Power of a test for assessing interlaboratory consensus of nominal and ordinal characteristics of a substance, material, or object

Power of a test for assessing interlaboratory consensus of nominal and ordinal characteristics of a substance, material, or object

OPEN ACCESS
IOP Publishing | Bureau International des Poids et Mesures Metrologia

Metrologia **61** (2024) 045004 (13pp) https://doi.org/10.1088/1681-7575/ad5846

Power of a test for assessing interlaboratory consensus of nominal and ordinal characteristics of a substance, material, or object

Tamar Gadrich[1], Yariv N Marmor[1], Francesca R Pennecchi[2], D Brynn Hibbert[3], Anastasia A Semenova[4] and Ilya Kuselman[5,*]

[1] Department of Industrial Engineering and Management, Braude College of Engineering, PO Box 78, 51 Snunit St., 2161002 Karmiel, Israel
[2] Istituto Nazionale di Ricerca Metrologica (INRIM), Strada delle Cacce 91, 10135 Turin, Italy
[3] School of Chemistry, UNSW Sydney, Sydney NSW 2052, Australia
[4] V.M. Gorbatov Federal Research Center for Food Systems, 26 Talalikhina St., 109316 Moscow, Russia
[5] Independent Consultant on Metrology, 4/6 Yarehim St., 7176419 Modiin, Israel

E-mail: ilya.kuselman@bezeqint.net

Received 8 January 2024, revised 28 April 2024
Accepted for publication 14 June 2024
Published 27 June 2024

Abstract

A concept of the consensus among different laboratories participating in an interlaboratory comparison, classifying a substance, material, or object according to its nominal and ordinal (i.e. categorical) characteristics, is devised using decomposition of the total variation of the laboratory responses. One of the components of the total variation is caused by the between-laboratory differences, and the second—by conditions associated with the applied experimental design (for example, temperature of test items, technician experience, etc). This decomposition is based on the recently developed two-way CATANOVA for nominal variables and two-way ORDANOVA for ordinal variables. The consensus is tested as hypotheses about homogeneity, i.e. insignificance of the corresponding components of the total variation. The consensus power is taken to be the power of the homogeneity test. A methodology for evaluation of the consensus power and corresponding risks of false decisions versus the dataset size of categorical characteristics obtained in an interlaboratory comparison is detailed. Examples of evaluation of the power and risks are discussed using previously-published datasets of an interlaboratory comparison of identification of weld imperfections, and an examination of the intensity of the odor of drinking water. An example of computer code in the R programming environment is presented for the power calculations in the case of nominal variables, using a chi-square distribution. A newly developed tool for ordinal variables, an Excel spreadsheet with macros, which is based on Monte Carlo draws from a multinomial distribution, is also available.

Keywords: substance, material, categorical characteristics, interlaboratory consensus, homogeneity test, power, risk

* Author to whom any correspondence should be addressed.

1. Introduction

Interlaboratory studies are widely used for evaluation of calibration and measurement capabilities of national metrology institutes and designated institutes participating in key and supplementary comparisons [1]; for estimation of proficiency/competence of calibration and testing (including chemical analytical) laboratories [2]; and for development of certified reference materials [3]. When the reference value of the measurand is unknown, agreement (consistency) of the measured values obtained by the participating laboratories is investigated [4, 5]. If suitable agreement is observed and outliers are absent or treated, the laboratory results may then be used for estimating (building) a measurand consensus value applicable instead of the unknown reference value [6–8]. The consensus value typically is: an arithmetic mean of measured values, when their associated measurement uncertainties are approximately equal; a weighted mean with weights calculated considering the measurement uncertainties; a Bayesian estimator [9, 10]; or another kind of mean. When the reported measurement uncertainties do not sufficiently cover the actual differences between laboratory results, an interlaboratory 'dark' uncertainty component, which was not considered by the laboratories but contributes to the uncertainty of the consensus value, is evaluated [11]. Then the consensus value and its associated uncertainty are applied for determination of a laboratory success [12, 13]. Another application is to assign the measurand value and its uncertainty for a candidate reference material [14].

Consensus building for datasets of measured values of the same measurand obtained in different laboratories, in different years, by different measurement methods allows evaluation of a physical constant [15] or a quantitative substance property [16]. DerSimonian and Laird method, and other statistical procedures are used for meta-analysis of such datasets, including statistical samples of small size [17]. Meta-analysis is also widely applied in medical studies [18].

However, no algebraic operations and mathematical functions can be applied to categorical characteristics of a substance, material, or object [19]. Categorical variables are nominal/qualitative or ordinal/semi-quantitative. For example, kinds of weld imperfections [20] and descriptors of water odor [21] are nominal variables, whose occurrences can be only equal or unequal, i.e. can belong to the same or different categories. However, intensity of water odor or the taste of a sausage from very bad to excellent [22, 23], which are able to be 'equal/unequal' or 'greater than'/'less than', relate to ordinal variables. Since categories may be expressed verbally, and no algebraic operations and mathematical functions exist among them, a consensus numerical value (the equivalent of a mean) in an interlaboratory comparison or meta-analysis of categorical properties cannot be formulated.

In sociology, consensus of opinions within a given group of individuals is discussed as cohesiveness or closeness, i.e. the degree to which the members of the group agree [24, 25]. For example, it may be the cohesiveness of opinions of members of a society choosing one of a few candidates for the chair of the society, or one of the alternative programs for the society's activities. Ideal consensus by this concept means a lack of dispersion of opinions or choices, while a minimal consensus corresponds to their maximal dispersion reflecting a disagreement or dissension [26–28]. Consideration of such consensus is applied in studying decision making by experts [29–31], nursing care (clinical practice) [32, 33], psychology [34] and other fields. Likert (satisfaction) scales of expert responses, similarity functions describing the distance between opinions of the experts, rank aggregation (when members of a group decide which issue is collectively preferred), and kappa coefficients interpreting a consensus as a value on the interval from 0 to 1, are used in the cited references for a consensus 'measurement'.

Consensus of responses of different laboratories participating in an interlaboratory comparison, classifying a substance, material, or object according to its nominal and ordinal characteristics, could be also interpreted as cohesiveness. The recently developed two-way factorial analysis of variation of nominal variables CATANOVA and of ordinal variables ORDANOVA, applied first in [20] and [21–23], respectively, answers the question' is a consensus among participating laboratories achieved?' The answer is based on testing hypotheses about homogeneity of the between-laboratory and within-laboratory variation components, as well as the components caused by other factors under study. This is like in two-way ANOVA for continuous quantitative variables, but the variations are calculated here from probabilities (relative frequencies) of the responses for specified categories. Similar hypotheses about the influence of different factors on the laboratory responses (and on the consensus), according to the applied experimental design and decomposition of the total variation, are tested as hypotheses on homogeneity of corresponding variations. The homogeneity testing of nominal variables in the CATANOVA framework is based on the application of a χ^2-distribution. Similar testing of ordinal variables in ORDANOVA applies empirical distributions obtained using random Monte Carlo draws from a multinomial distribution. Since in many cases the number of participating laboratories is small, not only the level of confidence (probability of Type I error or α-risk [35]) but also the power of the test (probability of Type II error or β-risk [36]), is important for a correct interpretation of the test results [37–39].

The goal of the present paper is to evaluate power of the test for assessing consensus and corresponding risks of false decisions vs. the dataset size of nominal or ordinal characteristics of a substance, material, or object, obtained in interlaboratory comparisons.

2. Statistical method

2.1. Total variation

An expert response for a given property (characteristic of a substance, material, or object) can be modelled as a random quantity $\boldsymbol{Y}$ on an ordinal scale with $K \geqslant 2$ categories

(classes or levels) characterized by a probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$, where p_k with $k = 1, 2, \ldots, K$ denotes the theoretical probability of responses related to the $k-$th category, such that $\sum_{k=1}^{K} p_k = 1$. Then, F_k denotes the cumulative theoretical probability up to the $k-$th category, i.e. $F_k = \sum_{q=1}^{k} p_q$, and $F_K = 1$. The probability P of receiving a set (vector) of responses $\boldsymbol{n} = (n_1, n_2, \ldots, n_K)$, where $n_k \geqslant 0$ denotes the number of responses related to the $k-$th category, and $\sum_{k=1}^{K} n_k = N$ is the total number of responses calculated based on the multinomial distribution with parameters $(N, \boldsymbol{p})$ as the probability mass function $P(\boldsymbol{Y} = \boldsymbol{n})$ [40]. If only two n_k are different from zero ($K = 2$), the multinomial distribution simplifies to the binomial distribution, which is applicable as to ordinal as to nominal properties.

In interlaboratory comparisons for proficiency testing and other purposes, variability in the responses of $\boldsymbol{Y}$ may be explained by independent fixed effects of two main factors (two independent categorical variables). The first factor, i.e. the variable $X1$, has I levels (I laboratories participating in the comparison), and the second factor, the variable $X2$, has J levels (e.g. J different temperatures of the water samples, distributed to the laboratories). Each of the N possible responses falls into one of the I levels of the first factor $X1$, and into one of the J levels of the second factor $X2$, so that $IJ = N$. Besides, each of the responses belongs to one of K categories of $\boldsymbol{Y}$. This is a cross balanced design without replication at any cell. No interaction between the two factors is analyzed, since only one expert response at the specified levels of the factors is examined from each laboratory as required in ISO 17 043 [2].

In practice, responses to a categorical property of an object may be correlated with the quantitative parameters, for example of the object's chemical composition [22]. Moreover, responses to different properties of the same object may be correlated between them [23]. These possible correlations are not considered further in the present work for simplicity.

Treating N responses as a statistical sample, and n_{ijk} as a random variable, then $\hat{p}_{ijk} = n_{ijk}/N$ and $\hat{F}_{ijk} = \sum_{q=1}^{k} \hat{p}_{ijq}$ denote the sample (observed) relative frequency of responses belonging to the $k-$th category and the sample cumulative relative frequency of responses up to the $k-$th category in cell (i, j), respectively. The sample total cumulative relative frequency of all responses belonging to the $k-$th category is denoted by

$$\hat{F}_{..k} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{F}_{ijk},\ k = 1, 2, \ldots, K. \tag{1}$$

Here $\hat{F}_{i.k} = \frac{1}{J} \sum_{j=1}^{J} \hat{F}_{ijk}$ $(i = 1, 2, \ldots, I;\ k = 1, 2, \ldots, K)$ and $\hat{F}_{.jk} = \frac{1}{I} \sum_{i=1}^{I} \hat{F}_{ijk}$ $(j = 1, 2, \ldots, J;\ k = 1, 2, \ldots, K)$ denote the sample total cumulative relative frequency of responses up to the $k-$th category at level i of factor $X1$ and at level j of factor $X2$, respectively. Points in a subscript symbol mean the indices of summation (for averaging) of the frequencies, e.g. i and j in $\hat{F}_{..k}$.

The observed (sample) total variation of the response variable $\boldsymbol{Y}$, normalized on the [0, 1] interval, is estimated in the two-way ORDANOVA for ordinal variables [41] as

$$\hat{V}_{\mathrm{T}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \hat{F}_{..k} \left(1 - \hat{F}_{..k}\right). \tag{2}$$

A similar estimate in the two-way CATANOVA for nominal variables [42] is

$$\hat{V}_{\mathrm{T}} = \frac{K}{(K-1)} \left(1 - \sum_{k=1}^{K} \hat{p}_{..k}^2\right), \tag{3}$$

where $\hat{p}_{..k} = n_{..k}/N$ is the sample proportion (relative frequency) of data belonging to the $k-$th category and $\sum_{k=1}^{K} \hat{p}_{..k} = 1$.

2.2. Decomposition of the total variation

In the model without replication, considered in the present paper, the total sample variation $\hat{V}_{\mathrm{T}}$ is partitioned into the between (inter)-laboratory component $\hat{C}_{\mathrm{B}}$ and the within (intra)-laboratory component $\hat{V}_{\mathrm{W}}$, caused by the second factor and/or 'residual' variation of unknown reason(s). For ordinal data [41], this is

$$\hat{V}_{\mathrm{T}} = \hat{C}_{\mathrm{B}} + \hat{V}_{\mathrm{W}}, \tag{4}$$

where

$$\hat{C}_{\mathrm{B}} = \frac{1}{(K-1)/4} \times \sum_{k=1}^{K-1} \left[\frac{1}{I} \sum_{i=1}^{I} \left(\hat{F}_{i.k} - \hat{F}_{..k}\right)^2 + \frac{1}{J} \sum_{j=1}^{J} \left(\hat{F}_{.jk} - \hat{F}_{..k}\right)^2 \right] \tag{5}$$

and

$$\hat{V}_{\mathrm{W}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\hat{F}_{i.k} + \hat{F}_{.jk} - \hat{F}_{..k}\right) \times \left(1 - \left[\hat{F}_{i.k} + \hat{F}_{.jk} - \hat{F}_{..k}\right]\right). \tag{6}$$

The individual effects of factors $X1$ and $X2$ can be estimated using the next decomposition of the variation $\hat{C}_{\mathrm{B}}$:

$$\hat{C}_{\mathrm{B}} = \hat{C}_{X1}^{\mathrm{B}} + \hat{C}_{X2}^{\mathrm{B}}, \tag{7}$$

where

$$\hat{C}_{X1}^{\mathrm{B}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \frac{1}{I} \sum_{i=1}^{I} \left(\hat{F}_{i.k} - \hat{F}_{..k}\right)^2 \text{ and}$$

$$\hat{C}_{X2}^{\mathrm{B}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \frac{1}{J} \sum_{j=1}^{J} \left(\hat{F}_{.jk} - \hat{F}_{..k}\right)^2. \tag{8}$$

A similar decomposition for nominal variables [42] leads to

$$\hat{C}^{\mathrm{B}}_{X1} = \frac{K}{K-1}\sum_{k=1}^{K}\frac{1}{I}\sum_{i=1}^{I}\left(\hat{p}_{i.k}-\hat{p}_{..k}\right)^2 \text{ and}$$
$$\hat{C}^{\mathrm{B}}_{X2} = \frac{K}{K-1}\sum_{k=1}^{K}\frac{1}{J}\sum_{j=1}^{J}\left(\hat{p}_{.jk}-\hat{p}_{..k}\right)^2. \quad (9)$$

Such decomposition may include a component related to the possible interaction between the two factors. In addition, decomposition by response categories was discussed in papers [20–22]. Note that the sample estimators by equations (2)–(9) are biased from the corresponding population variations [41, 43].

3. Power of the test and risks of false decisions vs. the dataset size

3.1. The null and alternative hypotheses

The null hypothesis H_0 of homogeneity of the responses states that the probability of classifying the responses as belonging to the k-th category does not depend on the levels of the first factor (levels i) nor on those of the second factor (levels j), i.e. $p_{ijk} = p_k$ for all $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Under this hypothesis, the following relations are applicable for both nominal and ordinal variables:

$$\frac{E\left[\hat{V}_{\mathrm{T}}\right]}{df_{\mathrm{T}}} = \frac{E\left[\hat{C}^{\mathrm{B}}_{X1}\right]}{df_{X1}} = \frac{E\left[\hat{C}^{\mathrm{B}}_{X2}\right]}{df_{X2}} = \frac{V_{\mathrm{T}}}{N}, \quad (10)$$

where E is the expected value; $df_{\mathrm{T}} = N-1$, $df_{X1} = I-1$, and $df_{X2} = J-1$, are degrees of freedom. The numerator of the last term in equation (10) is the population total variation V_{T} corresponding to the probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$. The alternative hypotheses H_1 are that one or both the studied factors influence the probability vector $\boldsymbol{p}$, i.e.

$$\frac{E\left[\hat{C}^{\mathrm{B}}_{X1}\right]}{df_{X1}} > \frac{V_{\mathrm{T}}}{N} \text{ and/or } \frac{E\left[\hat{C}^{\mathrm{B}}_{X2}\right]}{df_{X2}} > \frac{V_{\mathrm{T}}}{N}. \quad (11)$$

To test the statistical significance of both the factor effects the following significance indices (test statistics) have been defined [41]:

$$\widehat{SI}_{X1} = \frac{\hat{C}^{\mathrm{B}}_{X1}/df_{X1}}{\hat{V}_{\mathrm{T}}/df_{\mathrm{T}}} \text{ and } \widehat{SI}_{X2} = \frac{\hat{C}^{\mathrm{B}}_{X2}/df_{X2}}{\hat{V}_{\mathrm{T}}/df_{\mathrm{T}}}. \quad (12)$$

3.2. Test for nominal variables based on application of a χ^2-distribution

Distributions of the statistics $df_l\,\widehat{SI}_{Xl}$, $l = 1, 2$, for nominal variables are asymptotically approximated by the chi-square distributions $\chi^2_{df_l}$ [20] with $df_1 = (K-1)(I-1)$ and $df_2 = (K-1)(J-1)$, respectively. They have the following expectations and variances:

$$E\left[df_l\widehat{SI}_{Xl}\right] = df_l \text{ and } \mathrm{VAR}\left[df_l\widehat{SI}_{Xl}\right] = 2df_l. \quad (13)$$

This approximation allows the application of a chi-square test for testing the null and alternative hypotheses [44]. The null hypothesis H_0 regarding the equivalence of the levels of factor $X1$ ($p_{i.k} = p_k$), i.e. insignificance of the effect of factor $X1$ on the response variable $\boldsymbol{Y}$, is rejected when $df_1\,\widehat{SI}_{X1}$ exceeds the critical value x_1 of the chi-square distribution $\chi^2_{df_1}$ at the $(1-\alpha)$ 100 % level of confidence, i.e. when the probability $P(df_1\,\widehat{SI}_{X1} > x_1) = \alpha$. Similarly, the H_0 regarding the levels of factor $X2$ ($p_{.jk} = p_k$) is rejected when $df_2\,\widehat{SI}_{X2}$ exceeds the critical value x_2 of the chi-square distribution $\chi^2_{df_2}$. In a different way, the null hypothesis H_0 related to factor Xl is rejected when $\widehat{SI}_{Xl}$ exceeds x_l/df_l at the level of confidence $(1-\alpha)$ 100 %.

The alternative hypothesis H_1 by equation (11) corresponds to the shifted/modified distribution of the statistics $df_l\,\widehat{SI}_{Xl}$ which would be valid under the null hypothesis H_0. The modified distribution is denoted further as $df_l\,\widehat{SI}_{Xl,\lambda}$, where λ is the parameter of non-centrality, i.e. the shift in the distribution. The following expectations and variances related to the modified distribution are:

$$E\left[df_l\,\widehat{SI}_{Xl,\lambda}\right] = df_l + \lambda,\ \mathrm{VAR}\left[df_l\,\widehat{SI}_{Xl,\lambda}\right] = 2df_l + 4\lambda, \quad (14)$$

and

$$E\left[\widehat{SI}_{Xl,\lambda}\right] = 1 + \frac{\lambda}{df_l},\ \mathrm{VAR}\left[\widehat{SI}_{Xl,\lambda}\right] = \frac{2}{df_l} + \frac{4\lambda}{df_l{}^2}. \quad (15)$$

This modified distribution is approximated by the noncentral chi-square distribution $\chi^2_{df_l,\lambda}$ [45]. The λ values are calculated as $\lambda = w^2N$, where w is the effect of the statistical sample size for the chi-square test. A value of $w = 0.1$ is considered as a small effect, 0.3 — medium, and 0.5 — a large effect [46]. As the sample size $N = IJ$ is equal for both factors $X1$ and $X2$, the same λ is applicable.

Then, values of the power of the homogeneity test of the responses at different levels of the factor $X1$ (levels i) and factor $X2$ (levels j) can be calculated as the power of the corresponding chi-square test [47, 48]:

$$P_1 = 1-\beta_1 = 1-\mathrm{CDF}\chi^2_{df_1,\lambda}(x_1) \text{ and}$$
$$P_2 = 1-\beta_2 = 1-\mathrm{CDF}\chi^2_{df_2,\lambda}(x_2), \quad (16)$$

where CDF means cumulative distribution function and β_l denotes probability of Type II error (β-risk).

An example computer code for the power calculations in the R programming environment, is available in appendix A. The calculated results for P_1 and P_2 vs. $I = 3$ to 50 and $J = 2$ to 10 are shown as the yellow and blue transparent surfaces in figures 1 and 2, respectively. The calculations are performed at the probability of Type I error $\alpha = 0.05$ and the medium effect of the sample size: $w = 0.3$. Plots (a)–(c) in each figure correspond to $K = 3$, 5 and 10, respectively. The lower limit of the range for K in the plots, set as binary categorical cases ($K = 2$) were discussed in previous publications [43]. Bilateral interlaboratory comparisons ($I = 2$) are a specific case in metrology [13, 49] which are not considered here. The range for J started at $J = 2$ as it is usual for testing influence of a comparison

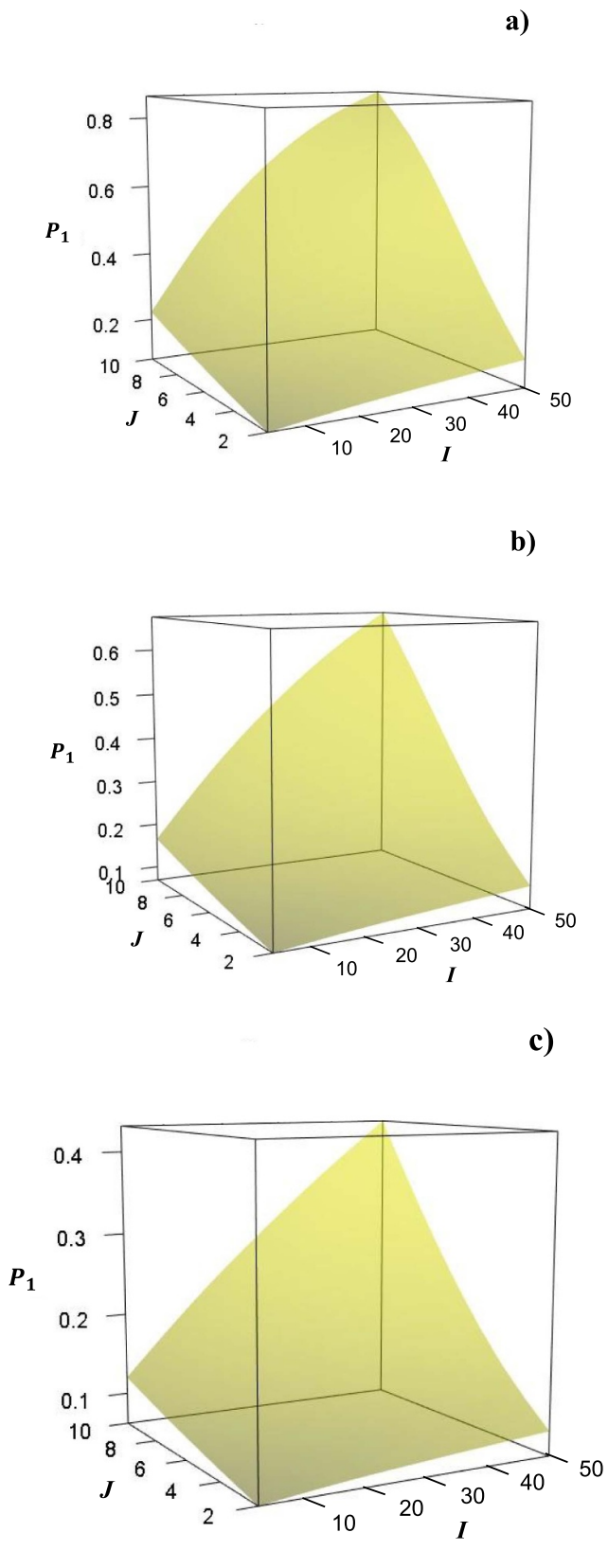

Figure 1. Power P_1 of the criterion for testing the hypothesis on significance of the effect of factor $X1$ in dependence on the number I of laboratories (levels of factor $X1$) and the number J of conditions (levels of factor $X2$). Plots (a)–(c) correspond to the number of response categories $K = 3,\ 5$ and 10, respectively.

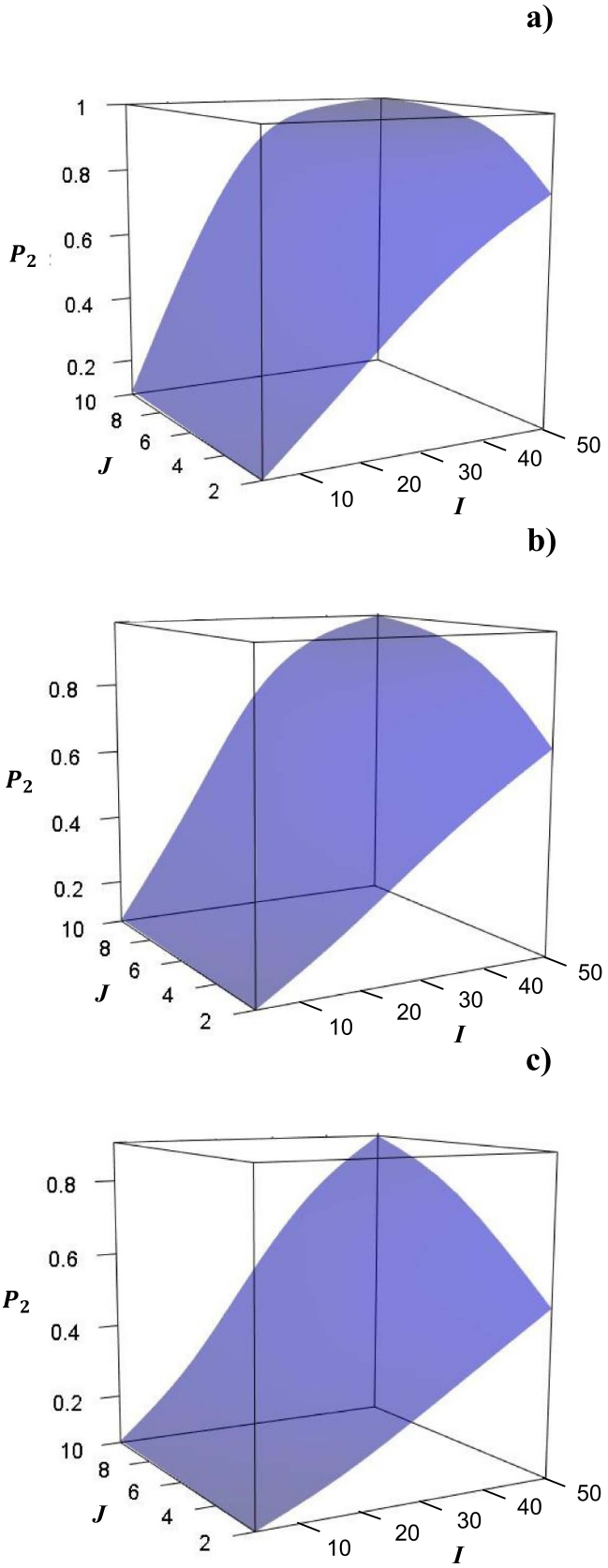

Figure 2. Power P_2 of the criterion for testing the hypothesis on significance of the effect of factor $X2$ in dependence on the number I of laboratories and the number J of conditions. Plots (a)–(c) correspond to the number of response categories $K = 3,\ 5$ and 10, respectively.

condition on the responses. Smoothing each surface plot from corresponding discrete values was performed using R [50].

Note that the axes of the plots in the figures do not start from zero, but correspond to the set ranges (from 3 for I, and from 2 for J) and the minimal calculated power values $P_l > 0$. The ranges of power P_1 values in figure 1 are from 0.08 to 0.86 for $K = 3$ in plot (a); 0.07–0.67 for $K = 5$ in plot (b); and 0.06–0.43 for $K = 10$ in plot (c). In figure 2 power P_2 values are from 0.09 to 1.00 for $K = 3$ in plot (a); 0.08–0.98 for $K = 5$ in plot (b); and 0.07–0.90 for $K = 10$ in plot (c). Thus, these figures indicate that increasing I and J, which form the statistical sample size $N = IJ$, increases the power of the test for both factors $X1$ and $X2$. Comparison of corresponding plots in figures 1 and 2 allows observing power values $P_1 < P_2$ at the same sample size N and number of categories K. This is due to variances (and corresponding standard deviations) in by equations (14) and (15), which are greater at the number of degrees of freedom $df_1 = (K-1)(I-1)$ than at $df_2 = (K-1)(J-1)$, when $I > J$. Besides, increasing K decreases the power at the same N, i.e. a larger number of categories requires a greater sample size for achieving the same power of the test. In other words, a more complex the task of a response category identification and evaluation of the influence of factors $X1$ and $X2$ on the responses requires a greater N for the task solution.

When two or more items (chemical samples, products, or objects such as photographs/images) are sent to each laboratory for examination according to the cross balanced design, there are the same number n of responses/examination results at any cell (i, j), and the total number of the results is $N = nIJ$. For example, in [20], photographs of $n = 14$ different weld features (imperfections) of $K = 5$ categories as proficiency testing items were distributed to each of the $I = 3$ participating laboratories, and examined by experienced technicians as well as by trained novices, $J = 2$. The 14 features' examination results were obtained from each experienced technician and each novice in the laboratories. The total number of photograph examinations was $N = nIJ = 84$. The vectors of the response frequencies are presented in [20]. Influence of both the factors, laboratories $X1$ and technician's experience $X2$, as well as their interaction, were found insignificant at the α-risk, i.e. the probability of Type I error, $\alpha = 0.05$.

Under hypothesis H_0 for the factor $X1$, the expectation according to equation (13) is $E[df_1\, \widehat{SI}_{X1}] = 8$, the variance is $\mathrm{VAR}[df_1 \widehat{SI}_{X1}] = 16$, and the standard deviation $sd_1 = \sqrt{\mathrm{VAR}[df_1 \widehat{SI}_{X1}]} = 4$. The distribution of $df_1\, \widehat{SI}_{X1}$ is approximated by χ^2_8. The critical value of χ^2_8 at 95 % level of confidence is $x_1 = 15.51$. Under hypothesis H_1, at the sample size $N = 84$ and the medium effect $w = 0.3$, the parameter of non-centrality of the distribution is $\lambda = w^2 N = 7.56$. By equation (14), $E[df_1\, \widehat{SI}_{X1,7.56}] = 15.56$, $VAR[df_1\, \widehat{SI}_{X1,7.56}] = 46.24$, and $sd_{1,7.56} = 6.80$. The distribution of $df_1\, \widehat{SI}_{X1,7.56}$ under H_1 is approximated by $\chi^2_{8,7.56}$. The probability density functions (PDFs) of the chi-square distributions for factor $X1$ are shown in figure 3, plot (a). On this plot, the blue line is the PDF of the chi-square distribution χ^2_8 under hypothesis H_0, and the red line is the PDF of $\chi^2_{8,7.56}$ under hypothesis H_1.

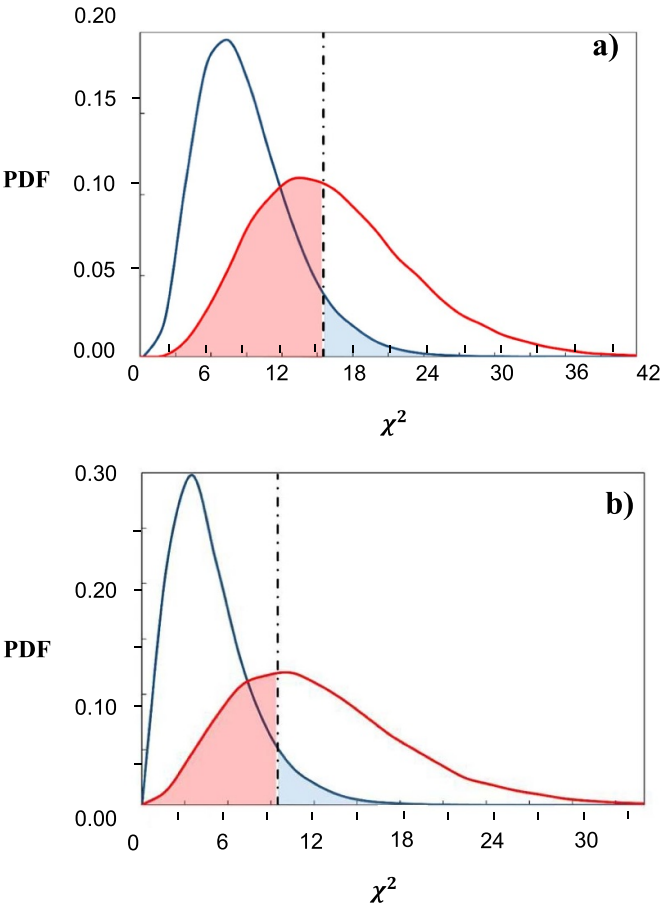

Figure 3. Probability density functions (PDFs) of the chi-square distributions. Plot (a) is for factor $X1$, and plot (b) is for factor $X2$. The blue lines demonstrate PDF of the chi-square distribution under hypothesis H_0, the red lines—under hypothesis H_1. The vertical black dashed lines indicate the critical values x_1 and x_2 for the level of confidence 95 %, for plot (a) and for plot (b), respectively. Probabilities of Type I error α are shown as the shaded area of transparent blue color, and probability of Type II error β—by the shaded area of transparent red color.

The vertical black dashed line indicates the critical value x_1. Probabilities of Type I error are shown by transparent blue area to the right of the dashed line, and of type II error — by the transparent red area to the left of the dashed line. The power of the test of insignificance of factor $X1$ is $P_1 = 0.45$

For factor $X2$ under hypothesis H_0, $E[df_2 \widehat{SI}_{X2}] = 4$, $\mathrm{VAR}[df_2 \widehat{SI}_{X2}] = 8$, and $sd_2 = 2.83$. The distribution of $df_2\, \widehat{SI}_{X2}$ is approximated by χ^2_4. The critical value of χ^2_4 at 95 % level of confidence is $x_2 = 9.49$. Under hypothesis H_1 and $\lambda = 7.56$, $E[df_2\, \widehat{SI}_{X2,7.56}] = 11.56$, $\mathrm{VAR}[df_2 \widehat{SI}_{X2,7.56}] = 38.24$, and $sd_{2,7.56} = 6.18$. The distribution of $df_2\, \widehat{SI}_{X2,7.56}$ is approximated by $\chi^2_{4,7.56}$. The corresponding PDFs of the chi-square distributions χ^2_4 and $\chi^2_{4,7.56}$ for factor $X2$ are shown on plot (b) in figure 3. The notations are the same as on plot (a) in this figure. The power of the test of insignificance of factor $X2$ is $P_2 = 0.58$.

In other words, a consensus of the laboratories in assessment of the weld imperfections, and a consensus between an experienced technician and a trained novice in a laboratory,

were accepted at the level of confidence $(1-\alpha)\ 100\ \% = 95\ \%$. Nevertheless, the probability of type II error, i.e. β-risk of a false consensus indication, was $\beta_1 = 1 - P_1 = 0.55$ concerning the laboratories, and $\beta_2 = 1 - P_2 = 0.42$ concerning the technicians.

3.3. Test for categorical variables based on a multinomial distribution and Monte Carlo simulations

Testing the null hypothesis H_0 on the effect significance for ordinal variables also requires knowledge of an asymptotical distribution for the indices $\widehat{SI}_{X1}$ and $\widehat{SI}_{X2}$ by equation (12), in order to calculate the critical values of the indices SI_{X1}^{crit} and SI_{X2}^{crit} at a given level of confidence $(1-\alpha)\ 100\ \%$. A calculation tool based on at least $3{\cdot}10^5$ Monte Carlo simulations was proposed for the two-way ORDANOVA in [41]. In case of no interaction between the two factors, the tool calculates from the empirical data the sample vector of relative frequencies $\hat{\boldsymbol{p}} = (\hat{p}_{..1}, \hat{p}_{..2}, \ldots, \hat{p}_{..K})$, as well as the variation components $(\hat{C}_{X1}^{\text{B}}, \hat{C}_{X2}^{\text{B}}, \hat{V}_{\text{W}}, \hat{V}_{\text{T}})$ and the values of the indices $\widehat{SI}_{X1}$ and $\widehat{SI}_{X2}$. At each iteration, the calculator performs random draws from the multinomial distribution with K categories and the vector of relative frequencies $\hat{\boldsymbol{p}}$, and stores the calculated values of the significance indices.

Finally, for each significance index an empirical cumulative distribution function CDF is constructed and relative frequency (%) plots of the simulated values (empirical distributions of $\widehat{SI}_{Xl}$, $l = 1, 2$) are displayed. The critical values SI_{Xl}^{crit} for the significance indices, as an equivalent of x_l/df_l for nominal variables, are recovered as the points where $(1-\alpha)\ 100\ \%$ level of confidence of the empirical CDF is achieved. The null hypothesis H_0 is rejected when the significance index $\widehat{SI}_{Xl}$ exceeds the critical value SI_{Xl}^{crit} at $(1-\alpha)\ 100\ \%$ level of confidence.

The alternative hypothesis H_1 for factor Xl is represented by the shifted/modified empirical distribution of the significance index $\widehat{SI}_{Xl}^{M} = (1 + \lambda/df_l)\,\widehat{SI}_{Xl}$. Thus, the power value P_l of the criterion for testing homogeneity of the responses at different levels of the factor Xl is $P_l = 1 - CDF_{\widehat{SI}_{Xl}^{M}}\left(SI_{Xl}^{\text{crit}}\right)$. A newly developed tool — an Excel spreadsheet with macros for these power calculations — is freely available at the webpage [51]. A brief tool description is provided in appendix B.

To validate the proposed algorithm, the analytical results of calculation for nominal variables based on the chi-square distributions discussed in section 3.2 for the study of weld imperfections were compared with those obtained with the Monte Carlo simulations. A theoretical background of such a comparison is available in appendix C. Ten repetitions of the simulations were performed to check repeatability of the power calculations. In each simulation run of $5{\cdot}10^4$ values from noncentral chi-square distributions $\chi^2_{df_l,7.56}/df_l$ were generated. The obtained power average value and the standard deviation from the average for the first factor were $P_1 = 0.4492 \pm 0.0024$, and for the second factor they were $P_2 = 0.5760 \pm 0.0018$. These power values do not practically differ from the analytical ones in section 3.2 for both factors $X1$ and $X2$ (0.45 and 0.58, respectively). The results of similar calculations with application of the Monte Carlo simulations of modified empirical distributions $\widehat{SI}_{Xl}^{M}$ by the proposed algorithm were: $P_1 = 0.4344 \pm 0.0021$ and $P_2 = 0.5107 \pm 0.0020$. The gaps between the obtained power values and the corresponding analytical ones are explained in appendix C.

The proposed algorithm using random Monte Carlo draws from a multinomial distribution was applied for example with ordinal variables for evaluation of power of the consensus of 45 laboratories participated in an interlaboratory comparison of the intensity of chlorine and sulfurous odors of different drinking water samples [21]. The laboratory responses, classified into six categories, were obtained for each water sample at 20 and 60 °C. Thus, there were: factor $X1$—laboratory with $I = 45$ levels; factor $X2$ — temperature of a water sample with $J = 2$ levels; $K = 6$ categories/levels of chlorine or sulfurous odor intensity; $n = 1$ — one response from each laboratory related to a sample of the specified odor at the specified temperature; $N = IJ = 90$ responses in total for each chlorine odor and sulfurous odor. Probability of Type I error applied was $\alpha = 0.05$. The medium size effect to be used for the power calculations assumed equal to $w = 0.3$, hence $\lambda = w^2N = 8.10$. The vectors of the response frequencies are available in the open-access paper [21].

Critical values of the significance indices are $SI_{X1}^{\text{crit}} = 1.18$ and $SI_{X2}^{\text{crit}} = 3.06$ at the level of confidence 95 %. PDF of significance index $\widehat{SI}_{Xl}$ under hypothesis H_0 and of the index $\widehat{SI}_{Xl}^{M}$ modified under hypothesis H_1 for chlorine odor of the water samples are presented in figure 4. Plot (a) is related to factor $X1$, and plot (b) — to factor $X2$. The blue line shows the PDF of $\widehat{SI}_{Xl}$, the red line is the PDF of $\widehat{SI}_{Xl}^{M}$, the black vertical dashed line indicates the critical value SI_{Xl}^{crit}. Probability of Type I error α and probability of Type II error β are shown as in figure 3. The power of the test of insignificance of factor $X1$ is $P_1 = 0.10$, and for factor $X2$ it is $P_2 = 0.29$.

For sulfurous odor, the obtained PDF of the significance indices and their critical values $SI_{X1}^{\text{crit}} = 1.20$ and $SI_{X2}^{\text{crit}} = 3.23$ were close to those for chlorine odor. Therefore, the power values $P_1 = 0.10$, and $P_2 = 0.28$ are here practically the same as for chlorine odor.

Note that decreasing the level of confidence (increasing the α-risk) leads to increasing the power (decreasing the β-risk). For example, at the level of confidence 90 % ($\alpha = 0.10$) and the effect of the statistical sample size $w = 0.3$, the power values for the intensity of chlorine odor are $P_1 = 0.17$ and $P_2 = 0.34$, and for the intensity of sulfurous odor they are $P_1 = 0.17$ and $P_2 = 0.39$. Increasing w also increases the power. Hence, at the level of confidence 90 % and $w = 0.5$, the power values for the intensity of chlorine odor are $P_1 = 0.32$ and $P_2 = 0.58$, and for the intensity of sulfurous odor — $P_1 = 0.35$ and $P_2 = 0.66$.

The standard [2] defines an interlaboratory comparison as 'design, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions'. In practice, purposes of the comparisons may be different, the number of laboratories able and ready to participate in a comparison and the number of test items may be small or large. Thus,

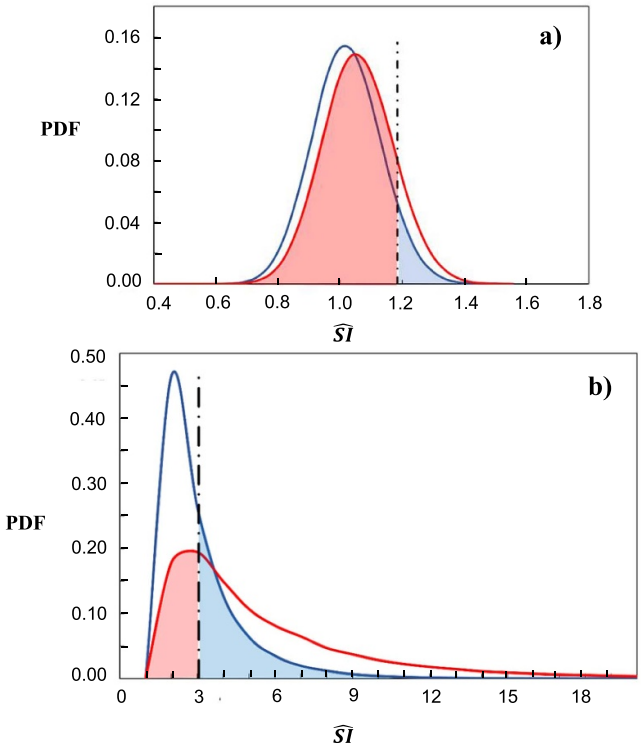

Figure 4. PDF of significance index under hypothesis H_0 and of the index modified under hypothesis H_1 for chlorine odor of the drinking water samples. Plot (a) is related to factor $X1$, and plot (b)—to factor $X2$. Blue line shows the PDF of $\widehat{SI}_{Xl}$, red line—the PDF of $\widehat{SI}^{M}_{Xl}$, black vertical dashed line indicated the critical value SI^{crit}_{Xl} at the level of confidence 95 %. Probability of Type I error α and probability of Type II error β are depicted as in figure 3.

the requirements for power of the applied tests and risks of false decisions may also be very different. In any case, it is important to have a mean for a correct evaluation of the power and the risks, as described above.

4. Conclusion

A decomposition of the total variation of the laboratory responses obtained in an interlaboratory comparison, classifying a substance, material, or object according to its nominal and ordinal characteristics, is applicable for testing a consensus of the participating laboratories. The consensus is tested as hypotheses about homogeneity, i.e. insignificance of the corresponding components of the total variation, based on the two-way analysis of variation CATANOVA for nominal variables and ORDANOVA for ordinal variables.

The consensus power is taken to be the power of the homogeneity test. For nominal variables, when a chi-square distribution is applied in the testing, increasing the number of laboratories participating in the comparison and the number of levels of a condition, at which the nominal characteristic is identified (their product is the dataset size), increases the test power. Increasing the number of categories/kinds of the studied characteristic decreases the power at the same dataset size, i.e. a larger number of categories requires greater dataset size for achieving the same power of the test.

The power evaluation for categorical (both nominal and ordinal) variables can be based on Monte Carlo draws from a multinomial distribution. This approach was validated successfully showing the power calculation results for nominal variables close to those analytical ones, obtained with application of the chi-square distribution.

Examples of evaluation of the power using earlier-published datasets of an interlaboratory comparison of identification of weld imperfections, and examination of intensity of drinking water odor supported the findings. An example computer code for the power calculations in the R programming environment for nominal variables using a chi-square distribution, and a newly developed tool for categorical variables based on Monte Carlo draws from a multinomial distribution (Excel spreadsheet with macros), are provided.

Acknowledgments

This work was supported in part by the International Union of Pure and Applied Chemistry, Project 2021-017-2-500.

Appendix A. Example computer code in R programming environment for calculation of the test power for nominal variables based on application of the chi-square distribution

```
################################################
# Power calculation (w = 0.3, K = 3, 5, 10)
# when df1 = (K-1)*(I-1)

library(DescTools)
rm(list = ls())

alpha = 0.05
w = 0.3
#K = 3, 5, 10
Ivec = 3:50
Jvec = 2:10

# N = I*J
Nvec = (Ivec[1])*(Jvec)
for (i in 2:length(Ivec))
  {Nvec = c(Nvec,(Ivec[i])*(Jvec))}
Nvec
length(Nvec) # 432 == length(Ivec)*length(Jvec)

# df1= (K-1)*(I-1)
K = 3
df1vec_K3 = c()
for (i in 1:length(Ivec))
    {df1vec_K3 = c(df1vec_K3,(K-1)*c(rep(Ivec[i]-1,length(Jvec))))}
df1vec_K3
length(df1vec_K3) # 432

Pvec_K3 = c()
for (i in 1:length(df1vec_K3))
        {Pvec_K3 = c(Pvec_K3, power.chisq.test(w = w, df = df1vec_K3[i], n = Nvec[i], sig.level = alpha)$power)}
Pvec_K3
length(Pvec_K3) # 432

# df1 = (K-1)*(I-1)
K = 5
df1vec_K5 = c()
for (i in 1:length(Ivec))
    {df1vec_K5 = c(df1vec_K5,(K-1)*c(rep(Ivec[i]-1,length(Jvec))))}
df1vec_K5
length(df1vec_K5) # 432

Pvec_K5 = c()
for (i in 1:length(df1vec_K5))
        {Pvec_K5 = c(Pvec_K5, power.chisq.test(w = w, df = df1vec_K5[i], n = Nvec[i], sig.level = alpha)$power)}
Pvec_K5
length(Pvec_K5) # 432

# df1= (K-1)*(I-1)
K = 10
df1vec_K10 = c()
for (i in 1:length(Ivec))
    {df1vec_K10 = c(df1vec_K10,(K-1)*c(rep(Ivec[i]-1,length(Jvec))))}
```

```
df1vec_K10
length(df1vec_K10) # 432

Pvec_K10 = c()
for (i in 1:length(df1vec_K10))
        {Pvec_K10 = c(Pvec_K10, power.chisq.test(w = w, df = df1vec_K10[i], n = Nvec[i], sig.level = alpha)$power)}
Pvec_K10
length(Pvec_K10) # 432

##############################################
# Power calculation (w = 0.3, K = 3, 5, 10)
# when df2 = (K-1)*(J-1)

library(DescTools)
rm(list = ls())

alpha = 0.05
w = 0.3
#K = 3, 5, 10
Ivec = 3:50
Jvec = 2:10

# N = I*J
Nvec = (Ivec[1])*(Jvec)
for (i in 2:length(Ivec))
    {Nvec = c(Nvec,(Ivec[i])*(Jvec))}
Nvec
length(Nvec) # 432 == length(Ivec)*length(Jvec)

# df2 = (K-1)*(J-1)
K = 3
df2vec_K3 = (K-1)*rep(Jvec-1,length(Ivec))
df2vec_K3
length(df2vec_K3) # 432

Pvec_K3 = c()
for (i in 1:length(df2vec_K3))
        {Pvec_K3 = c(Pvec_K3, power.chisq.test(w = w, df = df2vec_K3[i], n = Nvec[i], sig.level = alpha)$power)}
Pvec_K3
length(Pvec_K3) # 432

# df2 = (K-1)*(J-1)
K = 5
df2vec_K5 = (K-1)*rep(Jvec-1,length(Ivec))
df2vec_K5
length(df2vec_K5) # 432

Pvec_K5 = c()
for (i in 1:length(df2vec_K5))
     {Pvec_K5 = c(Pvec_K5, power.chisq.test(w = w, df = df2vec_K5[i], n = Nvec[i], sig.level = alpha)$power)}
Pvec_K5
length(Pvec_K5) # 432

# df2 = (K-1)*(J-1)
K = 10
df2vec_K10 = (K-1)*rep(Jvec-1,length(Ivec))
df2vec_K10
length(df2vec_K10) # 432
```

```
Pvec_K10= c()
for (i in 1:length(df2vec_K10))
    {Pvec_K10= c(Pvec_K10, power.chisq.test(w = w, df = df2vec_K10[i], n = Nvec[i], sig.level = alpha)$power)}
Pvec_K10
length(Pvec_K10) # 432
```

Appendix B. Tool for calculation of the test power for categorical variables based on application of the multinomial distribution and Monte Carlo simulations—a brief description

When the tool is uploaded from the webpage [51], the following parameters can be set in 'Run Program' Excel sheet of the tool:

- model type (ordinal or nominal);
- number of iterations (simulations);
- probability of Type I error α;
- effect of the sample size w.

The file-example accompanying the tool, or another file of the same structure with the dataset in *.txt format should be uploaded also. Then, the chosen model type is initiated. The calculation time with a regular PC for the proposed example of the water chlorine odor intensity is about 3–5 min.

The output includes the total variation and its components, their degrees of freedom, the significance indices, their P-value (probabilities) according to [20] for nominal variables and [41] for ordinal variables, the critical values of these indices for given level of confidence $(1-\alpha)\,100\,\%$, and the test power as described in the present paper. Excel sheets 'Program Chart PDF' and 'Program Chart CDF' present plots of the PDFs and cumulative distribution functions of the significance indices, respectively.

The tool allows also to calculate characteristics of an interaction of two factors (variables) when the experiment was designed correspondingly.

Appendix C. Comparison of the approaches to modeling the empirical distributions under an alternative hypothesis H_1 for nominal variables

To simulate an empirical distribution of $df_l\ \widehat{SI}^M_{Xl}$ $(l=1,2)$ under an alternative hypothesis H_1, the following shift/modification on $df_l\ \widehat{SI}_{Xl}$ (approximated by the chi-square distribution $\chi^2_{df_l}$) is applied:

$$df_l\ \widehat{SI}^M_{Xl} = (1+\ \lambda/df_l\ )\ df_l\ \widehat{SI}_{Xl}. \tag{C1}$$

Then, the expectation and variance of $df_l\ \widehat{SI}^M_{Xl}$ are the following:

$$\begin{aligned} E\left[df_l\ \widehat{SI}^M_{Xl}\right] &= (1+\lambda/df_l)\ E\left[df_l\ \widehat{SI}_{Xl}\right] \\ &= (1+\lambda/df_l)\ df_l = df_l+\lambda, \text{ and} \end{aligned} \tag{C2}$$

$$\begin{aligned} VAR\left[df_l\ \widehat{SI}^M_{Xl}\right] &= \left(1+\frac{\lambda}{df_l}\right)^2 VAR\left[df_l\ \widehat{SI}_{Xl}\right] \\ &= \left(1+\frac{\lambda}{df_l}\right)^2 2\ df_l = \left(\frac{df_l+\lambda}{df_l}\right)^2 2\ df_l \\ &= \left(\frac{df_l^2+2\ df_l\ \lambda+\lambda^2}{df_l^2}\right) 2\ df_l = 2\ df_l + 4\ \lambda + \frac{2\ \lambda^2}{df_l}. \end{aligned} \tag{C3}$$

Thus, while the expectations $E[df_l\ \widehat{SI}_{Xl,\lambda}]$ by equation (14) and $E[df_l\ \widehat{SI}^M_{Xl}]$ by Eq. (C2) are equal, the absolute relative difference between the variances in equation (14) and Eq. (C3) is

$$\begin{aligned} R_{\mathrm{VAR}} &= \left|\frac{\mathrm{VAR}\left[df_l\widehat{SI}^M_{Xl}\right] - VAR\left[df_l\widehat{SI}_{Xl,\lambda}\right]}{VAR\left[df_l\widehat{SI}_{Xl,\lambda}\right]}\right| \\ &= \frac{2\ df_l+4\ \lambda+\frac{2\ \lambda^2}{df_l}-(2\ df_l+4\ \lambda)}{2\ df_l+4\ \lambda} \\ &= \frac{\frac{2\ \lambda^2}{df_l}}{2\ df_l+4\ \lambda} = \frac{\frac{\lambda^2}{df_l}}{df_l+2\ \lambda} = \frac{\frac{\lambda}{df_l}}{\frac{df_l}{\lambda}+2} \\ &= \frac{\frac{1}{\theta(Xl)}}{\theta(Xl)+2} = \frac{1}{\theta(Xl)\ \ [\theta(Xl)+2]}, \end{aligned} \tag{C4}$$

where $\theta(Xl) = df_l/\lambda$. At increasing $\theta(Xl)$, the relative difference R_{VAR} tends to zero. The same is valid for the proper empirical distribution $\widehat{SI}^M_{Xl}$ (not multiplied by df_l) according to equation (15).

For example, in the discussed study of weld imperfections the value $\theta(X1) = df_1/\lambda = 8/7.56 = 1.06$, while $\theta(X2) = df_2/\lambda = 4/7.56 = 0.53$. Power values of the test versus values of the corresponding noncentral chi-square distribution $\chi^2_{df_l,7.56}/df_l$ and the modified empirical distribution $\widehat{SI}^M_{Xl}$ are presented on figure C1 by red and black lines, respectively. Plot (a) is related to factor $X1$, and plot (b) — to factor $X2$.

The gap between the power values of the consensus between laboratories ($X1$), evaluated with the $\chi^2_{df_1,7.56}/df_1$ distribution and with the empirical distribution of $\widehat{SI}_{X1}$, demonstrated with dashed lines on plot (a) at the critical value $x_1/df_1 = SI^{\mathrm{crit}}_{X1} = 1.94$ is $|100\ (0.453-0.438)/0.453| = 3.3\ \%$. The gap between the power values of the consensus related to technicians ($X2$) on plot (b) at the critical value $x_2/df_2 = SI^{\mathrm{crit}}_{X2} = 2.37$ is $|100\ (0.578-0.513)/0.578| = 11.2\ \%$.

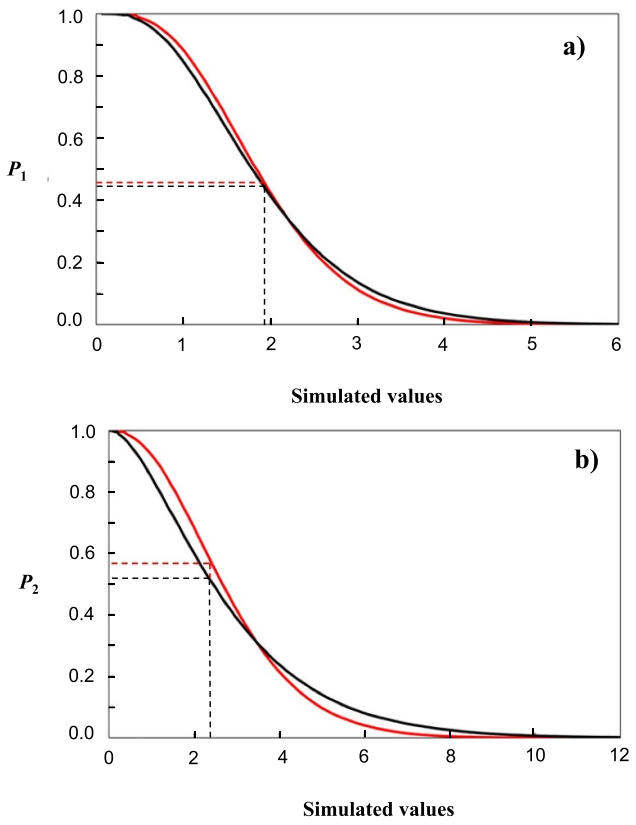

Figure C1. Power values of the test obtained with the noncentral chi-square distribution and the modified empirical distribution. Plot (a) is related to factor $X1$, and plot (b) — to factor $X2$. Power values versus the noncentral chi-square distribution values divided by numbers of degrees of freedom, $\chi^2_{df_l,7.56}/df_l$, are presented by red lines. The power values versus the modified empirical distribution of $\widehat{SI}^M_{Xl}$ values are shown by black lines. Dashed lines indicate the power for critical values of the chi-square distribution at $x_l/df_l = SI^{\text{crit}}_{Xl}$ and demonstrate corresponding gaps between the power values.

ORCID iDs

Tamar Gadrich https://orcid.org/0000-0001-6707-7510
Yariv N Marmor https://orcid.org/0000-0001-9200-5041
Francesca R Pennecchi https://orcid.org/0000-0003-1328-3858
D Brynn Hibbert https://orcid.org/0000-0001-9210-2941
Anastasia A Semenova https://orcid.org/0000-0002-4372-6448
Ilya Kuselman https://orcid.org/0000-0002-5813-9051

References

[1] The International Bureau of Weights and Measures (BIPM) *The BIPM Key Comparison Database (KCDB)* (available at: www.bipm.org/kcdb/) (Accessed 17 September 2023)
[2] The International Organization for Standardization 2023 ISO/IEC 17043 conformity assessment—general requirements for the competence of proficiency testing providers
[3] The International Organization for Standardization 2016 ISO 17034 general requirements for the competence of reference material producers
[4] CCQM Guidance note 2013 *Estimation of a Consensus KCRV and Associated Degrees of Equivalence* (available at: www.bipm.org/documents/20126/28430045/working-document-ID-5794/49d366bc-295f-18ca-c4d3-d68aa54077b5) (Accessed 17 September 2023)
[5] Ellison S L R 2022 Consistency plots: a simple graphical tool for investigating agreement in key comparisons *Accred Qual. Assur.* **27** 341–8
[6] Koepke A, Lafarge T, Possolo A and Toman B 2017 Consensus building for interlaboratory studies, key comparisons, and meta-analysis *Metrologia* **54** S34–62
[7] Possolo A 2020 Interlaboratory consensus building challenge *Anal. Bioanal. Chem.* **412** 3955–6
[8] Possolo A 2021 Solution to interlaboratory consensus building challenge *Anal. Bioanal. Chem.* **413** 3–5
[9] Tutmez B 2023 Relative uncertainty-based Bayesian interlaboratory consensus building *Sci. Total Environ.* **870** 161977
[10] Bodnar O and Bodnar T 2023 Bayesian estimation in multivariate inter-laboratory studies with unknown covariance matrices *Metrologia* **60** 054003
[11] Thompson M and Ellison S L R 2011 Dark uncertainty *Accred Qual. Assur.* **16** 483–7
[12] Thompson M 2017 A properly developed consensus from a proficiency test is, for all practical purposes, interchangeable with a certified value for a matrix reference material derived from an interlaboratory comparison *Geostand Geoanal. Res.* **42** 12195
[13] The International Organization for Standardization 2022 ISO 13528 statistical methods for use in proficiency testing by interlaboratory comparison
[14] The International Organization for Standardization 2017 ISO guide 35 reference materials—guidance for characterization and assessment of homogeneity and stability
[15] Merkatas C, Toman B, Possolo A and Schlamminger S 2019 Shades of dark uncertainty and consensus value for the Newtonian constant of gravitation *Metrologia* **56** 054001
[16] Hodges J T *et al* 2019 Recommendation of a consensus value of the ozone absorption cross-section at 253.65 nm based on a literature review *Metrologia* **56** 034001
[17] Jackson D, Bowden J and Baker R 2010 How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *J. Stat. Plan. Inference* **140** 961–70
[18] Hoffman J I E 2019 *Meta-Analysis Basic Biostatistics for Medical and Biomedical Practitioners* 2nd edn (Academic) ch 36 pp 621–9
[19] Hibbert D B (ed) 2023 *Compendium of Terminology in Analytical Chemistry* (IUPAC Orange Book 4th edn (RSC, CPI Group Ltd) ISBN:978-1-78262-947-4) (https://doi.org/10.1039/9781788012881)
[20] Gadrich T, Kuselman I and Andrić I 2020 Macroscopic examination of welds: interlaboratory comparison of nominal data *SN Appl. Sci.* **2** 2168
[21] Gadrich T, Kuselman I, Pennecchi F R, Hibbert D B, Semenova A A, Cheow P S and Naidenko V N 2022 Interlaboratory comparison of the intensity of drinking water odor and taste by two-way ordinal analysis of variation without replication *J. Water Health* **20** 1005–16
[22] Gadrich T, Pennecchi F R, Kuselman I, Hibbert D B, Semenova A A and Cheow P S 2022 Ordinal analysis of variation of sensory responses in combination with multinomial ordered logistic regression vs. chemical composition: a case study of the quality of a sausage from different producers *J. Food Qual.* **2022** 4181460

[23] Gadrich T, Pennecchi F R, Kuselman I, Hibbert D B, Semenova A A and Salikova M 2023 A novel multisensory quality index of a food product: an analysis of a sausage properties Chemometr *Intell. Lab Syst.* **237C** 104815
[24] Leik R K 1966 A measure of ordinal consensus *Pac. Soc. Rev.* **9** 85–90
[25] Keyton J and Springston J 1990 Redefining cohesiveness in groups *Small Group Res.* **21** 234–54
[26] Alcalde-Unzu1 J and Vorsatz M 2016 Do we agree? Measuring the cohesiveness of preferences *Theory Decis.* **80** 313–39
[27] Tastle W J, Wierman M J and Dumdum U R 2005 Ranking ordinal scales using the consensus measure *Issues Inf. Syst.* **6** 96–102
[28] Tastle W J and Wierman M J 2007 Consensus and dissention: a measure of ordinal dispersion *Int. J. Approx Reason.* **45** 531–45
[29] Chiclana F, Tapia García J M, Del Moral M J and Herrera-Viedma E 2013 A statistical comparative study of different similarity measures of consensus in group decision making *J. Inf. Sci.* **221** 110–23
[30] Colley R, Grandi U, Hidalgo C, Macedo M and Navarrete C 2023 Measuring and controlling divisiveness in rank aggregation *Proc. 32nd Int. Joint Conf. on Artificial Intelligence Main Track* pp 2616–23
[31] Perez I J, Cabrerizo F J, Alonso S and Herrera-Viedma E 2014 A new consensus model for group decision making problems with non-homogeneous experts *IEEE Trans. Syst. Man Cybern. Syst.* **44** 494–8
[32] Jakobsson U and Westergren A 2005 Statistical methods for assessing agreement for ordinal data *Scand. J. Caring Sci.* **19** 427–31
[33] Vituri D W and Évora Y D M 2014 Reliability of indicators of nursing care quality: testing inter-examiner agreement and reliability *Rev. Latino-Am. Enfermagem* **22** 234–40
[34] Schnuerch M, Haaf J M, Sarafoglou A and Rouder J N 2022 Meaningful comparisons with ordinal-scale items *Collabra Psychol.* **8** 11–15
[35] Mittag H-J and Rinne H 1993 *Statistical Methods of Quality Assurance* (Charman & Hall) (ISBN 0 412 55980 3)
[36] Hollebecq L-J 2023 β-risk in proficiency testing in relation to the number of participants *Acta IMEKO* **12** 1–9
[37] Kuselman I and Fajgelj A 2010 IUPAC/CITAC Guide: selection and use of proficiency testing schemes for a limited number of participants—chemical analytical laboratories *Pure Appl. Chem.* **82** 1099–135
[38] Stepanov A V and Chunovkina A G 2023 On testing of the homogeneity of variances for two-side power distribution family *Accred Qual. Assur.* **28** 129–37
[39] Jiménez-Gamero I and Analla M 2023 The importance of type II error in hypothesis testing *Int. J. Stat. Probab.* **12** 42–48
[40] Multinomial PDF *NIST/SEMATECH e-Handbook of Statistical Methods* (available at: www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/multpdf.htm) (Accessed 17 September 2023)
[41] Gadrich T and Marmor Y N 2021 Two-way ORDANOVA: analyzing ordinal variation in a cross-balanced design *J. Stat. Plan. Inference* **215** 330–43
[42] Anderson R J and Landis J R 1980 CATANOVA for multidimensional contingency tables: nominal-scale response *Commun. Stat.—Theory Methods* **9** 1191–206
[43] Gadrich T, Bashkansky E and Kuselman I 2013 Comparison of biased and unbiased estimators of variances of qualitative and semi-quantitative results of testing *Accred Qual. Assur.* **18** 85–90
[44] NIST/SEMATECH *e-Handbook of Statistical Methods—Chi-Square test for the variance* (available at: www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm) (Accessed 29 October 2023)
[45] Owen D B 1962 *Handbook of Statistical Tables* (Addison-Wesley) pp 49–62
[46] Zaiontz C *Real Statistics Using Excel—Effect size for Chi-square test* (available at: https://real-statistics.com/chi-square-and-f-distributions/effect-size-chi-square/) (Accessed 21 October 2023)
[47] Zaiontz C *Real Statistics Using Excel—Power of Chi-square tests* (available at: https://real-statistics.com/chi-square-and-f-distributions/power-chi-square-tests/) (Accessed 21 October 2023)
[48] Cran R *package documentation—Power calculations for ChiSquared tests* (available at: https://rdrr.io/cran/DescTools/man/power.chisq.test.html) (Accessed 21 October 2023)
[49] EURAMET Guide on Comparisons 2021 EURAMET Guide No. 4, Ver. 2.0 (available at: www.euramet.org/publications-media-centre/euramet-guides) (Accessed 17 September 2023)
[50] Oconnell J *Creating surface plots using R* (available at: https://copyprogramming.com/howto/how-to-create-surface-plot-in-r) (Assessed 9 November 2023)
[51] Marmor Y N *Research Areas 6—Factor Analysis Calculator Tool for Categorical Data* (available at: https://w3.braude.ac.il/lecturer/dr-yariv-n-marmor/) (Assessed 24 April 2024)