mean) and m is the average of the cell means obtained by all laboratories (the mean of the means).

Wilrich [28] provides a detailed definition of the distribution of the h statistic and its critical values. Let $(x_1, x_2, \ldots, x_L)$ be a sample of L observations, where L is the number of laboratories and x_l, with $l = 1, 2, \ldots, L$, the realizations of the random variables X_l, with $l = 1, \ldots, L$, independently and identically distributed according to a normal distribution $N(\mu, \sigma^2)$. Then, the distribution of the h statistic approximates a Student's t distribution with $L - 2^{\circ}$ of freedom,

$$h \approx t_{L-2}.$$

Taking into account this result, the critical value for the Mandel's h statistic [28] can be defined by

$$h_{l;1-\frac{\alpha}{2}} = \frac{(L-1)t_{L-2;1-\frac{\alpha}{2}}}{\sqrt{L\left(t^2_{L-2;1-\frac{\alpha}{2}} + L - 2\right)}},$$

whereby $t_{L-2;1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$ quantile of *twithν* $= L - 2^{\circ}$ of freedom.

2.2.2. Mandel's k

Mandel's k statistic measures intra-laboratory consistency, comparing the internal variability of an individual laboratory with respect to those corresponding to the remaining laboratories. In the case where $l = 1, 2, \ldots, L$ test laboratories are available, performing n replicates of a given experiment on the same material, the Mandel's k statistic for the l-th laboratory is defined by the expression

$$k_l = \frac{S_i}{S_r} = \frac{\sqrt{\sum_{i=1}^{n} \frac{(x_i - x)^2}{n-1}}}{\sqrt{\frac{\sum_{l=1}^{L} S_l^2}{L}}},$$

where S_i is the standard deviation of the number of replicates made by the laboratory (also called intra-cell standard deviation). On the other hand, S_r accounts for the repeatability standard deviation, defined as the square root of the mean of the variances corresponding to each of the L laboratories.

Similarly, one can define the distribution of the k statistic and thus its critical value [28]. Let $\left(S_1^2, S_2^2, \ldots, S_L^2\right)$ be the L sample variances, each calculated from n observed values. Then, under the hypothesis that the observations x_{ji}, with $j = 1, 2 \ldots, L$ and $i = 1, 2, \ldots, n$ are realizations of X_{ji} random variables, identically distributed, independent and distributed according to a normal distribution $N\left(\mu_j, \sigma^2\right)$, the sampling variances S_j^2; $j = 1, 2, \ldots, L$ divided by σ^2 follow a χ^2/ν distribution with $\nu = n - 1^{\circ}$ of freedom. As a result, the k statistic can be approximated to a Snedecor's F distribution as

$$k \approx F_{\nu_1;\nu_2},$$

with $\nu_1 = (L-1)(n-1)$ y $\nu_2 = n - 1$ the degrees of freedom, also defining the critical value of the Mandel's k [28] as

$$k_{l,n;1-\alpha} = \sqrt{\frac{L}{\left(1 + \frac{L-1}{F_{\nu_1,\nu_2;\alpha}}\right)}},$$

whereby $F_{\nu 1,\nu 2;\alpha}$ is the α-quantile of the F distribution, with $\nu_1 = (L-1)(n-1)$ y $\nu_2 = n - 1^{\circ}$ of freedom.

2.3. Approach to repeatability and reproducibility hypothesis tests

The application of a one-way ANOVA is the most usual analysis to detect outliers in an ILS [3]. The starting assumption is that the measurements taken by all laboratories should come from the same population and must be distributed according to a normal probability distribution. Furthermore, the mean and standard deviation between groups should be similar for each laboratory. This is equivalent to saying, in the context of an ANOVA model, that the means and standard deviations of all laboratories are approximately equal for each level of measurement.

The above is equivalent to the reproducibility and repeatability hypotheses, according to which there are no differences between the means of the measurements obtained by each laboratory, nor between the standard deviations, respectively. By means of the h and k statistics, the reproducibility and repeatability hypotheses can be tested, respectively, obtaining as a result the identification or not of outlier laboratories.

The null hypothesis of reproducibility states that there is no difference between the means of the measurements obtained by each laboratory, that is,

$$H_0 = \mu_1 = \mu_2 = \cdots = \mu_L,$$

where $l = 1, 2, 3, \ldots, L$ is the number of laboratories, and μ_l is the mean of the measurements made by each laboratory.

To test the null hypothesis of reproducibility, we will use Mandel's h statistic, assuming that $h \approx t_{L-2}$.

The null hypothesis of repeatability states that there is no difference between the deviations of the measurements taken by each laboratory, i.e.

$$H_0 = \sigma_1 = \sigma_2 = \cdots = \sigma_L,$$

where $l = 1, 2, \ldots, L$ is the number of laboratories and σ_l is the standard deviation of each laboratory.

To test the null hypothesis of repeatability, Mandel's k statistic is used, assuming that $k \approx F_{(L-1)(n-1);(n-1)}$.

3. Outlier detection methodology based on bootstraped Mandel's h and k statistics

As shown in the previous section, when estimating Mandel's h and k statistics, it is assumed that the measured variable follows a normal distribution. If this hypothesis is satisfied, one can equally well assume that the h and k statistics also follow parametric probability distributions, specifically a Student's t and a Snedecor's F, respectively. Therefore, it is expected that, the more different the distribution of X is with respect to a normal one, the less reliable the critical values of the h and k statistics, calculated from the t and F distributions, will be. Ultimately, this fact justifies the development of robust approximations, in the face of non-normality of the measurements, of the h and k statistics. This can be achieved by nonparametric estimation of their distributions using bootsrap resampling procedures. The proposed methodology is based on two main steps consisting, on the one hand, in the development of an algorithm to estimate the bootstrap distribution of Mandel's h and k statistics and, on the other hand, in the formulation and testing of the reproducibility and repeatability hypotheses applied on the measurements taken by the laboratories participating in the ILS.

3.1. Algorithm to estimate the critical values corresponding to the bootstrap distribution of the h and k statistics

In practice, when applying the h and k statistics for the detection of outlier laboratories in the scope of an ILS, the compliance with the starting hypotheses is not always verified before estimating their critical values (at a pre-specified significance level, α) from the t and F parametric distributions, respectively. In order to dispense with the assumption of normality of the measurements, this section proposes an algorithm based on the application of bootstrap resampling methods for the estimation of the distributions of the h and k statistics. This algorithm allows us to calculate the critical values, limits beyond which a laboratory (identified as an outlier) will be considered to provide data inconsistent with those provided by the others.

A nonparametric bootstrap method is applied. The main purpose is to obtain an estimate of the bootstrap distribution of the statistic of interest, to consequently obtain information about the true distribution of the statistic. This similarity can be established on the basis of the plug-in principle. The idea of the plug-in principle is that, to estimate a population parameter, the corresponding statistic is used for each sample [18]. That is, it is possible to use the bootstrap distribution of a statistic to estimate the value of a population parameter, which is one of the reasons why the bootstrap distribution is so important.

Bootstrap distributions have the same shape and dispersion as the statistic sampling distribution, but centered on the value of the statistic in

the original sample [18]. Therefore, it is very important that the original sample is representative of the population. The bootstrap distribution is centered on the value of the statistic in the original sample. Thus, the bootstrap distribution does not provide information about the center of the sampling distribution of the statistic, but it does provide important information about the bias, since, although the bootstrap and sampling distribution do not have the same distribution in the original sample, the bootstrap distribution does not provide information about the center of the sampling distribution of the statistic, they do have the same skewness [18].

The steps of the proposed methodology to determine the difference between laboratories, using the nonparametric bootstrap method, are detailed below:

1. The data from the different laboratories are grouped into a single set of size $L \times n$.
2. All atypically high or low measurements are discarded. For this purpose, a box plot can be used, keeping only those observations that are not outside the maximum and minimum values of the box plot.
3. Sampling with replacement is performed from the data that were not separated to then generate a resample of the same size as the original dataset.
4. From the dataset created in the previous step, n observations are randomly assigned to a number L of different groups.
5. For each resample, calculate the value of the statistic of interest (h or k) for each laboratory.
6. Repeat steps 3, 4, 5,6 a large number of times in order to best estimate the distribution of the h and k statistics, e.g., $B = 500$. As a result, the bootstrap distribution of the statistic is obtained.

In the case of h.

7. The critical values of the h statistic, at the α significance level, are determined, using the quantiles of the bootstrap distribution of the h statistic for the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ levels, i.e., $h_{\frac{\alpha}{2}}^{boot}$ and $h_{1-\frac{\alpha}{2}}^{boot}$.
8. The value of the Mandel's h is calculated for each of the $l = 1, 2, \ldots L$ laboratories. If any of the h_l meets $h_l < h_{\alpha/2}^{boot}$ or $h_l > h_{1-\alpha/2}^{boot}$, the corresponding laboratory is considered as an outlier.

In the case of k.

9. The critical value of the k statistic is estimated, at the α significance level, using the quantile of the bootstrap distribution of the k statistic, i.e., k_{α}^{boot}.
10. Values of the k statistic are computed for each of the $l = 1, 2, \ldots k$ laboratories. If any $k_l > k_{\alpha}^{boot}$, the corresponding l laboratory will be considered as an outlier, i.e. it provides inconsistent results with respect to those obtained by the other laboratories.

4. Validation of the proposed methodology in a simulation study

The proposed methodology is validated in very different simulation scenarios. They are defined in order to compare the performance of the classic parametric tests based on the h and k statistics and their bootstrap approximations, in a wide variety of possible situations, plausible with the features of the data that can be actually obtained by experimental methods. The simulated scenarios are defined by the variation in the parameters of three probability distributions, by which the various observations obtained by the laboratories are simulated, including the presence of outliers. The three distributions chosen are the normal distribution, the Laplace distribution and the skew normal distribution, which differ both in their skewness and kurtosis features.

1. Standard normal distribution, $X \sim N(0, 1)$, with the density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

2. Laplace distribution, $X \sim L(0, 1)$, with the density function:

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b} \right).$$

3. Skew normal distribution, $X \sim SN(\varepsilon, \omega^2, \gamma)$, defined by the following density function:

$$f(x) = \frac{2}{\omega\sqrt{2\pi}} e^{-\frac{(x-\varepsilon)^2}{2\omega^2}} \int_{-\infty}^{\alpha\left(\frac{x-\varepsilon}{\omega}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Considering the following values for their location, scale and shape parameters as follows: $\varepsilon = 0$, $\omega = 1$, and $\gamma = 1$, respectively.

4.1. Description of the simulation study

The algorithm described in Section 3 was applied to the different scenarios proposed for the simulation study. Each of these scenarios is defined by its corresponding shift with respect to the mean and standard deviation under the null hypothesis. These shifts, measured in units of the simulated magnitude, will be different according to the distribution being simulated.Non-compliance with the reproducibility hypothesis.

To simulate inconsistent laboratories, the following changes in the mean for the normal distribution and the Laplace distribution have been considered:

$$\mu \in \{-3, -2, -1, 0, 1, 2, 3\}$$

In the case of the skew normal distribution, we proceeded to vary the location parameter ε as shown below.

$$\varepsilon \in \{-3, -2, -1, 0, 1, 2, 3\}.$$

Non-compliance with the repeatability hypothesis.

To simulate the laboratories that provide inconsistent results with respect to those obtained by the other laboratories, the following values have been considered for the standard deviation in the normal distribution, with $\sigma = 1$ corresponding to the null hypothesis:

$$\sigma \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}.$$

For the skew normal distribution the ω scale parameter is varied as

$$\omega \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}.$$

In the case of Laplace distribution, different values of the b scale parameter are chosen:

$$b \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}.$$

Regarding the simulated laboratories, L consistent laboratories were generated, while one ($m = 1$) laboratory was generated under the alternative hypothesis (laboratory with inconsistent results), defined by the variation of the parameter values of the study distributions with respect to those corresponding to H_0. With a total of $L' = (L + m)$ laboratories, $mc = 1000$ ILS are performed on a single material. Within each ILS, when evaluating the performance of the approximate h and k by resampling, $B = 500$ bootstrap resamples are performed. As for the number of simulated laboratories, scenarios defined by $L \in \{5, 10\}$ with $m = 1$ have been generated, taking a number of observations equal to $n \in \{3, 6\}$, maintaining a significance level of $\alpha = 0.01$.

Table 1
Rejection proportion of the H_0 of reproducibility, for Mandel's h, from data assuming normal distribution.

μ Bootstrap Method									μ Parametric Method						
L	n	−3	−2	−1	0	1	2	3	−3	−2	−1	0	1	2	3
5	3	0.613	0.256	0.051	0.010	0.058	0.274	0.611	0.627	0.266	0.051	0.009	0.057	0.282	0.637
5	6	0.884	0.561	0.115	0.010	0.143	0.552	0.883	0.886	0.561	0.116	0.010	0.143	0.553	0.883
10	3	0.925	0.569	0.109	0.017	0.113	0.567	0.930	0.919	0.565	0.100	0.013	0.108	0.556	0.923
10	6	0.999	0.906	0.272	0.009	0.238	0.900	0.999	0.999	0.897	0.267	0.009	0.235	0.894	0.999

Table 2
Rejection proportion of the H_0 of reproducibility, for Mandel's h, from data assuming Laplace distribution.

μ Bootstrap Method									μ Parametric Method						
L	n	−3	−2	−1	0	1	2	3	−3	−2	−1	0	1	2	3
5	3	0.337	0.141	0.047	0.012	0.051	0.127	0.341	0.359	0.156	0.047	0.012	0.052	0.141	0.364
5	6	0.599	0.285	0.074	0.011	0.079	0.269	0.594	0.615	0.293	0.075	0.011	0.081	0.282	0.621
10	3	0.616	0.247	0.071	0.016	0.063	0.260	0.612	0.630	0.258	0.071	0.016	0.063	0.269	0.631
10	6	0.908	0.532	0.124	0.018	0.112	0.542	0.912	0.910	0.540	0.125	0.018	0.113	0.546	0.914

Table 3
Rejection proportion of the H_0 of reproducibility, for Mandel's h, from data assuming skew normal distribution.

μ Bootstrap Method									μ Parametric Method						
L	n	−3	−2	−1	0	1	2	3	−3	−2	−1	0	1	2	3
5	3	0.779	0.414	0.093	0.009	0.100	0.377	0.749	0.793	0.424	0.090	0.008	0.100	0.406	0.780
5	6	0.968	0.734	0.179	0.008	0.187	0.734	0.967	0.963	0.734	0.178	0.008	0.194	0.746	0.969
10	3	0.984	0.803	0.198	0.010	0.195	0.773	0.986	0.982	0.783	0.185	0.010	0.190	0.778	0.985
10	6	1.000	0.982	0.409	0.011	0.374	0.976	1.000	1.000	0.981	0.397	0.011	0.379	0.976	1.000

Once the hypothesis test has been performed, the power of the test is calculated for the different scenarios, in order to evaluate the proportion of rejection of the null hypothesis. That is, the proportion of times that a laboratory is detected as an outlier when its data come from a distribution different from that of the remaining laboratories.

It is important to note that the scenarios corresponding to the parameter values under the null hypothesis are also included in the reproducibility and repeatability tests. In the latter case, the proportion of times that H_0 is rejected when it is actually true will be shown.

The aim is to make a comparison of the power values obtained in the test of the repeatability and reproducibility hypotheses obtained, on the one hand, following the classical procedure (assuming a parametric distribution for h and k) and, on the other hand, using the proposed methodology (estimating the distribution of the h and k statistics by bootstrap). For this purpose, the critical values of the h and k statistics are calculated using the parametric classical method and the new proposed method. To calculate the critical values of Mandel's h and k statistics in the parametric classical way, as required by ISO 5725, we use the package `metRology` [29] available at https://cran.r-project.org.

4.2. Study of the power curves of the nonparametric and classical parametric test approaches

4.2.1. Study and comparison of the h statistic power

The following tables show rejection ratios of the H_0 of reproducibility using the test based on the h statistic.

Tables 1–3 compare the power values obtained for the h statistic using the bootstrap method and the parametric classical method when the data come from a normal distribution, Laplace distribution and skewed distribution, respectively. It is observed that the power values of the two methods are similar in all three cases. However, the power of the bootstrap method exceeds in most cases, with a very slight difference, the power of the classical parametric method.

Fig. 1 shows the behavior of the power curve for the h statistic when the data come from a) normal distribution, b) Laplace distribution and c) skew normal distribution. Different sample sizes (segmented line when $n = 3$ and continuous for $n = 6$) and different number of laboratories are considered. For the case of the three distributions a), b) and c), we can observe that, when the laboratory sample size and the number of laboratories involved in the ILS increase, the test power also significantly increases.

Additionally, in the case of the Laplace distribution, b), the classical parametric h statistic test and the bootstrap approximation of the test are characterized by very similar powers. The power of the bootstrap alternative is slightly higher in those cases where the mean of the simulated random variable is smaller than its value under the H_0, while it is very slightly lower when the mean values exceed the one corresponding to the H_0 (see Table 2).

Summarizing, although the proposed methodology has slightly higher power than the parametric alternative, the differences in terms of power between the alternatives is minimal for the h statistic.

4.2.2. Study and comparison of the k statistic power

Tables 4–6 show rejection ratios of the H_0 of repeatability by using the test based on the k statistic.

The results of the power curves corresponding to the scenarios defined by the simulation of normal distributions, Laplace distributions and skew normal distributions, respectively, are shown in Figs. 2 and 3. Fig. 2 shows that the power is higher the greater the number of laboratories and the greater the number of experimental tests or replicates they perform.

In fact, as far as classical parametric tests are concerned, ASTM or ISO standards could have relied on results similar to those presented here to define the criteria for the minimum number of participating laboratories. Similarly, we see that as the variation in standard deviation increases, the ability for outlier detection of the bootstrap resampling-based test increases.

In Fig. 3, we observe the behavior of the test powers for each method in the three cases of study, normal distribution, a), Lapace distribution, b), and skewed normal distribution, c). In general, if we compare the

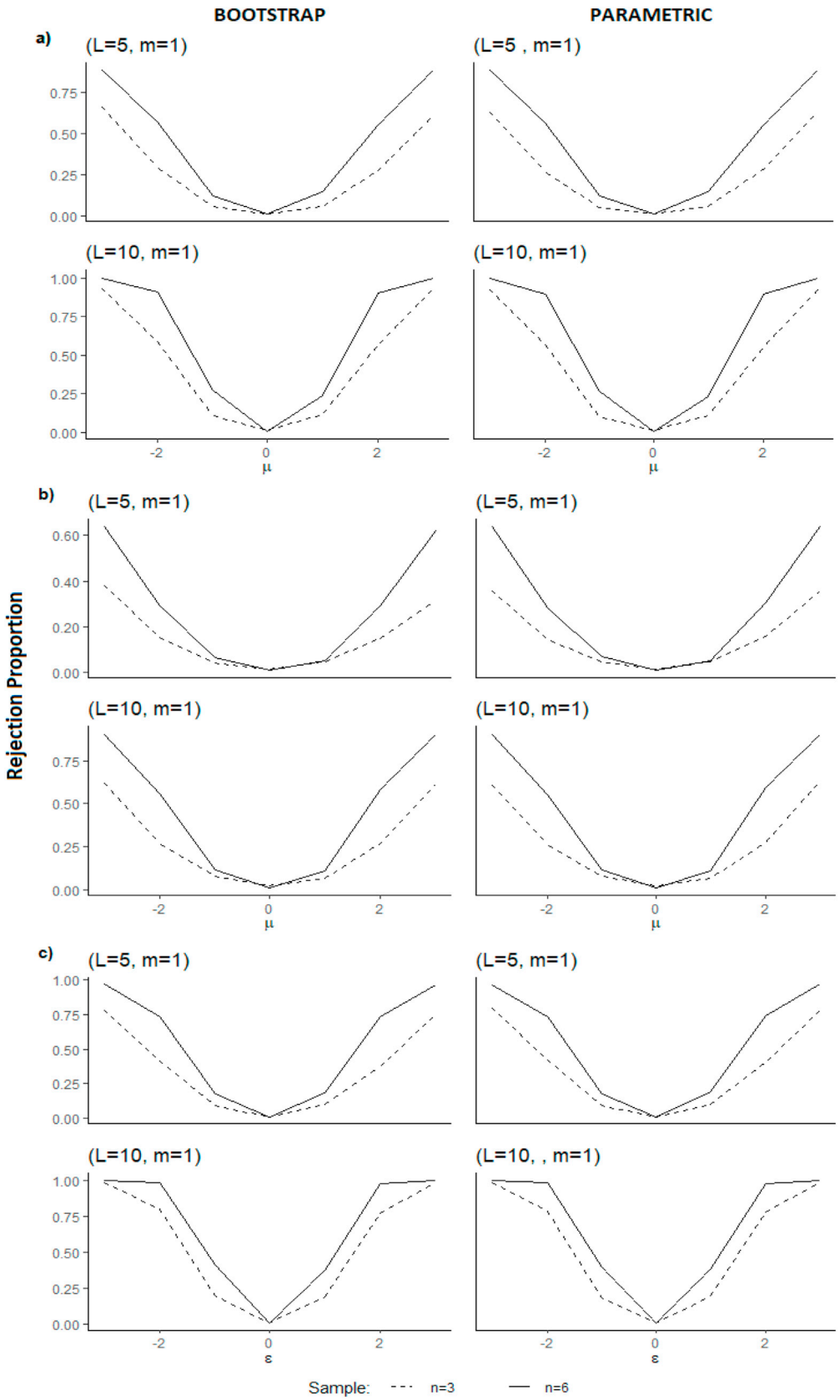

Fig. 1. Power of the test for *h* Mandel's statistic, the bootstrap method on the left and parametric method on the right. With L laboratories under null hypothesis and *m* laboratories under the alternative hypothesis, varying the sample size with $n = 3$ (dashed curve) and $n = 6$ (solid curve), and considering the three distributions of data in the simulation study: a) normal distribution, b) Laplace distribution, and c) skew normal distribution.

Table 4
Rejection proportion of the H_0 of repeatability, for Mandel's k, from data assuming normal distribution.

σ Bootstrap Method									σ Parametric Method						
L	n	−3	−2	−1	0	1	2	3	−3	−2	−1	0	1	2	3
5	3	0.012	0.088	0.221	0.349	0.462	0.568	0.678	0.007	0.071	0.193	0.324	0.438	0.545	0.647
5	6	0.01	0.224	0.517	0.707	0.852	0.914	0.941	0.007	0.182	0.456	0.652	0.823	0.889	0.927
10	3	0.014	0.122	0.292	0.455	0.535	0.658	0.735	0.008	0.103	0.265	0.421	0.507	0.637	0.698
10	6	0.024	0.253	0.565	0.771	0.872	0.957	0.958	0.013	0.192	0.501	0.734	0.852	0.937	0.949

Table 5
Rejection proportion of the H_0 of repeatability, for Mandel's k, from data assuming Laplace distribution.

σ Bootstrap Method									σ Parametric Method						
L	n	−3	−2	−1	0	1	2	3	−3	−2	−1	0	1	2	3
5	3	0.034	0.115	0.202	0.301	0.411	0.516	0.58	0.034	0.115	0.201	0.3	0.41	0.516	0.576
5	6	0.05	0.264	0.419	0.62	0.714	0.821	0.871	0.046	0.25	0.411	0.613	0.71	0.815	0.868
10	3	0.04	0.11	0.257	0.362	0.486	0.552	0.591	0.035	0.105	0.252	0.358	0.479	0.548	0.584
10	6	0.057	0.25	0.449	0.651	0.778	0.829	0.904	0.05	0.235	0.439	0.64	0.762	0.82	0.896

Table 6
Rejection proportion of the H_0 of repeatability, for Mandel's k, from data assuming skew normal distribution.

σ Bootstrap Method									σ Parametric Method						
L	n	−3	−2	−1	0	1	2	3	−3	−2	−1	0	1	2	3
5	3	0.014	0.098	0.241	0.378	0.474	0.572	0.625	0.011	0.083	0.219	0.356	0.449	0.553	0.610
5	6	0.020	0.187	0.445	0.662	0.808	0.891	0.905	0.013	0.148	0.385	0.619	0.774	0.868	0.884
10	3	0.016	0.118	0.271	0.457	0.592	0.633	0.742	0.012	0.101	0.238	0.413	0.545	0.607	0.712
10	6	0.019	0.248	0.560	0.774	0.885	0.923	0.966	0.013	0.197	0.508	0.732	0.859	0.905	0.958

results of the parametric methodology with the proposed nonparametric procedure for the tests of the h and k statistics, we observe that the power of the proposed bootstrap procedure is always higher for all the simulation scenarios, even when the sample size is small. The powers take higher values with respect to the classical parametric test, these differences being more noticeable for the case of the normal distribution, a), and, mainly, for skew normal, c). For the Laplace distribution (b), the power values corresponding to the two methods were very similar.

To summarize, the highest power values for the test of the k statistic performed with the proposed bootstrap methodology are obtained when the simulated variable comes from a skew normal distribution and the sample size is small. Therefore, the new procedure, in addition to being able to be used in those cases in which the normality hypotheses are satisfied, is especially recommended in those cases in which each laboratory provides a relatively small number of measurements of a variable distributed according to an asymmetric probability distribution.

5. Validation of the proposed methodology with real case studies

In this section, we apply the classical parametric tests of Mandel's h and k statistics, as well as the new bootstrap alternative proposed in this work, to two real databases from two different case studies:

- A case study of blood biometry in which the measurements of several variables taken in blood tests are shown.
- Case study defined by experimental blood glucose measurements taken from various patients. This database, called "Glucose", is included in the `R` package `ILS` [14].

Regarding the software available to detect outlier laboratories within an ILS, through parametric contrasts of the h and k statistics, there are two packages available in the `R` software repository, which follow the procedures described in the ISO 5725 and ASTM E-691 standards. These are the package `ILS` [14] and the package `metRology` [29], available from the Comprehensive `R` Archive Network [30]. Both will be used in the applications shown below, comparing their results with those obtained using the proposed bootstrap methodology.

5.1. Blood biometrics

The objective of this ILS is to determine the accuracy of a blood analysis method based on computerized blood biometry techniques. For this purpose, the procedures established in the ASTM E-691 and ISO 5725 standards were followed, considering as the study population a group of 9 clinical laboratories in the city of Ibarra (Ecuador), which have taken observations of 11 different variables: hemoglobin level (A), hematocrit amount (B), red blood cell amount (C), white blood cell amount (D), neutrophil percentage (E), lymphocyte percentage (F), intermediate cell percentage (G), mean globular hemoglobin (H), mean modular hemoglobin concentration (I), mean globular volume (J) and platelet amount (K).

5.1.1. Detection of outlier laboratories by classical parametric test of using the h and k statistics

The results of the application of the classical parametric test based on Mandel's h and k statistics to the measurements obtained by the different laboratories for each of the quantities analyzed, which in this context are referred to as "materials", are presented below.

5.1.1.1. Results of the parametric test based on the h statistic. Fig. 4 shows that laboratories 1, 2, 9 have been detected as outliers (provide inconsistent results with respect to those provided by the other laboratories) when analyzing "materials" F, K and E, respectively. In addition, laboratory 4 shows a characteristic negative bias in the values of the h statistics for the analyzed materials.

5.1.1.2. Results of the parametric test based on the k statistic. Fig. 5 shows that laboratory 1 provides atypical results compared to those presented by the other laboratories when the materials being analyzed are A and I.

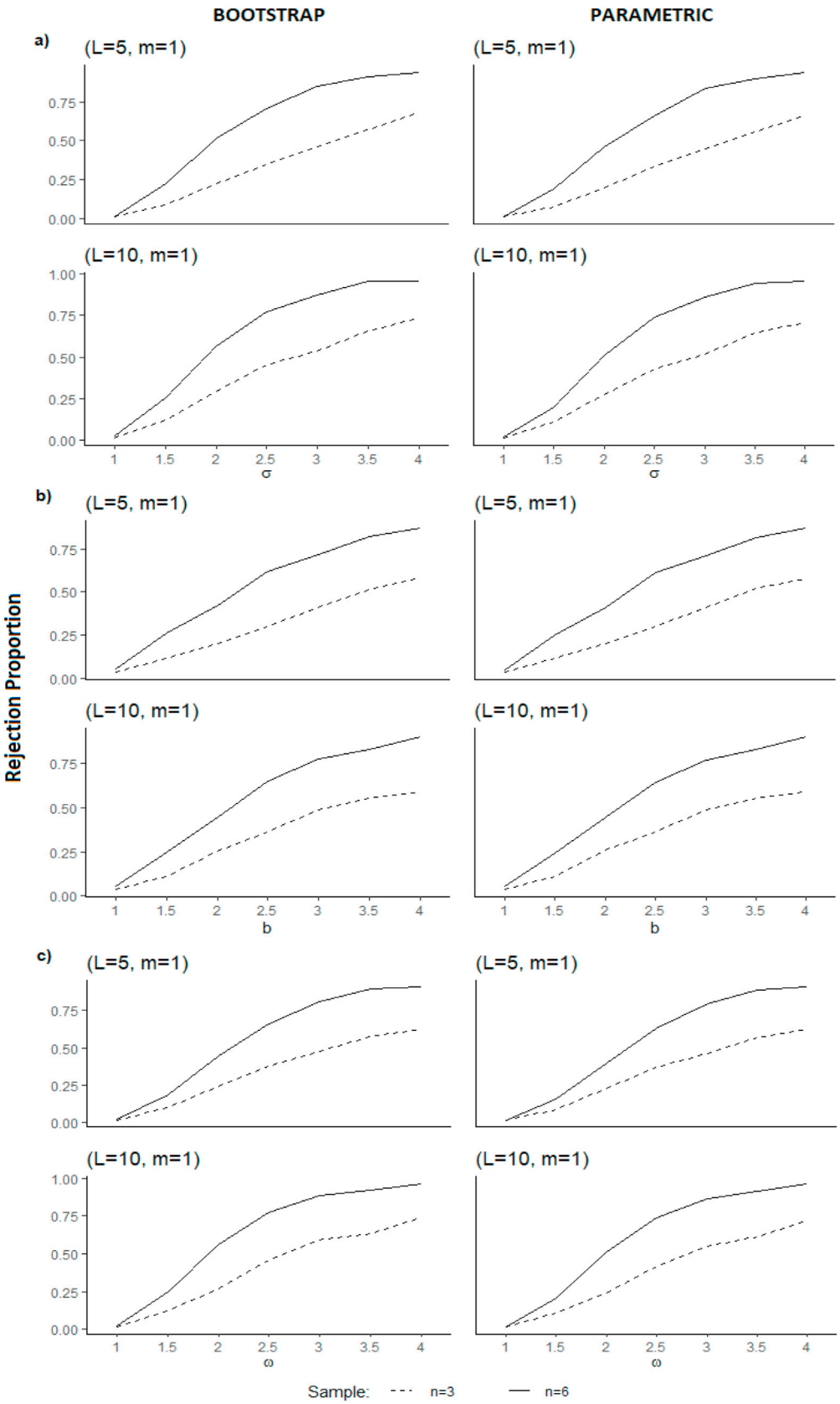

Fig. 2. Power of the test for k Mandel's statistic, the results of the bootstrap method are on the left, and those of the parametric method on the right. L laboratories under H_0 and m laboratories under H_1 are assumed, varying the sample size for each lab with $n = 3$ (dashed curve) and $n = 6$ (solid curve), and considering the three simulated distributions: a) normal distribution, b) Laplace distribution, and c) skew normal distribution.

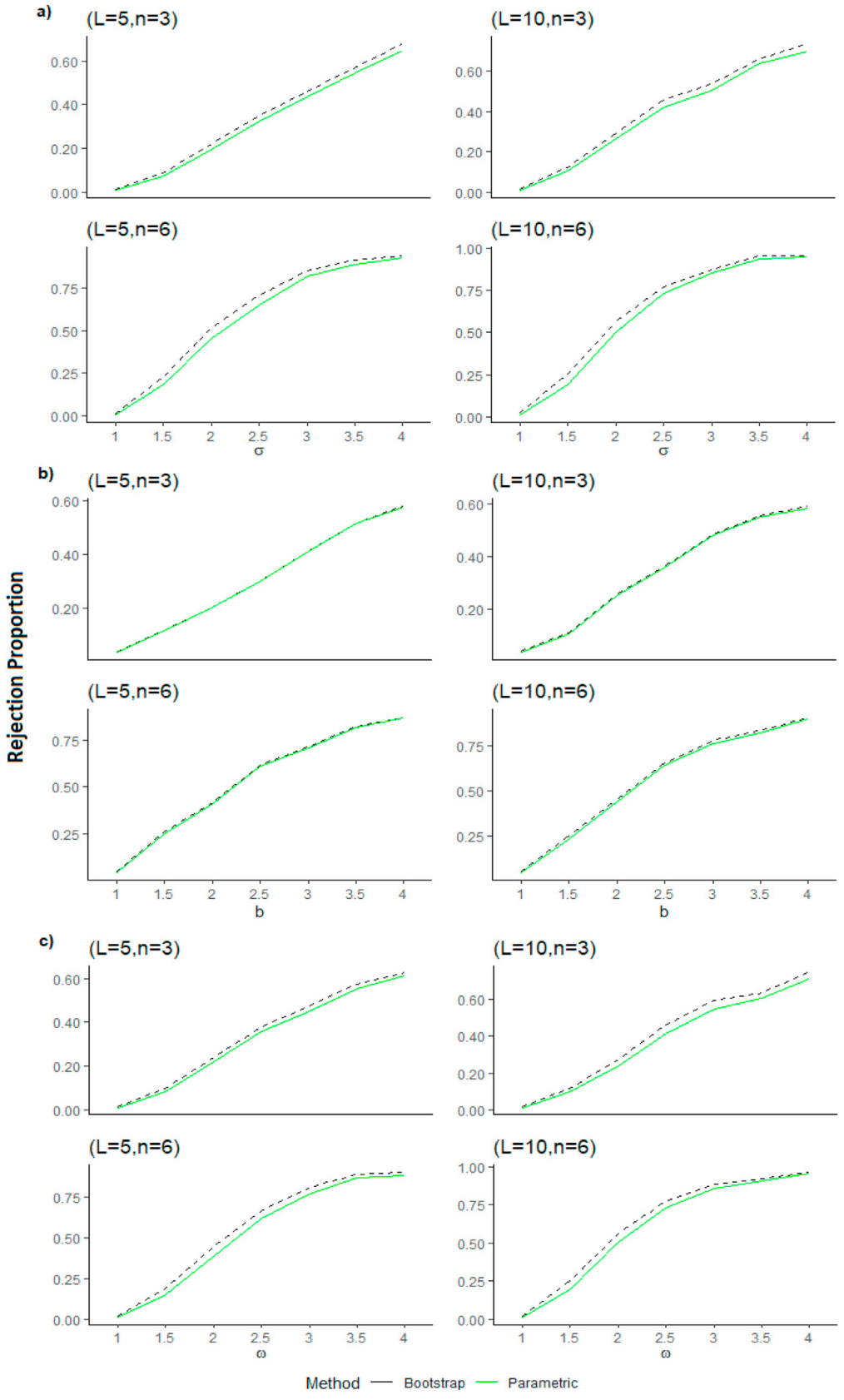

Fig. 3. Power of the test for *k* Mandel's *k* statistic using the proposed Bootstrap method (dashed curve) and also assuming parametric distribution for *k*. We assume L laboratories under H_0 and *m* laboratories under H_1, with two different values for the sample size for each lab, $n = 3$ and $n = 6$. The results are shown assuming the three different simulated distributions for the data: a) normal distribution, b) Laplace distribution, and c) skew normal distribution.

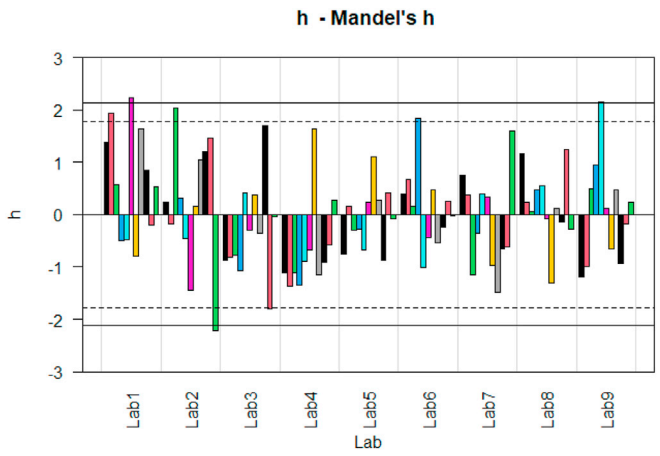

Fig. 4. Bar chart for Mandel's h statistic applied to blood biometrics dataset.

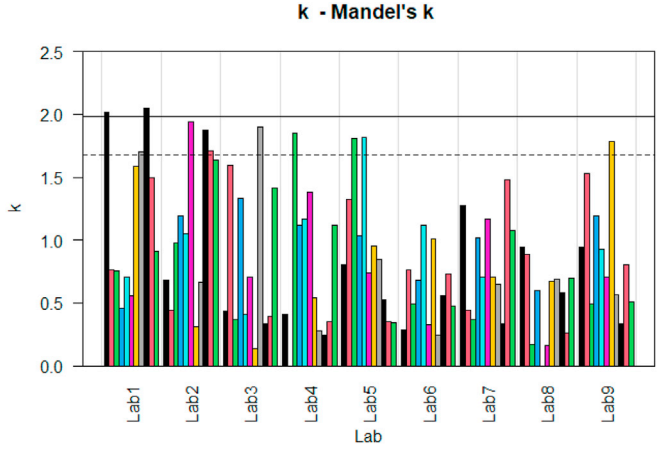

Fig. 5. Bar chart for Mandel's k statistic applied to blood biometrics dataset.

Table 7
Critical values for Mandel's statistics.

Level	$h_{1-\frac{\alpha}{2}}$	$h_{\frac{\alpha}{2}}$	$k_{1-\alpha}$
A	2.11	−2.063	1.742
B	2.163	−1.991	1.929
C	2.197	−1.992	2.006
D	2.178	−2.062	1.923
E	2.248	−1.964	2.036
F	1.989	−2.172	1.931
G	2.087	−2.049	1.795
H	2.124	−2.057	1.842
I	2.192	−1.858	1.801
J	2.108	−2.136	1.885
K	2.146	−2.04	1.879

5.1.2. Detection of outlier laboratories by nonparametric boostrap test of the h and k statistics

The critical values of the h and k statistics are determined using the bootstrap distribution of the statistics, considering that a laboratory will be detected as an outlier if $h_l < h^{boot}_{\frac{\alpha}{2}}$ or $h_l > h^{boot}_{1-\frac{\alpha}{2}}$ for the h statistic, and if $k_l > k^{boot}_{\alpha}$ for the k statistic. The critical values of the statistics, $h^{boot}_{\frac{\alpha}{2}}$, $h^{boot}_{1-\frac{\alpha}{2}}$ and k^{boot}_{α}, are given in Table 7.

5.1.2.1. Results of the nonparametric bootstrap test based on Mandel's h statistic. To perform the Interlaboratory Study, $B = 1000$ bootstrap resamples were generated for each material; and for each resample, the value of the h statistic was calculated, with which the bootstrap distribution of the statistic was generated. Fixing a significance level of $\alpha = 0.01$, we obtained the critical values for each material detailed in Table 7.

As was the case when applying the classical parametric test of the h statistic, Fig. 6 shows that the nonparametric bootstrap test identifies as outliers the results that laboratory 1 obtains from material F and the measurements that laboratory 2 provides from material K. However, at a significance level of $\alpha = 0.01$, laboratory 9 is not detected as an outlier when it analyzes material F (even if only slightly), as was the case with the classical contrast. This is to be expected, given the narrow margin by which lab 9 is detected as an outlier by the parametric test, also taking into account the very slight differences in the power of the parametric and bootstrap tests shown in the simulation study for the h statistic.

5.1.2.2. Results of the nonparametric bootstrap test based on Mandel's k statistic. In Table 7, the critical values of the k statistic that were determined by the new bootstrap methodology are presented.

In Fig. 7 we can notice that the laboratories detected as inconsistent were the first laboratory for material A and I, the third laboratory for material H, and the second laboratory for material F and I. It is found, therefore, that the proposed bootstrap methodology provides a more powerful test for the k statistic than the parametric test, in fact we detected two additional outlier laboratories, one for material F and one for I, which in the classical method were at the limit of being considered as outliers with an $\alpha = 0.01$.

5.2. Glucose analysis in blood

In this section, the case study of blood glucose measurement is shown, the results of which are available in the database called "Glucose", available in the `ILS` package. The `ILS` package implements several tests to identify laboratories that show inconsistent results when compared with others in the framework of an interlaboratory study. It shows alternatives for univariate and functional data analysis. The univariate approach is based on ASTM E691-08, implementing techniques such as Mandel's h and k statistics, and Cochran's and Grubbs' tests, as well as including the ANOVA table with F and Tukey's tests.

The "Glucose" dataset is composed of the results of experimental tests to measure blood glucose concentration for the control and prevention of diseases such as diabetes. The ILS is composed of eight laboratories that tested five different blood samples, each one labelled with its reference, ranging from low to very high sugar content. Three replicates were obtained for each sample.

This case study appears in the text of ASTM E691-08 [20], having identified the outlier laboratories by applying the parametric tests of Mandel's h and k statistics. This will allow us to validate the proposed bootstrap methodology by comparing its results with those presented in the standard, obtained using the parametric tests.

5.2.1. Detection of outlier laboratories by classical parametric tests for h and k statistics

The results for the h and k statistics are shown in Fig. 8 and Fig. 9, respectively, assuming a significance level of $\alpha = 0.01$, which is the one used in the ASTM standard.

For the h statistic (Fig. 8), laboratory 4 and material C were identified as inconsistent, while for the k statistic (Fig. 9), laboratory 2, when measuring material E, and laboratory 4, when analyzing material C, were identified as outliers.

5.2.2. Detection of outlier laboratories by nonparametric bootstrap tests for h and k statistics

The new proposed methodology is applied with a significance level of $\alpha = 0.01$, obtaining, applying the bootstrap approach, the following results for the h and k statistics (Table 8).

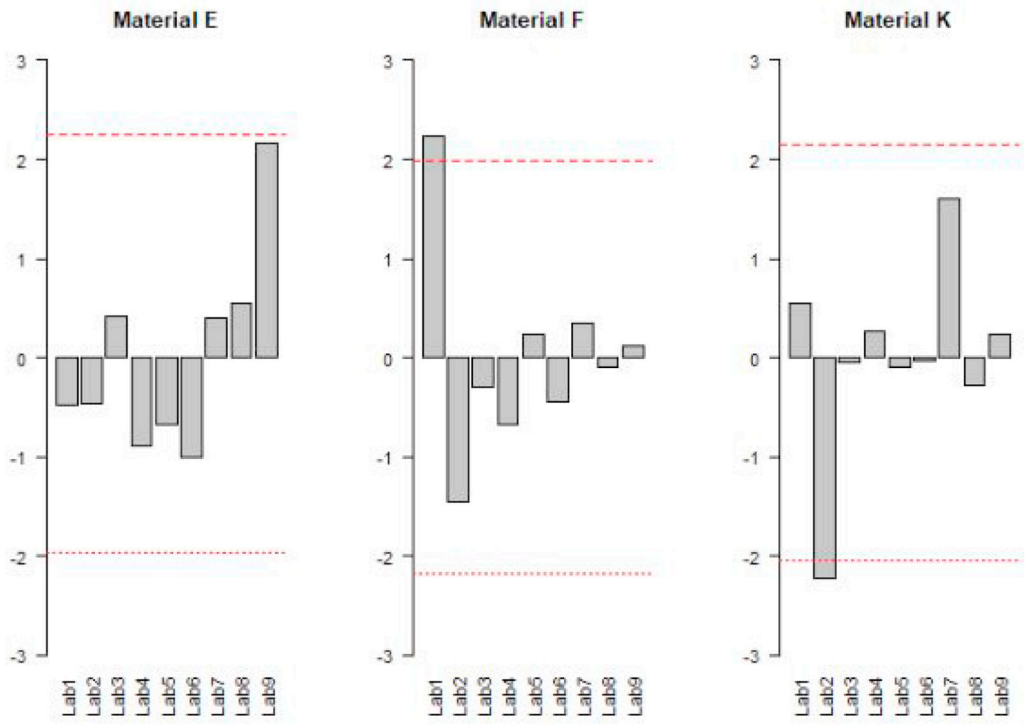

Fig. 6. Materials detected as outlier for the *h* statistic.

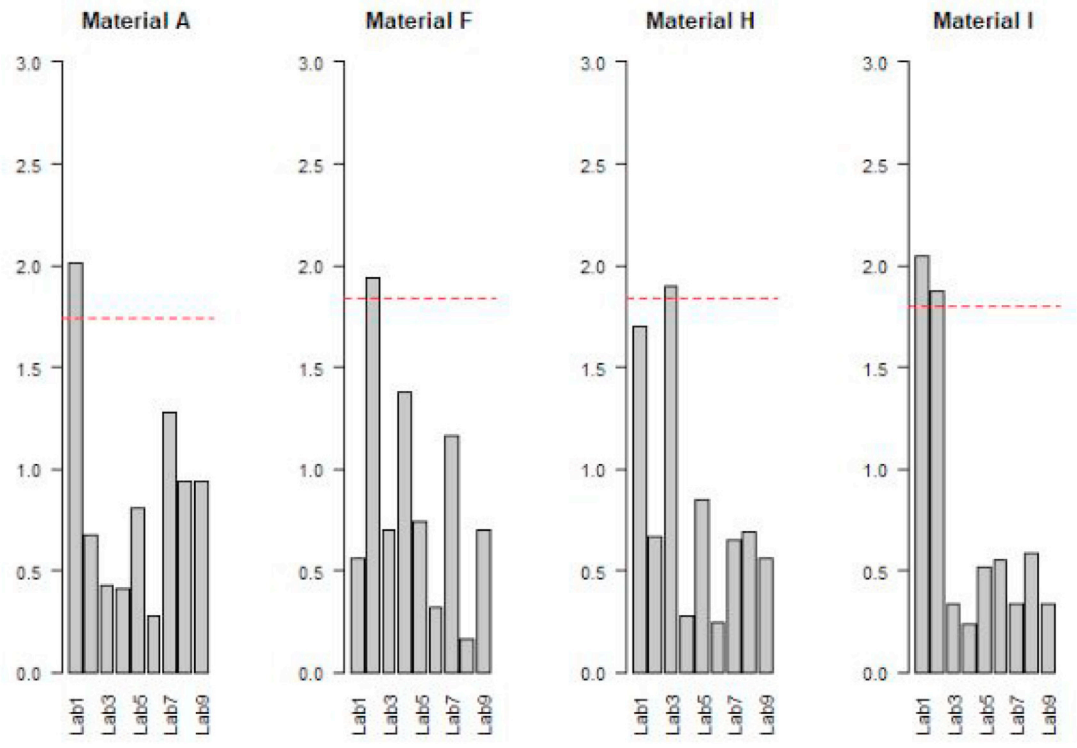

Fig. 7. Materials detected as outlier for the *k* statistic.

Figs. 10 and 11 show that the same laboratories and materials that had been identified by the classical parametric tests were identified as outliers. Therefore, it is concluded that the two methods, bootstrap and classical, provide the same results with regard to the detection of outlier laboratories. This fact help to validate the results obtained by the bootstrap-based test proposed in this study.

6. Conclusions

In this work, a new nonparametric alternative for the detection of outlier laboratories in the context of ILS has been proposed. Specifically, nonparametric tests of the *h* and *k* statistics have been proposed, for which their distributions were estimated using the bootstrap procedure.

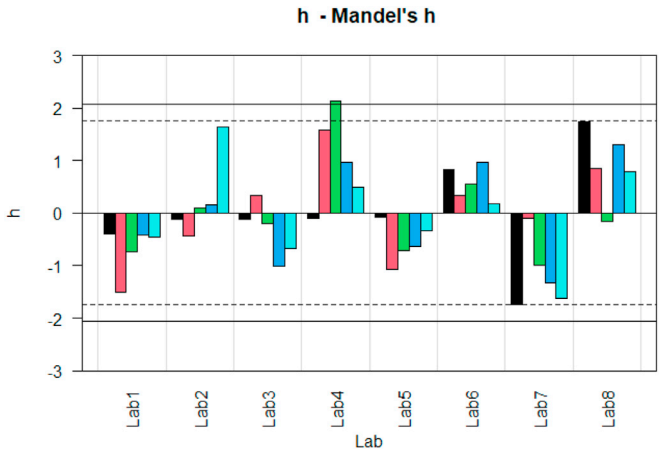

Fig. 8. Bar chart for Mandel's h statistic.

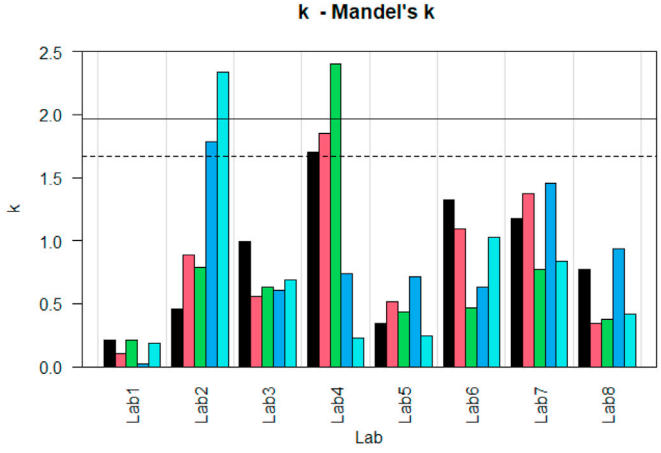

Fig. 9. Bar chart for Mandel's k statistic.

Table 8
Critical values for Mandel's statistics.

Level	$h_{1-\frac{\alpha}{2}}$	$h_{\frac{\alpha}{2}}$	$k_{1-\alpha}$
A	2.026	−2.042	1.975
B	2.036	−2.068	1.835
C	2.013	−2.044	1.838
D	2.063	−2.079	1.908
E	1.964	−2.145	1.935

They are, therefore, an alternative to the classical parametric tests, which start from the hypothesis of normality for the variable measured by the laboratories, being the distributions of the h and k statistics also parametric, under the null hypothesis of reproducibility and repeatability, respectively.

Resampling techniques have proved to be very useful for estimating the empirical distributions of the Mandel's h and k statistics and, consequently, for calculating their critical values. In this case, their application allowed us to approximate the distributions and calculate the quantile of Mandel's h and k statistics. To approximate these distributions, an algorithm was developed based on the null hypothesis that all the data come from the same distribution. Thus, it is necessary to resample the data by previously eliminating the outliers.

A simulation study has been carried out to validate this new nonparametric approach to perform reproducibility and repeatability tests of measurements in an ILS, detecting laboratories that provide outlier data, based on the h and k statistics. Three different probability distributions (normal, skew normal and Laplace) have been simulated, varying their position, variability, degree of skewness and kurtosis by modifying their parameters. Taking into account that the number of laboratories (at two levels, 5 and 10) and the number of replicates obtained by each one (3 and 6) were also varied, a total of 18 different scenarios were simulated. In each of them, the power of the tests based on the h and k statistics was estimated, either in their classical parametric form or for the nonparametric alternative, in which their distributions are estimated by bootstrap resampling. Among the most outstanding results, it is observed that the proposed nonparametric bootstap test based on the k statistic is significantly more powerful than the classical parametric version, in all the simulated scenarios, while the nonparametric bootstrap test based on the h statistic has a similar power (tending to be, very slightly, higher) to its classical parametric version. It is also important to note that the differences in the power curves also depend on the probability distribution of the measured variable. Thus, in terms of power, the difference between the bootstrap and parametric tests of Mandel's k is greater for the skew normal and normal distributions. It was also observed that the highest power values were obtained when the data come from a skew normal distribution for the h statistic, while for the k statistic the highest power values were obtained when the data come from a skew normal distribution and the sample size is small. On the other hand, a better performance of bootstrap tests has been obtained in ILS defined with a small number of laboratories obtaining a small number of samples. Finally, increasing the number of laboratories and the number of samples generally increases the power of the tests.

Apart from the simulation study, the new bootstrap alternatives for the tests of the Mandel's h and k statistics have also been validated by applying them to two real case studies. These correspond to two ILS that evaluate procedures for the measurement of different substances in blood. The bootstrap alternatives for the Mandel h and k tests provide practically the same results as the classical parametric tests, identifying, in general, the same outlier laboratories for each of the materials studied. The bootstrap alternative for the Mandel k test tends to be slightly more powerful than the classical test, as had already been verified in the simulation study.

CRediT authorship contribution statement

Miguel Flores: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft. **Génesis Moreno:** Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft. **Cristian Solórzano:** Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft. **Salvador Naya:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. **Javier Tarrío-Saavedra:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

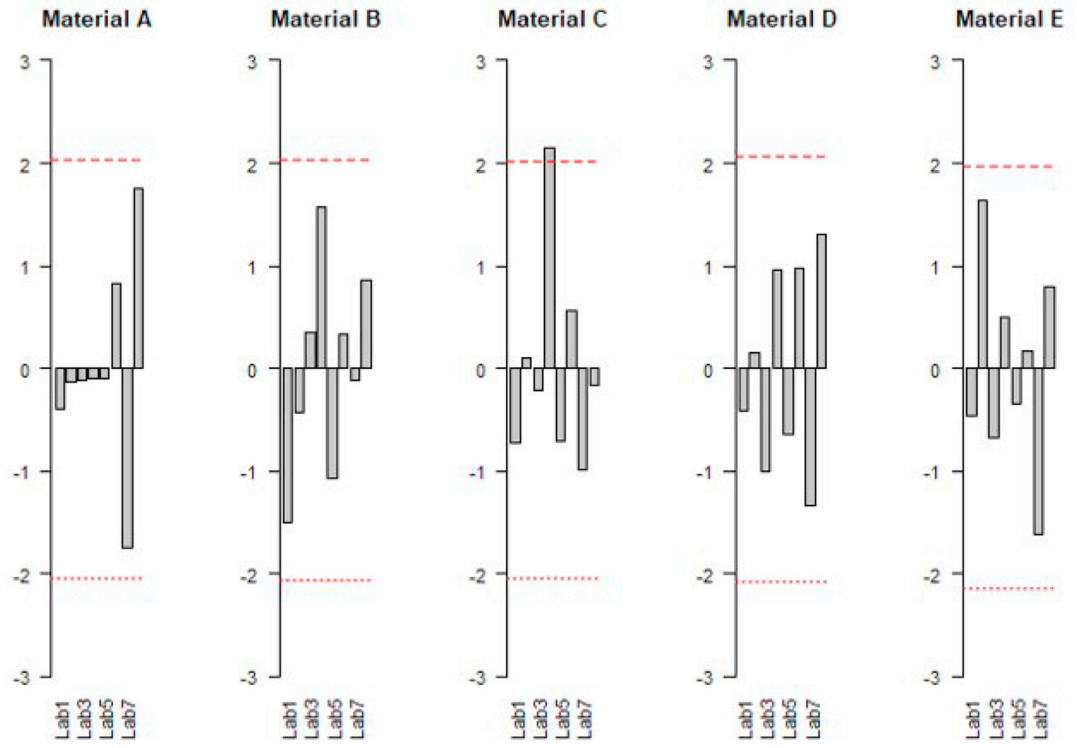

Fig. 10. Bar chart for Mandel's h Bootstrap.

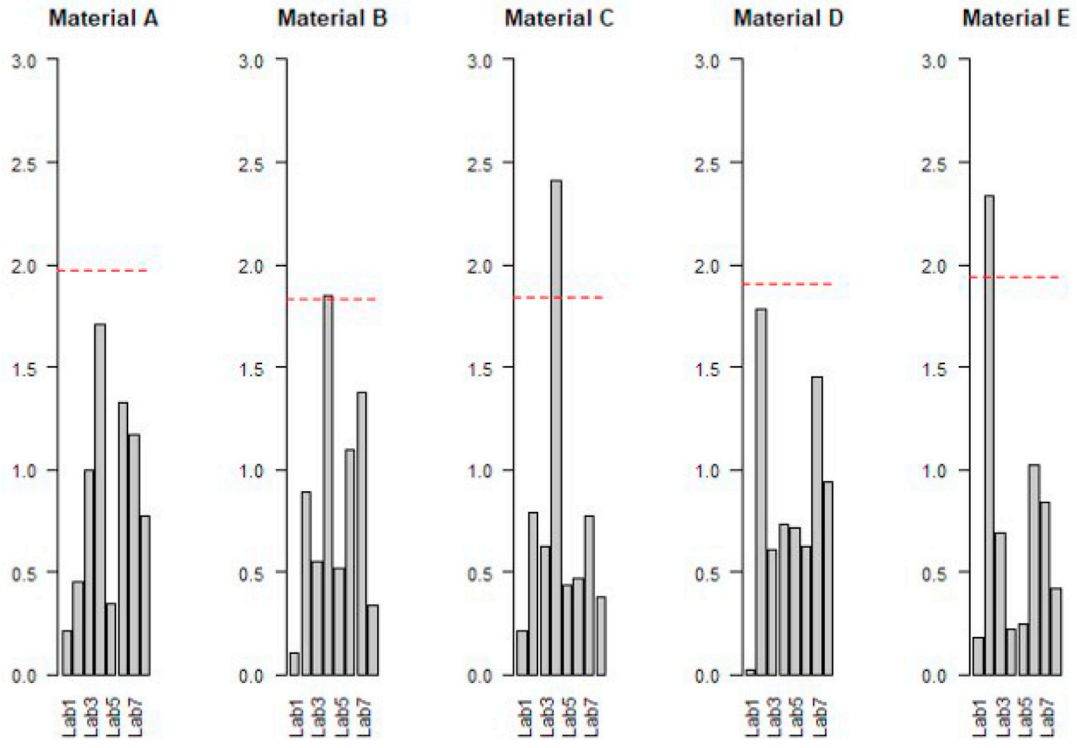

Fig. 11. Bar chart for Mandel's *k* Bootstrap.

Acknowledgements

The research has been supported by MINECO grant MTM2017-82724-R, Ministerio de Ciencia e Innovación grant PID2020-113578RB-100, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema universitario de Galicia ED431G2019/01), all of them through the ERDF.

References

[1] E. Maier, P. Quevauviller, Interlaboratory Studies and Certified Reference Materials for Environmental Analysis: the BCR Approach, Elsevier, 1999.

[2] M. Flores, J. Tarrio-Saavedra, R. Fernandez-Casal, S. Naya, Functional extensions of mandel's h and k statistics for outlier detection in interlaboratory studies, Chemometr. Intell. Lab. Syst. 176 (2018) 134–148.

[3] Y. Vander Heyden, J. Smeyers-Verbeke, Set-up and evaluation of interlaboratory studies, J. Chromatogr. A 1158 (2007) 158–167.

[4] ASTM International, Interlaboratory study program of ASTM, Last checked, https://www.astm.org/ILS/, 2021. (Accessed 15 May 2021).

[5] R.S. Kenett, E. Rahav, D.M. Steinberg, Bootstrap analysis of designed experiments, Qual. Reliab. Eng. Int. 22 (2006) 659–667.

[6] R.P. Browne, R.J. MacKay, S.H. Steiner, Two-stage leveraged measurement system assessment, Technometrics 51 (2009) 239–249.

[7] N.T. Stevens, S.H. Steiner, R.P. Browne, R.J. MacKay, Gauge r&r studies that incorporate baseline information, IIE Trans. 45 (2013) 1166–1175.

[8] N.T. Stevens, R. Browne, S.H. Steiner, R.J. MacKay, Augmented measurement system assessment, J. Qual. Technol. 42 (2010) 388–399.

[9] E. Hund, D.L. Massart, J. Smeyers-Verbeke, Inter-laboratory studies in analytical chemistry, Anal. Chim. Acta 423 (2000) 145–165.

[10] ISO 5725-2, Accuracy, (trueness and Precision) of Measurement Methods and Results-Part 2: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method, 1994.

[11] P.C. Kelly, Outlier detection in collaborative studies, J. Assoc. Off. Anal. Chem. 73 (1990) 58–64.

[12] P.L. Davies, Statistical evaluation of interlaboratory tests, Fresenius' Z. für Anal. Chem. 331 (1988) 513–519.

[13] B. Ripley, Robust statistics—How not to reject outliers. Part 1. Basic concepts, Analyst 114 (1989) 1693–1697.

[14] M. Flores, R. Fernández-Casal, S. Naya, J. Tarrío-Saavedra, R. Bossano, ILS: an R package for statistical analysis in interlaboratory studies, Chemometr. Intell. Lab. Syst. 181 (2018) 11–20.

[15] L. Xu, C. Gotwalt, Y. Hong, C.B. King, W.Q. Meeker, Applications of the fractional-random-weight bootstrap, Am. Statistician 74 (2020) 345–358.

[16] M. Flores, S. Naya, J. Tarrío-Saavedra, R. Fernández-Casal, Functional data analysis approach of mandel's h and k statistics in interlaboratory studies, in: Functional Statistics and Related Fields, Springer, 2017, pp. 123–130.

[17] S. Naya, J. Tarrío-Saavedra, J. López-Beceiro, M. Francisco-Fernández, M. Flores, R. Artiaga, Statistical functional approach for interlaboratory studies with thermal data, J. Therm. Anal. Calorim. 118 (2014) 1229–1243.

[18] T. Hesterberg, D. Moore, S. Monaghan, A. Clipson, R. Epstein, D. Moore, G. McCabe, Bootstrap methods and permutation tests, in: D. Moore, G. McCabe, W. Duckworth, L. Alwan (Eds.), The Practice of Business Statistics, WH Freeman, New York, 2008, pp. 1–85.

[19] W.C. Navidi, Statistics for Engineers and Scientists, McGraw-Hill Higher Education, New York, NY, USA, 2008.

[20] ASTM International 691-20, Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method, 14.05, Annual book of ASTM standards, 2020.

[21] L.B. Barrentine, Concepts for R&R Studies, Quality Press, 2003.

[22] R.S. Kenett, G. Shmueli, Clarifying the terminology that describes scientific reproducibility, Nat. Methods 12 (2015), 699–699.

[23] J. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, Pearson education, 2018.

[24] H.E. Plesser, Reproducibility vs. replicability: a brief history of a confused terminology, Front. Neuroinf. 11 (2018) 76.

[25] P. Patil, R.D. Peng, J.T. Leek, A statistical definition for reproducibility and replicability, BioRxiv (2016), 066803.

[26] J.D. Nichols, M.K. Oli, W.L. Kendall, G.S. Boomer, Opinion: a better approach for dealing with reproducibility and replicability in science, Proc. Natl. Acad. Sci. Unit. States Am. 118 (2021).

[27] National Academies of Sciences, Engineering, and Medicine, Reproducibility and Replicability in Science, National Academies Press, 2019.

[28] P.-T. Wilrich, Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic, AStA Adv. Stat. Analy. 97 (2013) 1–10.

[29] S.L.R. Ellison, metRology: Support for Metrological applications, 2018. https://CRAN.R-project.org/package=metRology, r package version 0.9-28-1.

[30] R. R Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020. https://www.R-project.org/.