# Interlaboratory comparison of categorical characteristics of a substance, material, or object

2025 IMEKO Joint Conference
TC8 - TC11 - TC25
Torino, Italy, September 14-17, 2025

# Interlaboratory comparison of categorical characteristics of a substance, material, or object

Ilya Kuselman[1], Tamar Gadrich[2], Francesca R. Pennecchi[3], D. Brynn Hibbert[4], Anastasia A. Semenova[5], and Angelique Botha[6]

[1] *Independent Consultant on Metrology, Israel, ilya.kuselman@bezeqint.net*
[2] *Braude College of Engineering, Israel, tamarg@braude.ac.il*
[3] *Istituto Nazionale di Ricerca Metrologica (INRIM), Italy, f.pennecchi@inrim.it*
[4] *School of Chemistry, UNSW Sydney, Australia, b.hibbert@unsw.edu.au*
[5] *V.M. Gorbatov Federal Research Center for Food Systems, Russia, a.semenova@fncps.ru*
[6] *National Metrology Institute of South Africa (NMISA), South Africa, abotha@nmisa.org*

***Abstract*** – **When a reference value for the measurand in an interlaboratory study is unknown, the laboratory results may be used to estimate/build a consensus value instead of a reference value. Since no algebraic operations and mathematical functions exist among categorical (nominal and ordinal) values, a numerical consensus value cannot be formulated. Consensus of responses of experts of different laboratories participating in an interlaboratory comparison, classifying a substance, material, or object according to its nominal and ordinal characteristics, could be interpreted as the degree to which the experts agree. Two-way factorial analysis of variation of nominal variables CATANOVA and of ordinal variables ORDANOVA answer the question 'is a consensus of participating laboratories achieved or not?' The answer is based on testing hypotheses about homogeneity of the between- and within-laboratory variation components, as well as the variation components caused by other factors under study.**

## I. INTRODUCTION

Data obtained in an interlaboratory comparison study, used for the evaluation of the proficiency/competence of calibration and testing laboratories, characterization of certified reference materials or other purposes [1, 2], may be quantitative (numerical) or categorical (non-quantitative). In Fig. 1, such data are shown as a magenta shape combining the blue color of the shapes of quantitative and the red color of the shapes of categorical data. Quantitative data include discrete and continuous data, the latter consisting of ratio and interval data. Categorical data are nominal (qualitative) or ordinal (semi-quantitative). They are expressed in words, by alphanumeric codes, barcodes, or pictograms. When a reference value for the measurand in an inter-laboratory comparison study is unknown, the laboratory results may be used to estimate/build a consensus value instead of the reference value [3, 4]. The consensus value for quantitative data typically is: the arithmetic mean of the measured values, when their distribution is approximately symmetric and associated measurement uncertainties are approximately equal; a weighted mean of the values with weights depending on their measurement uncertainties; or a robust estimator of the population mean; or another estimator.

However, no algebraic operations nor mathematical functions can be directly applied to the outcome of categorical characteristics of a substance, material, or object. For example, different kinds of weld imperfections and descriptors of water odor are nominal variables, whose occurrences can only be equal or not equal, i.e. can belong to the same or to different categories. At the same time, intensity of water odor or sausage taste from very bad to excellent relate to ordinal variables, which can be "equal/unequal" or "greater than/less than." Nominal variables are studied in identification tasks and detection (presence/absence) tasks, while ordinal variables are used for characterization of properties of a substance, material or object and its quality, e.g., in sensory analysis. A consensus numerical value (an equivalent of a mean) in an interlaboratory comparison of categorical characteristics is not applicable.

In such a case, the consensus could be interpreted as the degree to which the experts agree. The two-way factorial analysis of variation of nominal variables CATANOVA and of ordinal variables ORDANOVA answer the question 'is a consensus of participating laboratories achieved or not?' 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.

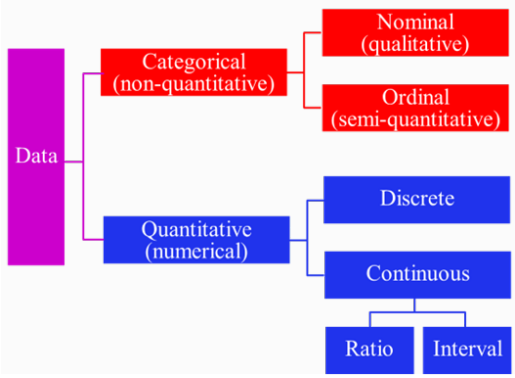


*Fig. 1. Kinds of data in an interlaboratory comparison.*

## II. DESIGN OF EXPERIMENT

### *A. Modeling responses*

An expert response for a given property (characteristic of a substance, material, or object) can be modelled as a random quantity $\boldsymbol{Y}$ on a categorical scale with $K \geq 2$ categories (classes or levels) characterized by a probability vector $\boldsymbol{p} = (p_1, p_2, \dots, p_K)$, where $p_k$ with $k = 1, 2, \dots, K$ denotes the theoretical probability of responses related to the $k$-th category, such that $\sum_{k=1}^{K} p_k = 1$. Then, for ordinal values, $F_k$ denotes the cumulative theoretical probability up to the $k$-th category, i.e., $F_k = \sum_{q=1}^{k} p_q$, and $F_K = 1$. In practice there is a set (vector) of response frequencies $\boldsymbol{n} = (n_1, n_2, \dots, n_K)$, where $n_k \geq 0$ denotes the number (frequency) of responses related to the $k$-th category, and $\sum_{k=1}^{K} n_k = N$ is the total number of responses. The probability $P$ of receiving such set of response frequencies can be evaluated based on the multinomial distribution with parameters $(N, \boldsymbol{p})$ as the probability mass function (PMF) [5]:

$$P(\boldsymbol{Y} = \boldsymbol{n}) = \frac{N!}{n_1!\, n_2! \dots n_K!} p_1^{n_1} p_2^{n_2} \dots p_K^{n_K} \qquad (1)$$

Note that the multinomial distribution, being a generalization of the binomial distribution, is applicable to nominal as well as ordinal variables.

### *B. Factors influencing the responses*

In an interlaboratory comparison study, 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 $X1$, having $I$ levels (laboratories participating in the comparison, denoted as $i$ = 1, 2, …, $I$), and the second factor, the variable $X2$, having $J$ levels (e.g., $J$ different temperatures of the water samples for examination of the water odor, denoted as $j$ =1, 2, …, $J$). 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 $X2$. Besides, each of the responses belongs to one of the $k = 1, 2, \dots, K$ categories of $\boldsymbol{Y}$.

As a rule, an interaction between such factors as a laboratory and a fixed condition of the item examination (like a temperature of a water sample) is unrealistic. Therefore, only one response at the specified levels of the factors is required in ISO/IEC 17043 [1], when an interlaboratory comparison study is used for proficiency testing of the participating laboratories. However, in a case of another simultaneous aim, e.g., checking abilities of a new trained technician vs. an experienced one (expert) for examination of the items in the same laboratory, the absence of an interaction between the factors is less obvious and may need to be tested.

### *C. Cross-balanced design*

A design of an interlaboratory comparison study without replication at any $(i, j)$ cell, when $IJ = N$, is the simplest cross-balanced design. It is shown in Table 1, where $n_{ijk}$ denotes the number of responses obtained in the $i$-th laboratory at the $j$-th condition, related to a $k$-th category. No interaction between the two factors can be analysed when all $n_{ijk} = 1$.

*Table 1. Cross-balanced design without replication.*

| Factor $X1$– laboratories | Factor $X2$ – condition levels 1 | … | $j$ | … | $J$ | Total |
|---|---|---|---|---|---|---|
| 1 | $n_{11k}$ | … | $n_{1jk}$ | … | $n_{1Jk}$ | $J$ |
| … | … | … | … | … | … | … |
| $i$ | $n_{i1k}$ | … | $n_{ijk}$ | … | $n_{iJk}$ | $J$ |
| … | … | … | … | … | … | … |
| $I$ | $n_{I1k}$ | … | $n_{Ijk}$ | … | $n_{IJk}$ | $J$ |
| Total | $I$ | … | $I$ | … | $I$ | $N$ |

In general, a cross-balanced design may contain $(i, j)$ cells with the same number $n > 1$ of replicate responses and the total number of responses $N = nIJ$. The design with replication allows testing of the interaction between the factors [6, 7].

## III. ANALYSIS OF THE RESPONSE VARIATION

### *A. Total variation*

Treating $N$ responses as a statistical sample, and the number of responses $n_{ijk}$ 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:

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

Here $\hat{F}_{i.k} = \frac{1}{J}\sum_{j=1}^{J}\hat{F}_{ijk}$ $(i = 1, 2, \dots, I;\ k = 1, 2, \dots, K)$ and $\hat{F}_{.jk} = \frac{1}{I}\sum_{i=1}^{I}\hat{F}_{ijk}$ $(j = 1, 2, \dots, J;\ k = 1, 2, \dots, K)$ denote the sample 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. Dots in a subscript symbol mean the indices of summation (for averaging) of the corresponding 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 two-way ORDANOVA for ordinal variables [7] as

$$\hat{V}_{\mathrm{T}} = \frac{1}{(K-1)/4}\sum_{k=1}^{K-1}\hat{F}_{..k}\left(1 - \hat{F}_{..k}\right) \tag{3}$$

with degrees of freedom $df_{\mathrm{T}} = N - 1$. A similar estimate in two-way CATANOVA for nominal variables [8] is

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

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

*B. Decomposition of total variation*

In the model without replication, 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 due to unknown sources(s):

$$\hat{V}_{\mathrm{T}} = \hat{C}_{\mathrm{B}} + \hat{V}_{\mathrm{W}} \tag{5}$$

The degrees of freedom of the variation components are $df_{\mathrm{B}} = (I-1) + (J-1)$ and $df_{\mathrm{W}} = (I-1)(J-1)$, respectively.

The individual effects of factors $X1$ and $X2$ can be estimated using decomposition of the component due to between-laboratory variation into the following sub-components [7]:

$$\hat{C}_{\mathrm{B}} = \hat{C}_{X1}^{\mathrm{B}} + \hat{C}_{X2}^{\mathrm{B}}, \text{ where} \tag{6}$$

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

$$\hat{C}_{X2}^{\mathrm{B}} = \frac{1}{(K-1)/4}\sum_{k=1}^{K-1}\frac{1}{J}\sum_{j=1}^{J}\left(\hat{F}_{.jk} - \hat{F}_{..k}\right)^2 \tag{7}$$

with degrees of freedom $df_{X1} = I - 1$ and $df_{X2} = J - 1$, respectively.

A similar decomposition for nominal data [8] leads to

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

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

with degrees of freedom $df_{X1} = I - 1$ and $df_{X2} = J - 1$, respectively.

Such decompositions may include a component related to the possible interaction between the two factors. In addition, decomposition by response categories can be discussed [6, 7].

*C. Null and alternative hypotheses*

The null hypothesis $H_0$ of the 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 $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$. According to this hypothesis, the following relations are applicable for both nominal and ordinal data:

$$\frac{E\left[\hat{V}_{\mathrm{T}}\right]}{df_{\mathrm{T}}} = \frac{E\left[\hat{C}_{\mathrm{B}}\right]}{df_{\mathrm{B}}} = \frac{E\left[\hat{V}_{\mathrm{W}}\right]}{df_{\mathrm{W}}} = \frac{E\left[\hat{C}_{X1}^{\mathrm{B}}\right]}{df_{X1}} = \frac{E\left[\hat{C}_{X2}^{\mathrm{B}}\right]}{df_{X2}} = \frac{V_{\mathrm{T}}}{N}, \tag{9}$$

where $E$ is the expected value of a random variable. The numerator of the last term in Eq. (9) is the population total variation $V_{\mathrm{T}}$ corresponding to the probability vector $\boldsymbol{p} = (p_1, p_2, \dots, p_K)$. The alternative hypotheses $H_1$ are that one or both the studied factors influence the probability vector $\boldsymbol{p}$, i.e.

$$\frac{E\left[\hat{C}_{X1}^{\mathrm{B}}\right]}{df_{X1}} > \frac{V_{\mathrm{T}}}{N} \quad \text{and/or} \quad \frac{E\left[\hat{C}_{X2}^{\mathrm{B}}\right]}{df_{X2}} > \frac{V_{\mathrm{T}}}{N}. \tag{10}$$

To test the statistical significance of both the factor effects, the following significance indices (test statistics) have been defined [7]:

$$\widehat{SI}_{X1} = \frac{\hat{C}_{X1}^{\mathrm{B}}/df_{X1}}{\hat{V}_{\mathrm{T}}/df_{\mathrm{T}}} \quad \text{and} \quad \widehat{SI}_{X2} = \frac{\hat{C}_{X2}^{\mathrm{B}}/df_{X2}}{\hat{V}_{\mathrm{T}}/df_{\mathrm{T}}}. \tag{11}$$

*D. Approximations with chi-square distributions*

Distributions of the statistics $df_l\,\widehat{SI}_{Xl}$, $l = 1, 2$, for nominal variables are asymptotically approximated by the chi-square distributions $\chi^2_{df_l}$ [6] with $df_1 = (K-1)(I-1)$ and $df_2 = (K-1)(J-1)$, respectively. They have the following expectations and variances:

$$E\left[df_l\,\widehat{SI}_{Xl}\right] = df_l \text{ and } VAR\left[df_l\,\widehat{SI}_{Xl}\right] = 2\,df_l. \quad (12)$$

This approximation allows testing the hypotheses by a chi-square test [5]. 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 $P\left(df_1\,\widehat{SI}_{X1} > x_1\right) = \alpha$. It is the probability of a Type I error, which may be interpreted as the $\alpha$-risk of a false decision that a consensus of the laboratories is absent, when it is actually achieved. 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 $\chi^2_{df_2}$. This also means that the null hypothesis $H_0$ related to factor $Xl$ is rejected when $\widehat{SI}_{Xl}$ exceeds $x_l/df_l$ at the $(1-\alpha)\,100\,\%$ level of confidence. The $\alpha$-risk here is the probability of a false decision of the significance of the influence of factor $Xl$ on the responses, when it is insignificant.

The alternative hypothesis $H_1$ by Eq. (10) 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 $\lambda$ is the parameter of non-centrality, i.e., the shift in the distribution. The following expectations and variances related to the modified distribution are:

$$E\left[df_l\,\widehat{SI}_{Xl,\lambda}\right] = df_l + \lambda\,,\ VAR\left[df_l\,\widehat{SI}_{Xl,\lambda}\right] = 2\,df_l + 4\,\lambda, \quad (13)$$

and

$$E\left[\widehat{SI}_{Xl,\lambda}\right] = 1 + \frac{\lambda}{df_l}\,,\ VAR\left[\widehat{SI}_{Xl,\lambda}\right] = \frac{2}{df_l} + \frac{4\,\lambda}{df_l{}^2}. \quad (14)$$

This modified distribution is approximated by the noncentral chi-square distribution $\chi^2_{df_l,\,\lambda}$. The $\lambda$ values are calculated as $\lambda = w^2N$, where $w$ is the effect of the statistical sample size on the chi-square test. Conventionally, a value of $w = 0.1$ is considered as a small effect, $w = 0.3$ – a medium effect, and $w = 0.5$ – a large effect. As the sample size $N = IJ$ is equal for both factors $X1$ and $X2$, the same $\lambda$ 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 $P_l$, $l = 1$ or 2, of the corresponding chi-square test [9]:

$$P_l = 1 - \beta_l = 1 - \mathrm{CDF}\chi^2_{df_l,\lambda}(x_l), \quad (15)$$

where CDF means cumulative distribution function, and $\beta_l$ denotes the probability of a Type II error. It may be interpreted in the case of factor $X1$ as the $\beta$-risk of a false decision of a consensus of the laboratories, when the consensus was not achieved. If the influence of factor $X2$ is tested, this is the $\beta$-risk of a false decision of the factor insignificance, when it is significant.

### *E. Approximations with multinomial distributions*

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 Eq. (11), in order to calculate the critical values of the indices $SI_{X1}^{\mathrm{crit}}$ and $SI_{X2}^{\mathrm{crit}}$ at the $(1-\alpha)\,100\,\%$ level of confidence.

A tool using random Monte Carlo (MC) draws from a multinomial distribution – an Excel spreadsheet with macros [10] – 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 $(\hat{C}_{X1}^{\mathrm{B}}, \hat{C}_{X2}^{\mathrm{B}}, \hat{V}_{\mathrm{W}}, \hat{V}_{\mathrm{T}})$ 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 CDF is constructed and relative frequency (%) plots of the simulated values (empirical distributions of $\widehat{SI}_{Xl}^{\mathrm{MC}}$, $l = 1, 2$) are displayed. The critical values $SI_{Xl}^{\mathrm{crit}}$ for the significance indices, as an equivalent of $x_l/df_l$ for nominal variables, are recovered as the points where the $(1-\alpha)\,100\,\%$ level of confidence of the empirical CDF is achieved. The null hypothesis $H_0$ is rejected when the significance index $\widehat{SI}_{Xl}^{\mathrm{MC}}$ exceeds the critical value $SI_{Xl}^{\mathrm{crit}}$ at the $(1-\alpha)\,100\,\%$ level of confidence.

The alternative hypothesis $H_1$ is represented by the shifted/modified empirical distribution of the significance index $\widehat{SI}_{Xl,\lambda}^{\mathrm{MC}} = (1 + \lambda/df_l)\widehat{SI}_{Xl}^{\mathrm{MC}}$. Thus, the power value $P_l$ of the criterion for testing homogeneity of the responses at different levels of the factor $Xl$ is $P_l = 1 - \mathrm{CDF}_{\widehat{SI}_{Xl,\lambda}^{\mathrm{MC}}}(SI_{Xl}^{crit})$.

## IV. APPLICATIONS

More details and applications of the discussed approach to interlaboratory comparisons, shown schematically in Fig. 2, are described in the IUPAC/CITAC Guide [11].

### *A. Comparison of weld imperfections*

The picture on the right in Fig. 2 is related to the

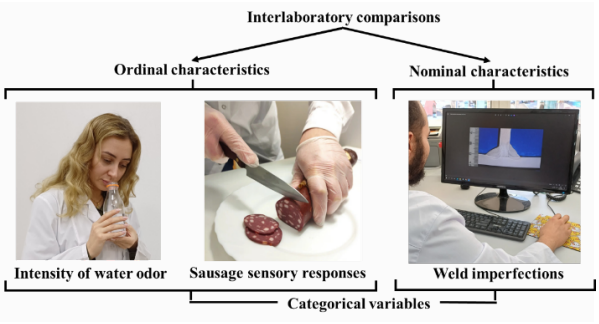


*Fig. 2. Examples of the interlaboratory comparisons.*

implementation of two-way CATANOVA with replication for a case study of nominal variables in an interlaboratory comparison of responses of technicians who categorized weld imperfections on the images for macroscopic examination. It is also an example of the evaluation of the consensus of the comparison participants when their responses are nominal values.

Three accredited laboratories participated in the comparison, $I = 3$, and were asked to recognize and classify weld imperfections according to the following five categories/classes of the weld features, $K = 5$: 1) cracks, 2) cavities, 3) inclusions, 4) lack of fusion/penetration, and 5) geometrical shape errors. Such imperfections, caused by failures in the welding process, were visible on the 12 images/macroscopic photographs of different welded joints used as the test items, sent to each participating laboratory [6]. Ten items had only one feature (imperfection) to detect, and each of the other two items had two different imperfections. Thus, $n = 14$ examination results, i.e., classes of weld imperfections by opinion of a laboratory technician – nominal characteristics of the welds, were expected from a participating laboratory. In addition, the laboratories were interested in comparing the examination results from an experienced technician and a trained novice, $J = 2$. Therefore, two datasets, each containing 14 examination results, were considered from each participating laboratory. The total number of examinations was $N = nIJ = 84$.

### *B. Comparison of the intensity of odors of drinking water*

The objective of the example demonstrated by the picture on the left in Fig. 2 was the implementation of two-way ORDANOVA without replication for a case study of ordinal variables in an interlaboratory comparison of sensory responses to the intensity of the odor of drinking water samples. The consensus of the comparison participants when their responses are ordinal values was also evaluated.

Two test items, 1 and 2, were prepared for examination of the intensity of a chlorine and a sulfurous odor, respectively. The components of these items were purchased bottled drinking water (from the same producer and batch) in a plastic container for each item, and the initial solutions of the pure reagents in glass vials: sodium hypochlorite providing a chlorine odor for test item 1, and sodium sulfide providing a sulfurous odor for test item 2. The components of items 1 and 2 were distributed to experienced ecological laboratories [12]. The examination of the items was performed at a participating laboratory immediately after preparation of the final solutions in the same conditions as for routine water samples, in six categories of the intensity for both the water odors. There were: a) imperceptible odor, b) very weak, c) weak – does not cause a disapproving response about the water, d) noticeable – causes a disapproving response, e) distinct – a tester wishes not to drink, and f) very strong – the water is not potable. To each category, the respective numeric score $k$ from 0 to 5 was assigned. The temperature of a test item was measured at 20 °C and 60 °C. Finally, 45 laboratories reported the responses. Thus, there were factor $X1$ – laboratory with $I = 45$ levels; factor $X2$ – temperature of a water sample with $J = 2$ levels ($j = 1$ at 20 °C and $j = 2$ at 60 °C); $K = 6$ categories/levels of chlorine or sulfurous odor inensity ($k$ from 0 to 5); $n = 1$ – 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.

### *C. Multinomial ordered logistic regression of sensory responses to the quality of a sausage vs. its composition*

The objective of the example represented by the picture in the middle of Fig. 2 was the implementation of two-way ORDANOVA without replication in combination with a multinomial ordered logistic regression of sensory responses to the quality of a sausage from different producers, influenced not only by variability of the testing laboratories, or their experts, but also by the chemical composition of the object under examination.

Samples of the sausage from $I = 16$ producers were purchased on the market practically simultaneously for the comparative testing of the sausage as a consumer product [13]. Its main chemical components were protein, fat, moisture, and salt. All samples were examined before their expiration dates (set by the producers) by $J = 3$ experienced assessors/experts. Five sensory quality characteristics of the samples were evaluated: 1) appearance and packaging; 2) consistency; 3) color and appearance of cut sausage; 4) taste, and 5) smell. An expert response related to each quality property was ordered by $K = 5$ categories from "very bad" to "excellent" ($k = 1, 2, \ldots, 5$). A total number $N = IJ = 48$ responses was obtained for each property, and $48 \times 5 = 240$ responses for the five properties. Contents (mass fractions expressed in %) of the $m = 4$ main components were taken from the certificates of the producers. In total $mI = 64$ continuous quantitative values were obtained.

### *D. Multisensory quality index of a product*

The IUPAC/CITAC Guide [11] also describes a method for evaluation of a quality index summarizing the responses to different properties of a commercial product. It can be useful in comparative testing of the same product from different manufacturers and for prediction models of a consumer choice. The product quality index is formulated as the negative common logarithm (– lg) of the estimate of the joint probability of the event when the product characteristics have the required/set values (categories).

For example, two-way ORDANOVA without replication was implemented for calculation of the multisensory multinomial quality index of a sausage (as in the previous example) considering possible correlation of the responses to the different quality properties of the same product [14]. However, this time the data were accumulated from each of two manufacturers during two years of production. There were $I_1 = 26$ batches $i = 1, 2, \ldots, 26$ from the first sausage manufacturer, named hereafter "producer 1", and $I_2 = 54$ batches $i = 1, 2, \ldots, 54$ from the second manufacturer, named "producer 2". Five quality sensory properties of the sausage in a batch were examined without replication at each producer's factory by its $J = 5$ experienced experts $j = 1, 2, \ldots, 5$: a) appearance and packaging; b) consistency; c) color and appearance of cut sausage; d) taste, and f) smell. An expert response related to each quality property was ordered by $K = 5$ categories $k$ from "very bad" to "excellent", $k = 1, 2, \ldots, 5$. A total of $N_1 = I_1 \times J = 130$ responses were obtained for each property, and hence $130 \times 5 = 650$ responses to the five properties of the sausage of producer 1, while for the sausage of producer 2 there were $N_2 = I_2 \times J = 270$ responses to each property and $270 \times 5 = 1350$ responses to the five properties. Contents (measured mass fractions expressed in %) of the $m = 4$ main components were taken from the batch certificates of the producer, included $I_1 \times m = 104$ quantitative values of producer 1 and $I_2 \times m = 216$ such values of producer 2, characterizing the chemical composition of the sausages.

## REFERENCES

[1] ISO/IEC 17043:2023, "Conformity assessment – General requirements for the competence of proficiency testing providers," International Organization for Standardization, Geneva, 2023.

[2] ISO 17034:2016, "General requirements for the competence of reference material producers," International Organization for Standardization, Geneva, 2016.

[3] A.Koepke, T.Lafarge, A.Possolo, B.Toman, "Consensus building for interlaboratory studies, key comparisons, and meta-analysis," Metrologia 54, S34, 2017, https://doi.org/10.1088/1681-7575/aa6c0e.

[4] ISO 13528:2022, "Statistical methods for Use in proficiency testing by interlaboratory comparison," International Organization for Standardization, Geneva, 2022.

[5] NIST/SEMATECH, "e-Handbook of Statistical Methods," https://www.itl.nist.gov/div898/handbook/.

[6] T.Gadrich, I.Kuselman, I.Andrić, "Macroscopic examination of welds: Interlaboratory comparison of nominal data," SN Appl. Sci. 2, 2168, 2020, https://doi.org/10.1007/s42452-020-03907-4.

[7] T.Gadrich, Y.N.Marmor, "Two-way ORDANOVA: Analyzing ordinal variation in a cross-balanced design," J. Stat. Plan. Inference. 215, 330, 2021, https://doi.org/10.1016/j.jspi.2021.04.005.

[8] R.J.Anderson, J.R.Landis, "CATANOVA for multidimensional contingency tables: Nominal-scale response," Comm. Statist. Theory Methods 9, 1191, 1980, https://doi.org/10.1080/03610928008827952.

[9] T.Gadrich, Y.N.Marmor, F.R.Pennecchi, D.B.Hibbert, A.A.Semenova, I.Kuselman, "Power of a test for assessing interlaboratory consensus of nominal and ordinal characteristics of a substance, material, or object," Metrologia **61**, 045004, 2024, https://doi.org/10.1088/1681-7575/ad5846.

[10] Y.N.Marmor, "Research Areas 6 – Factor Analysis Calculator Tool for Categorical Data", https://w3.braude.ac.il/lecturer/dr-yariv-n-marmor/.

[11] I.Kuselman, T.Gadrich, F.R.Pennecchi, D.B.Hibbert, A.A.Semenova, A.Botha, "IUPAC/CITAC Guide: Interlaboratory comparison of categorical characteristics of a substance, material, or object (IUPAC Technical Report)," Pure Appl. Chem., 2025.

[12] T.Gadrich, I.Kuselman, F.R.Pennecchi, D.B.Hibbert, A.A.Semenova, P.S.Cheow, V.N.Naidenko, "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, 2022, https://doi.org/10.2166/wh.2022.060.

[13] T.Gadrich, F.R.Pennecchi, I.Kuselman, D.B.Hibbert, A.A.Semenova, P.S.Cheow, "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, 2022, https://doi.org/10.1155/2022/4181460.

[14] T.Gadrich, F.R.Pennecchi, I.Kuselman, D.B.Hibbert, A.A.Semenova, M.Salikova, "A novel multisensory quality index of a food product: An analysis of a sausage properties," Chemometr. Intell. Lab. Syst. 237C, 104815, 2023, https://doi.org/10.1016/j.chemolab.2023.104815.