Macroscopic examination of welds: Interlaboratory comparison of nominal data

Tamar Gadrich¹ · Ilya Kuselman² · Ivana Andrić³

Received: 31 March 2020 / Accepted: 18 November 2020 / Published online: 7 December 2020
© Springer Nature Switzerland AG 2020

Abstract

A statistical technique is developed for interlaboratory comparisons of the macroscopic examination of weld imperfections caused by failures in the welding process. This technique allows comparisons of nominal data with more than two categories which are influenced by two factors (variables). The development is based on two-way categorical analysis of variation (CATANOVA). Decomposition of the between-laboratory variation for a cross-balanced design by these two factors (e.g., laboratory and technician experience) and their interaction, as well as by the categories of the response variable, is proposed. The developed technique was applied to an interlaboratory comparison using 12 images/photographs of welds with five categories of imperfections as the test items for macroscopic examination. The items were distributed to three participating laboratories and examined by both experienced technicians and novices. Analysis of the obtained data showed a consensus between laboratories and between technicians and no interaction between them. It was found also which categories of imperfections were more difficult to identify than others. This technique is applicable for other nominal properties, and can be adjusted for proficiency testing (PT) of laboratories participated in the comparison.

Keywords Welding failure · Macroscopic examination · Imperfections · Nominal data · Interlaboratory comparison · CATANOVA

1 Introduction

An interlaboratory comparison involves the organization, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions. This comparison is used, in particular, for PT of accredited laboratories, as well as laboratories preparing for accreditation [1]. The international standard of general requirements for PT [2] defines qualitative PT schemes as the evaluation of the performance of participating laboratories against established criteria by means of interlaboratory comparisons, where the objective is to identify or describe one or more nominal and ordinal properties or characteristics of the

test item in question. For example, the established criteria may consist in a consensus of the results obtained in the participating laboratories. The nominal property value of a phenomenon, body or substance is a word or alphanumeric code given for identification reasons, where the property has no magnitude (e.g., blood groups, colors, or weld imperfections) [3]. Nominal variables are coded by exhaustive and disjoint classes or categories with no natural ordering. Therefore, nominal data are related to categorical data [4, 5]. According to Stevens' scales of measurement [6], the only legitimate operations between any two nominal variables are equality or nonequality ($=$, $\neq$).

The statistical methods recommended for use in PT by interlaboratory comparisons [7] relate to statistical

✉ Ilya Kuselman, ilya.kuselman@bezeqint.net; Tamar Gadrich, tamarg@braude.ac.il; Ivana Andrić, ivana.andric@zit-zg.hr | ¹Department of Industrial Engineering and Management, ORT Braude College, POB 78, 2161002 Karmiel, Israel. ²Independent Consultant on Metrology, 4/6 Yarehim St., 7176419 Modiin, Israel. ³Mechanical and Metallographic Laboratory, Department of Welding Testing and Technology (ZIT Ltd.), 6 Rakitnica St., 10040 Zagreb, Croatia.

design, value assignment, performance evaluation and scoring for continuous-valued PT schemes. They are not applicable to qualitative nominal data. To date, there are no widely accepted procedures for statistical treatment of nominal data, which can lead to misunderstanding and illogical test result interpretation of nominal properties in a laboratory. This problem is recognized by international groups, such as ISO REMCO [8, 9], ISO TC69 SC 6 [10] and Eurachem/CITAC [11], which work on developing the corresponding guidelines.

The first statistical method for treatment of nominal values similar to one-way analysis of variance ANOVA for quantitative data, was most likely developed in the previous century [12] and re-labeled CATANOVA. Furthermore, CATANOVA was generalized for multidimensional contingency tables [13].

Earlier we studied the case of interlaboratory comparisons for a binary nominal property, i.e. with the number of categories $K = 2$, using one-way ordinal analysis of variation ORDANOVA, a methodology applicable to both binary nominal and ordinal (semi-quantitative) properties [14, 15]. Ordinal quantities are also related to categorical data. Ordinal data are defined as values for which a total ordering relation can be established, according to magnitude, with other quantities of the same kind, but for which no algebraic operations exist among those quantities [3]. Their legitimate operations can be “equal/unequal” and “greater/less than” ($=$, $\neq$, $>$, $<$) [6]. Examples of such relations are the Mohs hardness of minerals, octane numbers of petroleum fuels and colors of dipsticks for urine tests. One-way ORDANOVA was described thoroughly in papers [16, 17]. The study of binary properties is continued, particularly for analysis of collaborative (interlaboratory) results [18] and for PT [19]. A unifying approach for all the scales of measurement, including both nominal and ordinal (categorical) scales, one-way CATANOVA and one-way ORDANOVA, was proposed in ref. [20].

The aim of the present paper is to develop a statistical technique for interlaboratory comparisons of nominal data with $K > 2$ categories, influenced by two variables, applicable to macroscopic examination of weld imperfections caused by failures in the welding process, which could be adjusted for PT. As an example, an interlaboratory comparison of the examination results of the imperfections with $K = 5$ categories (classes according to ISO 17639 [21]) is analyzed. This comparison was organized in 2019 in Croatia by the Mechanical and Metallographic Laboratory, ZIT Ltd. [22], which used macroscopic photographs of cross-sections of different welded joints as the test items (artifacts). The same photographs were distributed simultaneously to the three participating laboratories (factor X_1) and examined visually by experienced technicians as well

as by novices (factor X_2). Important is that there is no any hierarchy of the categories and/or hierarchy of the factors.

The applied interlaboratory comparison scheme is a qualitative, simultaneous, single occasional exercise of data transformation and interpretation by ISO 17043 [2, Sect. 3.7]. Note that the scheme is possible for any number of participating laboratories equal to or more than two [2, Sect. 3.4]. However, the small number of participating laboratories leads to problems in the interpretation of quantitative results, requiring relevant mathematical and metrological solutions [7, 23]. The same is also true for nominal data.

Since the test items in the present study are not samples of a substance, material or a thing but are identical images, there is no question about their chemical or physical homogeneity. The assign value of the testing property and its uncertainty are not objectives in this study, as well as a score of a laboratory proficiency based on a deviation of the laboratory results from the assigned value, required from a PT provider [2, Sect. 4.4.1.3r] for mostly quantitative interlaboratory comparisons. The consensus only of the comparison participants examined the photographs is discussed here, and a laboratory proficiency is considered to be satisfactory when its examination results are in this consensus. Therefore, such a PT is similar to an interlaboratory comparison for evaluation of reproducibility of a test or measurement method.

2 Statistical technique for interlaboratory comparisons of nominal data with $K > 2$ categories, influenced by two factors

2.1 Description of the nominal data

Assume that the test item examination results [24] are classified according to a nominal scale with K categories (classes) for the response variable Y . The response variability is explained by the influence of two factors—random variables, X_1 and X_2 , and their possible interaction, $X_1 * X_2$. Variable X_1 indicates that I laboratories participated in the comparison, i.e., X_1 has I levels. Variable X_2 has J levels, which may be, for example, the different experience of the technicians, methods of examination used, or type of equipment. There is no any hierarchy of the K categories and/or hierarchy of the factors/variables X_1 and X_2 . When X_2 has only one level ($J = 1$), the two-way model is simplified to the form used in one-way CATANOVA. The total N examination results with K categories of Y are systemized into a cross-layout of $I \times J$ cells. Let n_{ijk} denote the number of results for the k -th category obtained at the i -th laboratory (at the i -th level of X_1) and at the j -th level of X_2 (e.g., by a technician having the j -th experience level). Thus, n_{ij}

is the number of examination results in cell (i, j) ($\sum_{k=1}^K n_{ijk} = n_{ij}$, $\sum_{i=1}^I \sum_{j=1}^J n_{ij} = N$).

Let $\hat{p}_{ijk} = n_{ijk}/n_{ij}$ denote the proportion of examination results in cell (i, j) belonging to the k -th category ($\sum_{k=1}^K \hat{p}_{ijk} = 1$). The value $n_{..k}$ denote the total number of examination results belonging to the k -th category in the comparison, and $\hat{p}_{..k} = n_{..k}/N$ represent the proportion of data belonging to the k -th category ($\sum_{k=1}^K \hat{p}_{..k} = 1$). The proportion of data in the (i, j) -th cell is nonrandom and is given by $\pi_{ij} = n_{ij}/N$, where $\sum_{i=1}^I \sum_{j=1}^J \pi_{ij} = 1$.

A detailed description of the model of the examination results and their multinomial distribution is available in the Appendix of this paper.

2.2 Analysis of the nominal data variation

The total sample variation of the response variable Y , normalized to the $[0, 1]$ interval, is defined in two-way CAT-ANOVA as

$$\hat{V}_T = \frac{K}{K-1} \left(1 - \sum_{k=1}^K \hat{p}_{..k}^2 \right). \quad (1)$$

The total sample variation $\hat{V}_T$ is partitioned in ref. [13] into the within (intra) variation $\hat{V}_W$ and the between (inter) variation $\hat{C}_B$ as follows:

$$\hat{V}_T = \hat{V}_W + \hat{C}_B, \quad (2)$$

where

$$\hat{V}_W = \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} \hat{V}_W^{ij} = \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} \frac{K}{K-1} \left(1 - \sum_{k=1}^K \hat{p}_{ijk}^2 \right) \quad (3)$$

and

$$\hat{C}_B = \frac{K}{K-1} \sum_{k=1}^K \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} (\hat{p}_{ijk} - \hat{p}_{..k})^2. \quad (4)$$

In the balanced case, when $n_{ij} = n$ and $\pi_{ij} = 1/IJ$, the within and between variations are

$$\begin{aligned} \hat{V}_W &= \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \hat{V}_W^{ij} = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \frac{K}{K-1} \left(1 - \sum_{k=1}^K \hat{p}_{ijk}^2 \right) \text{ and} \\ \hat{C}_B &= \frac{K}{K-1} \sum_{k=1}^K \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J (\hat{p}_{ijk} - \hat{p}_{..k})^2. \end{aligned} \quad (5)$$

The multiple influences of the factors X_1 and X_2 on the response variable Y are characterized by the ratio

$$R^2 = \frac{\hat{C}_B}{\hat{V}_T}, \quad 0 \leq R^2 \leq 1. \quad (6)$$

This term reflects the joint effect of the factors on Y .

2.3 Decomposition of the between-laboratory variation for a cross-balanced design

In the comparison framework, $\hat{C}_B$ characterizes the interlaboratory scattering of the test item examination results as the between-laboratory variation. To evaluate the individual effects of factors X_1 and X_2 , as well as the interaction effect $X_1 * X_2$, on the response variable Y , we suggest decomposing $\hat{C}_B$ into the following parts:

$$\hat{C}_B = \hat{C}_{X_1}^B + \hat{C}_{X_2}^B + \hat{C}_{X_1 * X_2}^B, \quad (7)$$

where

$$\begin{aligned} \hat{C}_{X_1}^B &= \frac{K}{K-1} \sum_{k=1}^K \frac{1}{I} \sum_{i=1}^I (\hat{p}_{i.k} - \hat{p}_{..k})^2 \text{ and} \\ \hat{C}_{X_2}^B &= \frac{K}{K-1} \sum_{k=1}^K \frac{1}{J} \sum_{j=1}^J (\hat{p}_{.jk} - \hat{p}_{..k})^2, \end{aligned} \quad (8)$$

while

$$\hat{C}_{X_1 * X_2}^B = \frac{K}{K-1} \sum_{k=1}^K \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J (\hat{p}_{ijk} - \hat{p}_{i.k} - \hat{p}_{.jk} + \hat{p}_{..k})^2. \quad (9)$$

The proposed decomposition allows us to evaluate all the effects separately, including the interaction effect of the factors, using the R^2 ratios of the components of the between-laboratory variation $\hat{C}_B$ by Eqs. (8) and (9) to the total variation $\hat{V}_T$ by Eq. (1):

$$R_{X_1}^2 = \frac{\hat{C}_{X_1}^B}{\hat{V}_T}, \quad R_{X_2}^2 = \frac{\hat{C}_{X_2}^B}{\hat{V}_T}, \quad \text{and} \quad R_{X_1 * X_2}^2 = \frac{\hat{C}_{X_1 * X_2}^B}{\hat{V}_T}. \quad (10)$$

Another $\hat{C}_B$ decomposition, helpful for evaluation of whether the capability of the participating laboratories to identify one category k is better or worse than their capabilities to identify other categories, consists of evaluating the following k -th parts of $\hat{C}_B$:

$$\hat{C}_B(k) = \sum_{i=1}^I \sum_{j=1}^J \frac{1}{IJ} (\hat{p}_{ijk} - \hat{p}_{..k})^2. \quad (11)$$

The greater $\hat{C}_B(k)$ is, the weaker the laboratories' ability to identify category k is. As mentioned above, when $J = 1$ (e.g., only one technician in each laboratory participates

in the examination of the test items), Eq. (11) is simplified to the form applicable to one-way CATANOVA.

2.4 Testing the null hypothesis on homogeneity of examination results

The null hypothesis of the homogeneity of examination results H_0 states that the probability of identifying a result in the (i, j) -th cell as related to the k -th category (class) does not depend on either i or j , i.e., $p_{ijk} = p_k$ for all $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$. In other words, this hypothesis states that all the laboratories participating in the comparison and their technicians are equivalent in terms of their performance regarding the test item examination. Under this hypothesis, the following relationships are correct:

$$\begin{aligned} \frac{E(\hat{V}_T)}{df_T} &= \frac{E(\hat{V}_W)}{df_W} = \frac{E(\hat{C}_B)}{df_B} = \frac{E(\hat{C}_{X1}^B)}{df_{X1}} \\ &= \frac{E(\hat{C}_{X2}^B)}{df_{X2}} = \frac{E(\hat{C}_{X1 \times X2}^B)}{df_{X1 \times X2}} = \frac{\frac{K}{K-1} \left(1 - \sum_{k=1}^K p_k^2 \right)}{N}, \end{aligned} \quad (12)$$

where E is the expected value of the random variable, $df_T = N - 1$, $df_W = N - IJ$, $df_B = IJ - 1$, $df_{X1} = I - 1$, $df_{X2} = J - 1$ and $df_{X1 \times X2} = (I - 1)(J - 1)$ are the degrees of freedom.

Testing the null hypothesis H_0 requires knowledge of at least one asymptotic distribution of the random variable, allowing us to set the test critical values at the given level of confidence. Light and Margolin [12] for one-way CATANOVA and Anderson and Landis [13] for two-way CATANOVA have shown that the following indicator can be applied for testing:

$$\hat{I} = (IJ - 1)(K - 1)\hat{S}P_B = (IJ - 1)(K - 1) \frac{\hat{C}_B/df_B}{\hat{V}_T/df_T}. \quad (13)$$

This is because the indicator distribution can be approximated asymptotically by the chi-square distribution $\chi_{(IJ-1)(K-1)}^2$ with degrees of freedom $df = (IJ - 1)(K - 1)$, while $\hat{S}P_B$ is the index of the segregation power.

Note that the approximate asymptotic chi-square distribution of the indicator follows from the multivariate normal distribution and the quadratic forms in normal variables to the multinomial distribution of the examination results. More details are available in the Appendix.

Thus, one can reject H_0 when the indicator $\hat{I}$ exceeds the critical value at the $(1 - \alpha) \cdot 100\%$ level of confidence, i.e., when $\hat{I} > \chi_{(IJ-1)(K-1)}^2(1 - \alpha)$, and conclude that the joint effect of the factors on the response variable Y is detected. In such cases, the obtained results do not support the equivalence of the examination performance by the participating laboratories or by the different technicians, or both.

In addition, we propose the following three indicators for testing the statistical significance of the factors and their interaction separately. The first indicator

$$\hat{I}_{X1} = (K - 1)(I - 1)\hat{S}P_{X1} = (K - 1)(I - 1) \frac{\hat{C}_{X1}^B/df_{X1}}{\hat{V}_T/df_T} \sim \chi_{(K-1)(I-1)}^2 \quad (14)$$

allows us to test the null hypothesis regarding the equivalence of the levels i of factor $X1$ ($p_{i,k} = p_k$), i.e., the equivalence of the examination of the test items at different laboratories when the laboratories have technicians with the same experience and the same equipment. The null hypothesis is rejected when $\hat{I}_{X1} > \chi_{(K-1)(I-1)}^2(1 - \alpha)$. In the case of rejection of the null hypothesis, the interlaboratory comparison task is to show which laboratory is significantly different from the others. If this laboratory is found and removed from the calculation of $\hat{C}_{X1}^B$ using Eq. (8), the null hypothesis is not rejected. When the number of laboratories I is large enough, more than one laboratory may have to be removed before the null hypothesis is accepted. Note that the homogeneous results of the rest of the laboratories form a consensus. When the interlaboratory comparison is applied for a PT, proficiency of a laboratory participating in this consensus is considered to be satisfactory. The question remains, however, whether the removed laboratory performed the examination more or less correctly than the rest of the laboratories. This occurs when the test items are not measurement standards and there is no metrological traceability to the International System of Units (SI). Thus, the removed laboratory is not 'bad', it is simply not a part of the consensus [25].

The second indicator,

$$\hat{I}_{X2} = (K - 1)(J - 1)\hat{S}P_{X2} = (K - 1)(J - 1) \frac{\hat{C}_{X2}^B/df_{X2}}{\hat{V}_T/df_T} \sim \chi_{(K-1)(J-1)}^2, \quad (15)$$

is helpful for testing the null hypothesis regarding the equivalence of the levels j of factor $X2$ representing the experience of the technicians or another condition in the

laboratories ($p_{jk} = p_k$). The null hypothesis is rejected when $\hat{I}_{X2} > \chi^2_{(K-1)(J-1)}(1 - \alpha)$.

The third indicator,

$$\begin{aligned}\hat{I}_{X1 \times X2} &= (K-1)(I-1)(J-1) \hat{S} \hat{P}_{X1 \times X2} \\ &= (K-1)(I-1)(J-1) \frac{\hat{C}_{X1 \times X2}^B / df_{X1 \times X2}}{\hat{V}_T / df_T} \sim \chi^2_{(K-1)(I-1)(J-1)}\end{aligned}\quad (16)$$

is for testing the null hypothesis regarding the absence of interaction between the levels i of factor $X1$ and the levels j of factor $X2$, influencing the examination of the test items in the participating laboratories ($p_{ijk} = p_k$). This null hypothesis is rejected when $\hat{I}_{X1 \times X2} > \chi^2_{(K-1)(I-1)(J-1)}(1 - \alpha)$. The rejection means that the impact of the technicians' experience or another condition on the examination results depends on the laboratories that participated in the comparison.

Note that the proposed calculations are based on the formulas requiring the simplest mathematical operations and can be performed using a routine Excel sheet.

3 Interlaboratory comparisons of macroscopic examinations of welds

3.1 Design of experiment

The three accredited laboratories that participated in the comparison (denoted $L1$, $L2$ and $L3$) were asked to recognize and classify weld imperfections according to ISO 6520-1 [26]. These imperfections, caused by failures in the welding process, were seen in 12 images/macroscopic photographs—the test items. Table 1 presents the categories/classes of the possible weld imperfections and their designations by macroscopic examination.

Note that the reference numbers in Table 1 are for labeling the imperfections by standard [26]. These numbers do not influence the definition of the obtained results as nominal (not ordinal) data.

An example from the test items is shown in Fig. 1. The photograph presents the macrostructure (magnification:

Fig. 1 An example of a test item: a photograph of the macrostructure of a fillet weld from both sides

max. $\times 10$) of a transverse cross-section of a fillet weld from both sides. The welded joint consisted of two plates with a thickness of 10 mm made from the same base material, non-alloyed structural high tensile strength steel S355J2 delivered in the normalized heat treatment condition. The joint was processed by multirun metal active gas welding with a flux core electrode BÖHLER Ti52-FD as a filler material. Before macroscopic photographs were taken for further visual examination, the specimen was subjected to grinding (500 grit) and then etched with 5% nitric acid for a few seconds [27] to ensure that any features in the weld were clearly revealed.

The same 12 test items (the same 12 macroscopic photographs of different welded joints) were sent to each participating laboratory. Ten items had only one feature (imperfection) to detect, and each of the other two items had two different features. Thus, 14 examination results—classes of weld imperfections—were expected from every participating laboratory.

Laboratory $L1$ was also interested in comparing the examination results from an experienced technician (A) and a novice (B). The participating laboratory thus provided two datasets (A and B), each containing 14 examination results.

3.2 The examination results

The results from the examination of the test items are presented in Table 2. Laboratories $L2$ and $L3$ did not provide datasets for novices (B). To demonstrate the developed statistical technique, exemplificative novice (B) data were added to Table 2 for these two laboratories.

The examination results for the test items are summarized in Table 3. In total, there were $N = 84$ results from the test item examinations, while the sample size including all technicians and laboratory results was $n_{ij} = 14$.

Table 1 Classification of features by macroscopic examination

Category/class of the weld imperfection	Feature—designation of the imperfection	Reference number of the feature
1	Cracks	100
2	Cavities	200
3	Inclusions	300
4	Lack of fusion/penetration	400
5	Geometrical shape errors	500

3.3 Discussion of the obtained results

The total sample variation of the examination results is $\hat{V}_T = 0.9524$ with $df_T = 83$ by Eq. (1); the within (intra) laboratory variation is $\hat{V}_W = 0.9056$ with $df_W = 78$, and the between- (inter) laboratory variation is $\hat{C}_B = 0.0468$ with $df_B = 5$ by Eq. (5). The ratio $R^2 = 0.0491$ by Eq. (6) indicates the joint influence of a laboratory and technician experience on the variability of the obtained results, is practically negligible. The test statistic (index of the segregation power) is $\hat{S}P_B = 0.8152$, and the indicator is $\hat{I} = 16.30$ by Eq. (13). The critical value of the chi-square distribution at the 95% level of confidence and 20 degrees of

freedom is $\chi^2_{(20)}(0.95) = 31.40$. Thus, there is no rejection of the null hypothesis of homogeneity H_0 at the 95% level of confidence: the laboratories and their technicians do not differ statistically. Additional details obtained using decomposition of the between-laboratory variation $\hat{C}_B$ by Eqs. (7)–(9) are given in Table 4, including the R^2 ratio values from Eq. (10), segregation power indices and indicators from Eqs. (14)–(16), appropriate degrees of freedom of indicators df and critical values of the chi-square distribution $\chi^2(0.95)$ at the 95% level of confidence.

Table 2 Identified classes of weld imperfections

Test item	Laboratory					
	L1		L2		L3	
	A	B	A	B	A	B
1	4	4	4	5	5	5
2	3	2	3	2	4	2
3	2; 5	2; 4	2; 5	4; 5	2; 5	4; 5
4	3; 5	1; 2	2; 5	1; 2	3; 5	1; 2
5	4	4	4	4	5	4
6	2	3	2	3	2	4
7	5	5	5	5	5	4
8	4	1	4	1	4	5
9	4	3	4	3	4	4
10	1	1	1	1	5	5
11	4	4	4	5	4	4
12	4	1	4	1	4	3

Table 3 Number of examination results by categories/classes

Class	Laboratory						Total
	L1		L2		L3		
	A	B	A	B	A	B	
1	1	4	1	4	0	1	11
2	2	3	3	2	2	2	14
3	2	2	1	2	1	1	9
4	6	4	6	2	5	6	29
5	3	1	3	4	6	4	21
Total	14	14	14	14	14	14	84

Table 4 Results of the between-laboratory variation decomposition and parameters for testing H_0 hypotheses on the statistical significance of the variation components

Factor	Variation component	R^2	$\hat{S}P$	$\hat{I}$	df	$\chi^2(0.95)$
Laboratory (X1)	0.0191	0.0201	0.8337	6.67	8	15.50
Technician experience (X2)	0.0149	0.0156	1.2969	5.19	4	9.49
X1 * X2	0.0128	0.0134	0.5558	4.45	8	15.50

All the variation components are small, but the component related to the laboratory factor is the largest. Nevertheless, the indicators in Table 4 do not exceed the critical chi-square values, and the null hypotheses are not rejected at the 95% level of confidence. Therefore, these statistical tests support the finding that the laboratories' and technicians' examination results for weld imperfections do not differ. There is also no statistically significant interaction between the 'laboratory' and 'experience of technician' as factors. In general, the obtained results show a consensus between laboratories and between technicians, and therefore also acceptable training of novices. Thus, the proficiency of the participants of the comparison can be considered as satisfactory.

Decomposition of the $\hat{C}_B$ by categories or classes using Eq. (11) leads to the values

$\hat{C}_B(1) = 0.0126$; $\hat{C}_B(2) = 0.0011$; $\hat{C}_B(3) = 0.0013$; $\hat{C}_B(4) = 0.0109$; $\hat{C}_B(5) = 0.0115$.

This means that the capability of the laboratories to identify weld imperfections is better for classes $k=2$ and 3 (cavities and inclusions, respectively) than for the rest of the classes. Cracks ($k=1$), lack of fusion ($k=4$) and geometric shape errors ($k=5$) are much more difficult to identify.

4 Conclusions

A statistical technique for interlaboratory comparisons of nominal data that are influenced by two factors (variables) was developed based on two-way CATANOVA. The technique includes decomposition of the total variation for a cross-balanced design according to the factors (e.g., laboratory and experience of technician) and their interaction, as well as according to the categories of the response variable.

Application of the developed technique was demonstrated for the interlaboratory comparison of the macroscopic examination of weld imperfections caused by failures in the welding process, with five categories. The comparison was organized using 12 photographs of the macrostructures of the welds as test items distributed to three participating laboratories and examined by experienced technicians, as well as by novices. Analysis of the obtained data showed a consensus between laboratories and between technicians and no interaction between them. Therefore, the proficiency of the participants of the comparison can be considered as satisfactory.

It was found that the weld imperfections from two categories (cavities and inclusions) were examined with low variation, while the examination results of imperfections belonging to the three other categories (cracks, lack

of fusion, and geometric shape errors) had significantly larger variations, i.e., were much more difficult to identify.

The proposed calculations are based on the formulas requiring the simplest mathematical operations and can be performed using a routine Excel sheet. The developed technique is applicable for other nominal properties, and can be adjusted for PT of laboratories participated in the comparison.

Compliance with ethical standards

Conflicts of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix: Model of the examination results and their distribution

A weld examination result is a nominal random phenomenon Y (the response variable) characterized by a probability vector $\mathbf{p}$ with K categories, i.e.: $\mathbf{p} = (p_1, p_2, \dots, p_K)$, where p_k denote the probability of data belonging to the k -th category ($\sum_{k=1}^K p_k = 1$). There are $K=5$ categories of weld imperfections listed in Table 1 ($k=1, 2, \dots, 5$). We consider the results of interlaboratory comparison of the weld examination results influenced by two independent variables (and possibly their interaction) on the nominal response variable. One of variables X_1 denotes the first factor—laboratories which participated in the comparison with $I=3$ levels, and the second variable X_2 denotes the second factor with $J=2$ levels (experienced technician versus a novice one). Assume we have N examination results with K categories of Y , each of them systemized into one of I levels of the first factor X_1 and into one of the J levels of the second factor X_2 . The number of results in the (i, j) -th cell belonging to the k -th category of Y is n_{ijk} , where the counts n_{ijk} are random. Thus, the (i, j) -th cell contains $n_{ij.} = \sum_{k=1}^K n_{ijk}$ examination results, and in total there are $\sum_{i=1}^I \sum_{j=1}^J n_{ij.} = N$ data. The proportion of data in the (i, j) -th cell is nonrandom and is given by $\pi_{ij.} = n_{ij.}/N$, where $\sum_{i=1}^I \sum_{j=1}^J \pi_{ij.} = 1$.

Let $\hat{p}_{ijk} = n_{ijk}/n_{ij.}$ denote the proportion of data belonging to the k -th category in the (i, j) -th cell ($\sum_{k=1}^K \hat{p}_{ijk} = 1$). Due to the nature of the counting procedure the random vector $(n_{ij1}, n_{ij2}, \dots, n_{ijk})$ follows the *multinomial* distribution [5] with $n_{ij.}$ and the probability vector $\mathbf{p}_{ij} = (p_{ij1}, p_{ij2}, \dots, p_{ijk})$. That is

$$P(n_{ij1}, n_{ij2}, \dots, n_{ijk}) = \frac{n_{ij.}!}{\prod_{k=1}^K n_{ijk}!} \prod_{k=1}^K p_{ijk}^{n_{ijk}}.$$

Hence, the expected value and variation of the proportion of data belonging to the k -th category in the (i, j) -th cell are

$$E(\hat{p}_{..k}) = E\left(\frac{n_{ijk}}{n_{ij.}}\right) = p_{ijk}, \quad \text{VAR}(\hat{p}_{ijk}) = \frac{p_{ijk}(1 - p_{ijk})}{n_{ij.}}.$$

The total number of examination results belonging to the k -th category is denoted by $n_{..k} = \sum_{i=1}^I \sum_{j=1}^J n_{ijk}$, and the counts $n_{..k}$ are random. The random vector $(n_{..1}, n_{..2}, \dots, n_{..K})$ follows the *multinomial* distribution with N and the probability vector $\mathbf{p} = (p_1, p_2, \dots, p_K)$. Furthermore, $\hat{p}_{..k} = n_{..k}/N$ denote the proportion of data belonging to the k -th category ($\sum_{k=1}^K \hat{p}_{..k} = 1$).

For the cross-balanced design, we assume that each of the (i, j) cells contains the same amount of examination results, that is $n_{ij.} = n$, thus the total amount of examination results equals $N = IJn$ and $\pi_{ij.} = 1/IJ$.

The null hypothesis of homogeneity of the examination results H_0 assumes that all results drawn from the same infinite population are characterized by the probability vector $\mathbf{p} = (p_1, p_2, \dots, p_K)$. This is, identifying an examination result in the (i, j) -th cell as related to the k -th category does not depend on either i or j , i.e., $p_{ijk} = p_k$ for all $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$. Therefore, under the null hypothesis, it follows that

$$E(\hat{p}_{..k}) = E\left(\frac{n_{..k}}{N}\right) = p_k, \quad \text{VAR}(\hat{p}_{..k}) = \frac{p_k(1 - p_k)}{N}.$$

Moreover, we obtain the following:

$$E(\hat{p}_{..k}^2) = \text{VAR}(\hat{p}_{..k}) + E^2(\hat{p}_{..k}) = \frac{p_k}{N} + \frac{(N-1)p_k^2}{N}$$

$$E(\hat{p}_{ijk}^2) = \text{VAR}(\hat{p}_{ijk}) + E^2(\hat{p}_{ijk}) = \frac{p_k}{N} + \frac{(N-1)p_k^2}{n}$$

Under this hypothesis, the relationships shown in Eq. (12) is justified:

$$E(\hat{V}_T) = E\left(\frac{K}{K-1} \left(1 - \sum_{k=1}^K \hat{p}_{..k}^2\right)\right)$$

$$= \frac{K}{K-1} \left(1 - \sum_{k=1}^K E(\hat{p}_{..k}^2)\right)$$

$$= df_T \frac{\frac{K}{K-1} \left(1 - \sum_{k=1}^K p_k^2\right)}{N}$$

where $df_T = N - 1$. In the same manner, the expected value of the within (intra) variation $\hat{V}_W$ is equal to:

$$E(\hat{V}_W) = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \frac{K}{K-1} \left(1 - \sum_{k=1}^K E(\hat{p}_{ijk}^2)\right)$$

$$= df_W \frac{\frac{K}{K-1} \left(1 - \sum_{k=1}^K p_k^2\right)}{N},$$

where $df_W = nIJ - IJ = N - IJ$. The expected value of the between (inter) variation $\hat{C}_B$ is equal to

$$E(\hat{C}_B) = E(\hat{V}_T) - E(\hat{V}_W) = df_B \frac{\frac{K}{K-1} \left(1 - \sum_{k=1}^K p_k^2\right)}{N},$$

where $df_B = IJ - 1$.

To calculate the expected value of the components that form the between (inter) variation $\hat{C}_B$ in Eq. (7), that is $\hat{C}_B = \hat{C}_{X1}^B + \hat{C}_{X2}^B + \hat{C}_{X1 \times X2}^B$, we set the following. Let $\hat{p}_{i.k}$ denote the proportion of examination results belonging to the k -th category of Y at level i of factor X_1 and let $\hat{p}_{.jk}$ denote the proportion of examination results belonging to the k -th category of Y at level j of factor X_2 . It follows that

$$E\left[(\hat{p}_{i.k} - \hat{p}_{..k})^2\right] = (I-1) \frac{p_k(1-p_k)}{N}$$

$$E\left[(\hat{p}_{.jk} - \hat{p}_{..k})^2\right] = (J-1) \frac{p_k(1-p_k)}{N}$$

Therefore,

$$E(\hat{C}_{X1}^B) = \frac{K}{K-1} \sum_{k=1}^K \frac{1}{I} \sum_{i=1}^I E\left[(\hat{p}_{i.k} - \hat{p}_{..k})^2\right]$$

$$= df_{X1} \frac{\frac{K}{K-1} \left(1 - \sum_{k=1}^K p_k^2\right)}{N},$$

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

$$= df_{X2} \frac{\frac{K}{K-1} \left(1 - \sum_{k=1}^K p_k^2\right)}{N}.$$

Then, $E(\hat{C}_{X1 \times X2}^B) = E(\hat{C}_B) - E(\hat{C}_{X1}^B) - E(\hat{C}_{X2}^B)$, where $df_{X1} = I - 1$, $df_{X2} = J - 1$ and $df_{X1 \times X2} = (I-1)(J-1)$.

References

1. ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories
2. ISO/IEC 17043:2010. Conformity assessment—General requirements for proficiency testing
3. JCGM 200:2012. International vocabulary of metrology—Basic and general concepts and associated terms (VIM 3)
4. Freund RJ, Wilson WJ, Mohr DL (2010) Chapter 12—Categorical data. In: Statistical methods, 3rd edn. Academic Press, Oxford, pp 633–661

5. Agresti A (2012) *Categorical data analysis*, 3rd edn. Wiley, Hoboken, NJ
6. Stevens SS (1946) On the theory of scales of measurement. *Science* 103:677–680. <http://www.jstor.org/stable/1671815>. Accessed 31 July 2020
7. ISO 13528:2015. Statistical methods for use in proficiency testing by interlaboratory comparisons
8. ISO/TR 79:2015. Reference materials—Examples of reference materials for qualitative properties
9. ISO/REMCO/WG 13. Reference materials for qualitative analysis—Testing of nominal properties. <https://www.iso.org/commitee/55002.html>. Accessed 3 September 2019
10. ISO/TC 69/SC 6. Measurement methods and results. <https://www.iso.org/committee/49808.html>. Accessed 3 Sep 2019
11. Eurachem/CITAC. Qualitative analysis WG. <https://www.eurachem.org/index.php/euwgs/wg-qa>. Accessed 31 July 2020
12. Light RJ, Margolin BH (1971) An analysis of variance for categorical data. *J Am Stat Assoc* 66:534–544. <https://doi.org/10.1080/01621459.1971.10482297>
13. Anderson RJ, Landis JR (1980) CATANOVA for multidimensional contingency tables: nominal-scale response. *Commun Stat Theory Methods* 9:1191–1206. <https://doi.org/10.1080/03610928008827952>
14. Bashkansky E, Gadrich T, Kuselman I (2012) Interlaboratory comparison of test results of an ordinal or nominal binary property: Analysis of variation. *Accred Qual Assur* 17:239–243. <https://doi.org/10.1007/s00769-011-0856-0>
15. Gadrich T, Bashkansky E, 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. <https://doi.org/10.1007/s00769-012-0939-6>
16. Gadrich T, Bashkansky E (2012) ORDANOVA: analysis of ordinal variation. *J Stat Plan Inference* 142:3174–3188. <https://doi.org/10.1016/j.jspi.2012.06.004>
17. Takeshita J, Arai Y, Ogawa M, Lu XN, Suzuki T (2019) New statistic for detecting laboratory effects in ORDANOVA. Cornell University, [arXiv:1904.06048v1](https://arxiv.org/abs/1904.06048v1)[stat.AP]. <https://arxiv.org/abs/1904.06048>. Accessed 4 Sept 2019
18. Takeshita J, Suzuki T (2020) Precision for binary measurement methods and results under beta-binomial distribution. Cornell University, [arXiv:2008.13619v1](https://arxiv.org/abs/2008.13619v1)[stat.AP]. <https://arxiv.org/abs/2008.13619>. Accessed 19 Sept 2020
19. Bashkansky E, Turetsky V (2016) Proficiency testing: binary data analysis. *Accred Qual Assur* 21:265–270. <https://doi.org/10.1007/s00769-016-1208-x>
20. Gadrich T, Bashkansky E, Zitikis R (2015) Assessing variation: a unifying approach for all scales of measurement. *Qual Quant* 49:1145–1167. <https://doi.org/10.1007/s11135-014-0040-9>
21. ISO 17639:2013. Destructive tests on welds in metallic materials. Macroscopic and microscopic examination of welds
22. Mechanical and Metallographic Laboratory, Department of Welding Testing and Technology (ZIT Ltd.). <http://www.zit-zg.hr/testing-department/mechanical-metallographic-laboratory/>. Accessed 3 Sept 2019
23. Kuselman I, Fajgelj A (2010) IUPAC/CITAC Guide: Selection and use of proficiency testing schemes for a limited number of participants—chemical analytical laboratories (IUPAC Technical Report). *Pure Appl Chem* 82:1099–1135. <https://doi.org/10.1351/PACREP-09-08-15>
24. Nordin G, Dybkaer R, Forsum U, Fuentes-Arderiu X, Pontet F (2018) IUPAC Recommendations Vocabulary on nominal property, examination and related concepts for clinical laboratory sciences (IFCC-IUPAC Recommendations 2017). *Pure Appl Chem* 90:913–935. <https://doi.org/10.1515/pac-2011-0613>
25. Kuselman I, Fajgelj A (2011) Key metrological issues in proficiency testing—response to “Metrological comparability—a key issue in further accreditation” by K. Heydorn. *Accred Qual Assur* 16:99–102. <https://doi.org/10.1007/s00769-010-0744-z>
26. ISO 6520-1:2007. Welding and allied processes—Classification of geometric imperfections in metallic materials, Part 1—Fusion welding
27. ISO/TR 16060:2003. Destructive tests of welds in metallic materials—Etchants for macroscopic and microscopic examination

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.