IUPAC/CITAC guide: interlaboratory comparison of categorical characteristics of a substance, material, or object (IUPAC Technical Report)

IUPAC Technical Report

Ilya Kuselman*, Tamar Gadrich, Francesca R. Pennecchi, D. Brynn Hibbert, Anastasia A. Semenova and Angelique Botha

IUPAC/CITAC guide: interlaboratory comparison of categorical characteristics of a substance, material, or object (IUPAC Technical Report)

https://doi.org/10.1515/pac-2025-0408
Received January 2, 2025; accepted April 24, 2025

Abstract: This Guide is intended for harmonization of interlaboratory comparisons of categorical – nominal (qualitative, i.e., non-quantitative) and ordinal (semi-quantitative) – characteristics of a substance, material, or object. It provides guidance for application of relevant methods of mathematical statistics for design of such interlaboratory comparisons and analysis of the obtained data, when the methods developed for continuous quantitative values (e.g., ANOVA – analysis of variance) cannot be used without violation of their basic assumptions. The proposed approach employs recently-developed two-way nominal analysis of variation CATANOVA and two-way ordinal analysis of variation ORDANOVA. The Guide also addresses correlation between the categorical characteristics, as well as correlation between these characteristics and the chemical composition of the material or object. A multisensory quality index of a product, combining information on its categorical characteristics, is detailed. It allows for comparison of the quality of the same material produced by different producers. The examples provided in the Guide are from the fields of macroscopic examination of weld imperfections, comparison of odor intensity of drinking water, and comparison of sensory (ordinal) characteristics of a sausage. A corresponding calculation tool with an Excel spreadsheet including macros, and programs written in the R environment, are available in the specified references.

Keywords: CATANOVA; chemical composition; correlation; interlaboratory comparison; material; nominal characteristics; object; ORDANOVA; ordinal characteristics; quality index.

Article note: Sponsoring body: IUPAC Analytical Chemistry Division: see more details on page 32. This manuscript was prepared in the framework of IUPAC projects 2021-017-2-500 and 2023-016-1-500.

***Corresponding author: Ilya Kuselman**, Independent Consultant on Metrology, 4/6 Yarehim St., 7176419 Modiin, Israel, e-mail: ilya.kuselman@bezeqint.net. https://orcid.org/0000-0002-5813-9051
Tamar Gadrich, Department of Industrial Engineering and Management, Braude College of Engineering, 51 Snunit St., 2161002 Karmiel, Israel, e-mail: tamarg@braude.ac.il. https://orcid.org/0000-0001-6707-7510
Francesca R. Pennecchi, Istituto Nazionale di Ricerca Metrologica (INRIM), Strada delle Cacce 91, 10135 Turin, Italy, e-mail: f.pennecchi@inrim.it. https://orcid.org/0000-0003-1328-3858
D. Brynn Hibbert, School of Chemistry, UNSW Sydney, Sydney, NSW 2052, Australia, e-mail: b.hibbert@unsw.edu.au. https://orcid.org/0000-0001-9210-2941
Anastasia A. Semenova, V.M. Gorbatov Federal Research Center for Food Systems, 26 Talalikhina St., 109316 Moscow, Russia, e-mail: semmm@mail.ru. https://orcid.org/0000-0002-4372-6448
Angelique Botha, National Metrology Institute of South Africa (NMISA), Private Bag X34, Lynnwood Ridge 0040, Pretoria, South Africa, e-mail: abotha@nmisa.org. https://orcid.org/0000-0003-3987-359X

CONTENTS

1 Introduction

Interlaboratory studies are widely used for estimation of proficiency/competence of calibration and testing (including chemical analytical and medical) laboratories in specific measurements, tests, calibrations, examinations, inspections or sampling;[1] for development of certified reference materials;[2] for validation of

measurement and testing methods;[3,4] and for evaluation of calibration and measurement capabilities of national metrology institutes and designated institutes participating in key and supplementary comparisons.[5] When the reference value for the measurand is unknown, agreement (consistency) of the measured values obtained by the participating laboratories is investigated.[6,7] If suitable agreement is observed and outliers are absent or treated, the laboratory results may then be used for estimating (building) a consensus value of the measurand that is applicable in the absence of a known reference value.[8–10] The consensus value typically is an arithmetic mean of 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; a Bayesian estimator;[11,12] or a robust estimator of a population mean.[1,13] When the reported measurement uncertainties do not sufficiently cover the actual differences between laboratory results, an interlaboratory "dark" uncertainty component, not considered by the laboratories but contributing to the uncertainty of the consensus value, is evaluated.[14] Then the consensus value and its associated uncertainty are applied for determination of a laboratory's success. Another application is to assign the measurand value and its uncertainty for a candidate reference material.[15,16] In a method validation, the consensus value is used for evaluation of the method reproducibility.[17]

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[18] or a quantitative substance property.[19] The method of DerSimonian and Laird, and other statistical procedures, are used for meta-analysis of such datasets, including statistical samples of small size.[20] Meta-analysis is also widely applied in medical studies.[21]

However, no algebraic operations and mathematical functions can be directly applied to the outcome of categorical characteristics of a substance, material, or object, whether the categories are expressed in words, by alphanumeric codes, barcodes, or pictograms.[22–24] Categorical variables are nominal/qualitative (non-quantitative) or ordinal/semi-quantitative. For example, kinds of weld imperfections[25] and descriptors of water odor[26] are nominal variables, whose occurrences can be only equal or unequal, i.e., can belong to the same or different categories. At the same time, intensity of water odor or sausage taste from very bad to excellent [27,28] relate to ordinal variables, which are able to be "equal/unequal" or "greater than/less than". Nominal variables are studied in identification tasks and detection (presence/absence) tasks,[29,30] while ordinal variables are used for characterization of properties of a substance, material, or object and its quality, e.g., in sensory analysis.[31,32] Such variables are also applied for modeling in psychology, clinical, and social sciences.[33] A consensus numerical value (an equivalent of a mean) in an interlaboratory comparison or meta-analysis of categorical properties is not applicable. Statistical techniques for interlaboratory comparisons of nominal and ordinal properties of a substance, material, or object are less studied and not harmonized.[34,35]

In sociology, consensus of opinions within a given group of individuals is described as "cohesiveness" or "closeness", i.e., the degree to which the members of the group agree.[36,37] 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 dissent.[38–40] Consideration of such consensus is applied in studying decision making by experts,[41–43] nursing care (clinical practice),[44,45] psychology,[46] 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 laboratories participating in an interlaboratory comparison, classifying a substance, material, or object according to its nominal and ordinal characteristics, could also be interpreted as cohesiveness. Recently developed two-way factorial analysis of variation of nominal variables CATANOVA [25] and of ordinal variables ORDANOVA,[47] applied first in refs.[25] and,[26–28] respectively, answer 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.[48] This is analogous to two-way ANOVA for continuous quantitative variables, but the variations are

calculated here from 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 the homogeneity of corresponding variations. Homogeneity testing of nominal variables in the CATANOVA framework is based on the application of a χ^2-distribution. Testing of ordinal variables in the ORDANOVA framework applies empirical distributions obtained using random Monte Carlo (MC) draws from a multinomial distribution. Since the number of participating laboratories is small in many cases, not only the level of confidence (probability of a Type I error or α-risk[49]) but also the power of the test (the probability of a Type II error or β-risk[50]) is important for a correct interpretation of the test results.[51–53]

The present Guide describes a harmonized approach to design of experiments for interlaboratory comparisons of categorical characteristics of a substance, material, or object and interpretation of the obtained data based on two-way CATANOVA and ORDANOVA. The examples provided are from the fields of macroscopic examination of weld imperfections, comparison of odor intensity of drinking water, and comparison of sensory (ordinal) characteristics of a sausage.

1.1 Scope and field of application

This Guide is developed for harmonization of interlaboratory comparisons of categorical characteristics of a substance, material, or object. It will be helpful also for a correct interpretation of categorical data on properties of substances, materials and objects, the validation of corresponding methods of characterization (e.g., methods of sensory analysis), the development of reference materials with categorical properties, and similar tasks.

The document is intended for quality control, measurement and testing chemical analytical laboratories, metrologists and analytical chemists, specialists involved in the laboratory accreditation activity, laboratory customers, quality managers, and regulators.

1.2 Terms and definitions

Terms and definitions used in this Guide are sourced from the International Vocabulary of Metrology (VIM),[23] the ISO Vocabulary of Statistics,[54–56] the ISO Quality Measurement Systems – Fundamentals and Vocabulary,[57] and the IUPAC Compendium of Terminology in Analytical Chemistry (The Orange Book).[58] A definition is given as a phrase that can be substituted in a sentence for the term, following ISO practice.[59]

The most relevant terms and definitions relating to the categorical characteristics of a substance, material, or object applied in this Guide are given below.

1.2.1 Categorical characteristic

Distinguishing feature described by a specified set of categories

NOTE 1 A category (a class or division of values) can be represented in words, by alphanumeric codes, barcodes, or pictograms.

NOTE 2 A categorical characteristic can be physical, chemical, biological, sensory (relating to smell, touch, taste, sight, hearing), etc.

NOTE 3 A categorical characteristic can be nominal (qualitative, i.e., non-quantitative) or ordinal (semi-quantitative).

NOTE 4 The term “value” regarding a categorical characteristic is intended in a broad sense including qualitative or semi-quantitative information.

NOTE 5 A categorical characteristic can be related to an inherent property of a substance or material. However, in a detection task (if a substance is present or absent) or in an identification task (if a substance is

identified or not), like determination of a blood group or a kind of weld imperfection, the categorical characteristics are related to the sample under examination, the tested environmental compartment, etc.

Adapted from the Vocabulary of Statistics [55clauses 1.1.1 and 1.1.3], and ISO 9000 [57clause 3.10].

1.2.2 Categorical data

Data of a *categorical characteristic*, each value of which is one of the specified categories

NOTE 1 Categorical data have neither measurement units nor quantity dimensions.

NOTE 2 No algebraic operations among categorical data can be performed. Their differences and ratios, where categorical data are expressed numerically, have no physical meaning.

NOTE 3 Categorical data can be *nominal data* or *ordinal data.*

NOTE 4 Binary categorical data (yes/no, detected/not detected, identified/not identified, etc.) can be classed as *nominal data* or as *ordinal data.*

Adapted from the Orange Book [58entry 2.4].

1.2.3 Consensus of laboratories

Interlaboratory consensus

Consensus

Cohesiveness (closeness, agreement) of responses of different laboratories participating in an *interlaboratory comparison* of *categorical data*

NOTE 1 The term "consensus" means statistical homogeneity of responses, which can be tested using relevant statistical methods for analysis of *categorical data* variation.

NOTE 2 Evaluation of a consensus is performed as estimation of a power of the homogeneity test of the responses and corresponding probabilities of false decisions on the homogeneity (if the consensus was achieved or not).

NOTE 3 When the purpose of the *interlaboratory comparison of categorical data* is characterization of a material (e.g., a candidate reference material[2,30]), the consensus achieved with the acceptable power and probabilities of false decisions can be used for assignment of categories to the examined properties of the material. If a laboratory is out of the consensus of other participating laboratories, responses of the outlying laboratory (inhomogeneous with other responses) should be investigated.

NOTE 4 The consensus achieved in a laboratory proficiency testing supports the proficiency of the participating laboratories. However, when the (homogeneous) responses of the laboratories differ from the assigned/certified category of the test item property,[1] an investigation of the reasons for the difference is necessary.

1.2.4 Interlaboratory comparison of *categorical data*

Comparison of *categorical data*

Design, performance, and evaluation of *categorical data* related to qualitative or semi-quantitative *categorical characteristics* of the same or similar items (results of their examination) by two or more laboratories in accordance with predetermined conditions

NOTE 1 The term "laboratories" is used to cover all organizations that provide information on items based on experimental observation, including inspection, sampling, measurement or testing, and examination.

NOTE 2 Interlaboratory comparison is a generic term; the purpose and detailed objectives of an interlaboratory comparison (e.g., proficiency testing; a procedure validation; characterization of a candidate reference material) must be specified.

Adapted from ISO/IEC 17043 [1clause 3.4]; and the Orange Book [58entry 13.62].

1.2.5 Nominal data

Categorical data with unordered labeled categories or categories ordered by convention

EXAMPLES Color of a spot test; sex; blood group; sequence of amino acids in a polypeptide; sensory response to a kind of a water smell or taste; type of fault.

NOTE 1 Nominal data have no magnitude; they can be only equal or unequal.

NOTE 2 Nominal data are used in chemical qualitative analysis [29–30,58entry 1.3].

Adapted from VIM [23, clause 1.30]; the Vocabulary of Statistics [55clause 1.1.6]; and the Orange Book [58entries 1.3 and 1.55].

1.2.6 Ordinal data

Categorical data with ordered categories according to the inherent magnitude of the data

EXAMPLES Rockwell C hardness; octane number for petroleum fuel; sensory response to intensity of a food smell or taste, severity of a fault according to an expert assessment.

NOTE 1 Ordinal data are arranged according to ordinal scales by the data categories, e.g., from very bad to excellent or from 1 to 5. However, numeric codes of categories should not be treated as continuous quantities since the distance between numbers 1 and 2 on an ordinal scale may not be the same as between 2 and 3, or 3 and 4.

NOTE 2 Ordinal data can have empirical relations only; they can be equal or unequal, greater than or less than.

NOTE 3 Ordinal data are used in chemical semi-quantitative analysis.[60–62]

Adapted from VIM [23clause 1.26]; the Vocabulary of Statistics [55clause 1.1.7]; and Orange Book [58entry 1.58].

1.2.7 Response variable of a *categorical characteristic*

Variable that represents the observed results of the examination of a *categorical characteristic*

NOTE 1 The observed results may be responses of experts participating in the examination.

NOTE 2 The distribution of the response variable of a *categorical characteristic* can be described by absolute or relative frequencies of responses of each of the specified set of categories.

Adapted from the Vocabulary of Statistics [55clause 3.5.14].

1.3 Symbols

α	risk to reject null hypothesis H_0 when it is true (probability of a Type I error)
β	risk of failure to reject the null hypothesis H_0 when in fact it is not true (probability of a Type II error)
β_l	risk of a failure to reject null hypothesis H_0 when in fact it is not true, related to factor Xl, l = 1 or 2
γ_{k0}	intercept (cutoff point) for category k in the ordinal logistic regression model
γ_1 to γ_m	logistic regression coefficients (slopes) of components contents c_1 to c_m
$\widehat{C}_B$	between(inter)-laboratory component of the sample total variation $\widehat{V}_T$
$\widehat{C}^B_{Xl}$	component of variation $\widehat{C}_B$ caused by factor Xl, l = 1 or 2
c_1 to c_m	measured values of contents of chemical components
df_B	number of degrees of freedom of variation $\widehat{C}_B$
df_T	number of degrees of freedom of variation $\widehat{V}_T$
df_W	number of degrees of freedom of variation $\widehat{V}_W$
df_{Xl}	number of degrees of freedom of variation $\widehat{C}^B_{Xl}$, l = 1 or 2
df_l	number of degrees of freedom of chi-square distribution $\chi^2_{df_l}$, l = 1 or 2
E	expected value
F_k	cumulative theoretical probability of ordinal responses up to the k-th category
$\widehat{F}_{ijk}$	sample (observed) cumulative relative frequency of ordinal responses up to category k at the i-th level of factor $X1$ and j-th level of factor $X2$
$\widehat{F}_{i.k}$	sample cumulative relative frequency of ordinal responses up to category k at level i of factor $X1$; a dot in a subscript symbol means the index of summation (for averaging) of the corresponding frequencies, e.g., j in $\widehat{F}_{i.k}$

$\widehat{F}_{.jk}$	sample cumulative relative frequency of ordinal responses up to category k at level j of factor $X2$
$\widehat{F}_{..k}$	sample total cumulative relative frequency of ordinal responses up to category k
H_0	null hypothesis
H_1	alternative hypothesis
η_1 to η_m	logistic regression coefficients (slopes) of components contents c_1 to c_m, equivalent of $-\gamma_1$ to $-\gamma_m$
i	index of a level of factor $X1$, $i = 1, 2, \ldots, I$
I	number of levels of factor $X1$
(i, j)	cell in a cross balanced design
j	index of a level of factor $X2$, $j = 1, 2, \ldots, J$
J	number of levels of factor $X2$
k	index of a category of the responses, $k = 1, 2, \ldots, K$
K	number of categories of the responses
λ	parameter of non-centrality of a distribution
l	index of a factor Xl, $l = 1$ or 2
L	estimated likelihood
m	number of considered chemical components
M_{full}	full regression model with predictors
$M_{\text{intercept}}$	regression model without predictors, i.e., containing only the intercept
n	number of replicate responses
$\boldsymbol{n}$	vector of response frequencies by categories $(n_1, n_2, \ldots, n_K)$
n_{ijk}	number (frequency) of responses of category k at i-th level of factor $X1$ and j-th level of factor $X2$
n_k	frequency of responses of category k
N	total number of responses
P	probability; used also with subscripts related to a property, e.g., properties p1 to p5
$\widehat{P}$	sample estimate of P
P_{joint}	probability of a joint event (intersection)
$\widehat{P}_{\text{joint}}$	sample estimate of P_{joint}
P_l	power of a test for assessing interlaboratory consensus, equal to probability $1 - \beta_l$, $l = 1$ or 2
$\boldsymbol{p}$	vector of the response probabilities by categories $(p_1, p_2, \ldots, p_K)$
p_k	probability of a response of category k
$\widehat{\boldsymbol{p}}$	vector of the sample probabilities (relative frequencies) of responses by categories $(\widehat{p}_{..1}, \widehat{p}_{..2}, \ldots, \widehat{p}_{..K})$
$\widehat{p}_{ijk}$	sample probability (relative frequency) of nominal responses of category k at i-th level of factor $X1$ and j-th level of factor $X2$
$\widehat{p}_{i.k}$	sample probability (relative frequency) of nominal responses of category k at i-th level of factor $X1$
$\widehat{p}_{.jk}$	sample probability (relative frequency) of nominal responses of category k at j-th level of factor $X2$
$\widehat{p}_{..k}$	sample probability (total relative frequency) of nominal responses of category k
pseudo-R^2	McFadden's statistics for evaluation of goodness-of-fit of logit models
q	index of a category of the responses (like k), $q = 1, 2, \ldots, K$
Q_{index}	quality index
$Q_{\text{index}}^{\text{exc}}$	index of a product having imagined excellent quality
$\widehat{SI}_{Xl}$	significance index (test statistics) for evaluation of effect of factor Xl, $l = 1$ or 2
SI_{Xl}^{crit}	critical value of significance index $\widehat{SI}_{Xl}$ under null hypothesis H_0, $l = 1$ or 2
$\widehat{SI}_{Xl}^{\text{MC}}$	significance index generated with the MC method, $l = 1$ or 2
$\widehat{SI}_{Xl,\lambda}$	significance index, shifted/modified under alternative hypothesis H_1 with parameter of non-centrality λ, $l = 1$ or 2
$\widehat{SI}_{Xl,\lambda}^{\text{MC}}$	significance index, shifted/modified under alternative hypothesis H_1, generated with MC method, $l = 1$ or 2
w	effect of the statistical sample size for the chi-square test
Xl	factor 1 or factor 2 ($l = 1$ or 2)
x_l	critical value of chi-square distribution $\chi^2_{df_l}$, $l = 1$ or 2
$\chi^2_{df_l}$	chi-square distribution with df_l degrees of freedom, $l = 1$ or 2
$\chi^2_{df_l,\lambda}$	non-central chi-square distribution shifted under alternative hypothesis H_1 with parameter of non-centrality λ, degrees of freedom df_l, and $l = 1$ or 2
VAR	variance
$\widehat{V}_T$	sample total variation of the response variable $\boldsymbol{Y}$, normalized on [0, 1] interval
$\widehat{V}_W$	within(intra)-laboratory component of $\widehat{V}_T$ caused by factor $X2$ and/or "residual" variation due to unknown reason(s)
$\boldsymbol{Y}$	random quantity on a categorical scale, $\boldsymbol{Y} = \boldsymbol{n}$; used also with subscripts related to a property, e.g., properties p1 to p5
$\cap$	intersection of events

1.4 Abbreviations

ANOVA	Analysis of Variance
CATANOVA	Categorical (nominal) Analysis of Variation
CDF	Cumulative Distribution Function
CITAC	Cooperation on International Traceability in Analytical Chemistry
exp	exponential function
IBM SPSS	Statistical Package for the Social Sciences of the International Business Machines Corporation
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
IUPAC	International Union of Pure and Applied Chemistry
logit	logarithmic odds of ordinal responses
MASS	Modern Applied Statistics with S programming environment
MC	Monte Carlo
ORDANOVA	Ordinal Analysis of Variation
PDF	Probability Density Function
PMF	Probability Mass Function
VIM	International Vocabulary of Metrology

2 Design of experiment

The provider of the interlaboratory comparison shall design and plan those activities which directly affect the validity of the comparison and shall ensure that activities are carried out in accordance with prescribed procedures as detailed, for example, in ISO/IEC 17043.[1] For an interlaboratory comparison in the field of sensory analysis according to ISO 6658,[31] there are important requirements for the qualification of the experts of participating laboratories (sensory assessors by ISO 8586 [63]) and conditions of examination of the test items.

2.1 Test items

Choice and preparation of test items having homogeneity and stability of the properties of interest fit for purpose of a planned interlaboratory comparison is a task of the comparison provider.[1] When test items are consumer products (e.g., samples of a packaged sausage) from different producers, purchased simultaneously from a market for comparison,[64] they shall be examined before their expiration dates.

2.2 Responses

2.2.1 Modeling responses

An expert response for a given property (characteristic of a substance, material, or object) can be modeled as a random quantity $\boldsymbol{Y}$ on a categorical scale with $K \geq 2$ categories (classes or levels) characterized by a probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$, 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, 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 $\boldsymbol{n} = (n_1, n_2, \ldots, n_K)$, 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):[65]

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

When some of the numbers n_k of responses related to the k-th category are equal to zero, factorial 0! = 1, and corresponding $p_k^{n_k} = 1$. Therefore, if two only n_k are different from zero (i.e. $K = 2$), the multinomial distribution in Eq. (1) simplifies to the binomial distribution. Note that the multinomial distribution, being a generalization of the binomial distribution, is applicable to nominal as well as ordinal variables.

2.2.2 Factors influencing the responses

In an interlaboratory comparison, 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 $i = 1, 2, \ldots, I$), 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 $j = 1, 2, \ldots, J$). Each of the N possible responses falls into one of the I levels i of the first factor $X1$, and into one of the J levels j of the second factor $X2$. Besides, each of the responses belongs to one of the $k = 1, 2, \ldots, K$ categories of $\boldsymbol{Y}$.

2.2.3 Interaction of the factors

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 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 versus 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.

2.3 Cross-balanced design

A design of an interlaboratory comparison 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 analyzed when all $n_{ijk} = 1$.

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.[25,47]

Mathematical issues of random effects of factors for a nominal scale were described recently in ref. [66].

Table 1: Cross-balanced design without replication.

Factor $X1$ – laboratories	Factor $X2$ – levels of a condition					Total counts
	1	…	***j***	…	*J*	
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 counts	I	…	I	…	I	N

3 Analysis of the response variation

3.1 Total variation

Treating N responses as a statistical sample, and the number of responses n_{ijk} as a random variable, then $\widehat{p}_{ijk} = n_{ijk}/N$ and $\widehat{F}_{ijk} = \sum_{q=1}^{k} \widehat{p}_{ijq}$ 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

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

Here, $\widehat{F}_{i.k} = \frac{1}{J} \sum_{j=1}^{J} \widehat{F}_{ijk}$ $(i = 1, 2, \ldots, I\,;\ k = 1, 2, \ldots, K)$ and $\widehat{F}_{.jk} = \frac{1}{I} \sum_{i=1}^{I} \widehat{F}_{ijk}$ $(j = 1, 2, \ldots, J\,;\ k = 1, 2, \ldots, K)$ 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 $\widehat{F}_{..k}$.

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 [47] as

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

with degrees of freedom $df_{\mathrm{T}} = N - 1$. A similar estimate in two-way CATANOVA for nominal variables [67] is

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

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

3.2 Decomposition of the total variation

In the model without replication, the total sample variation $\widehat{V}_{\mathrm{T}}$ is partitioned into the between(inter)-laboratory component $\widehat{C}_{\mathrm{B}}$ and the within(intra)-laboratory component $\widehat{V}_{\mathrm{W}}$, caused by the second factor and/or "residual" variation due to unknown reason(s). For ordinal data,[47] this is

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

where

$$\widehat{C}_{\mathrm{B}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \left[\frac{1}{I} \sum_{i=1}^{I} \left(\widehat{F}_{i.k} - \widehat{F}_{..k}\right)^2 + \frac{1}{J} \sum_{j=1}^{J} \left(\widehat{F}_{.jk} - \widehat{F}_{..k}\right)^2 \right] \tag{6}$$

and

$$\widehat{V}_{\mathrm{W}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\widehat{F}_{i.k} + \widehat{F}_{.jk} - \widehat{F}_{..k}\right)\left(1 - \left[\widehat{F}_{i.k} + \widehat{F}_{.jk} - \widehat{F}_{..k}\right]\right). \tag{7}$$

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 the next decomposition of the variation $\widehat{C}_{\mathrm{B}}$:

$$\widehat{C}_{\mathrm{B}} = \widehat{C}^{\mathrm{B}}_{X1} + \widehat{C}^{\mathrm{B}}_{X2}, \tag{8}$$

where

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

$$\widehat{C}^{\mathrm{B}}_{X2} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \frac{1}{J} \sum_{j=1}^{J} \left(\widehat{F}_{.jk} - \widehat{F}_{..k}\right)^2 \tag{9}$$

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

A similar decomposition for nominal data [67] leads to

$$\widehat{C}^{\mathrm{B}}_{X1} = \frac{K}{K-1} \sum_{k=1}^{K} \frac{1}{I} \sum_{i=1}^{I} \left(\widehat{p}_{i.k} - \widehat{p}_{..k}\right)^2 \text{ and } \widehat{C}^{\mathrm{B}}_{X2} = \frac{K}{K-1} \sum_{k=1}^{K} \frac{1}{J} \sum_{j=1}^{J} \left(\widehat{p}_{.jk} - \widehat{p}_{..k}\right)^2. \tag{10}$$

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.[25–27] Note also that the sample estimators by Eqs. (3–10) are biased from the corresponding population variations.[42,68]

3.3 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 data:

$$\frac{E\left[\widehat{V}_{\mathrm{T}}\right]}{df_{\mathrm{T}}} = \frac{E\left[\widehat{C}_{\mathrm{B}}\right]}{df_{\mathrm{B}}} = \frac{E\left[\widehat{V}_{\mathrm{W}}\right]}{df_{\mathrm{W}}} = \frac{E\left[\widehat{C}^{\mathrm{B}}_{X1}\right]}{df_{X1}} = \frac{E\left[\widehat{C}^{\mathrm{B}}_{X2}\right]}{df_{X2}} = \frac{V_{\mathrm{T}}}{N}, \tag{11}$$

where E is the expected value. The numerator of the last term in Eq. (11) 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.,

$$\frac{E\left[\widehat{C}^{\mathrm{B}}_{X1}\right]}{df_{X1}} > \frac{V_{\mathrm{T}}}{N} \quad \text{and/or} \quad \frac{E\left[\widehat{C}^{\mathrm{B}}_{X2}\right]}{df_{X2}} > \frac{V_{\mathrm{T}}}{N}. \tag{12}$$

To test the statistical significance of both the factor effects, the following significance indices (test statistics) have been defined:[47]

$$\widehat{SI}_{X1} = \frac{\widehat{C}^{\mathrm{B}}_{X1} \Big/ df_{X1}}{\widehat{V}_{\mathrm{T}} \Big/ df_{\mathrm{T}}} \quad \text{and} \quad \widehat{SI}_{X2} = \frac{\widehat{C}^{\mathrm{B}}_{X2} \Big/ df_{X2}}{\widehat{V}_{\mathrm{T}} \Big/ df_{\mathrm{T}}}. \tag{13}$$

3.4 Hypothesis testing for nominal variables

Distributions of the statistics $df_l\, \widehat{SI}_{Xl}$, $l = 1, 2$, for nominal variables are asymptotically approximated by the chi-square distributions $\chi^2_{df_l}$ [25] with $df_1 = (K-1)(I-1)$ and $df_2 = (K-1)(J-1)$, respectively. They have the following expectations and variances:

$$E\left[df_l\, \widehat{SI}_{Xl}\right] = df_l \text{ and } VAR\left[df_l\, \widehat{SI}_{Xl}\right] = 2\, df_l. \tag{14}$$

This approximation allows the application of a chi-square test for testing the null and alternative hypotheses.[69] 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 α-risk of a false decision that a consensus of laboratories is absent, when it is actually achieved. Similarly, the H_0 regarding the levels of factor $X2$ $\left(p_{.jk} = p_k\right)$ 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 α-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. (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

$$E\left[df_l \widehat{SI}_{Xl,\lambda}\right] = df_l + \lambda,\ VAR\left[df_l \widehat{SI}_{Xl,\lambda}\right] = 2\,df_l + 4\,\lambda, \tag{15}$$

and

$$E\left[\widehat{SI}_{Xl,\lambda}\right] = 1 + \frac{\lambda}{df_l},\ VAR\left[\widehat{SI}_{Xl,\lambda}\right] = \frac{2}{df_l} + \frac{4\,\lambda}{df_l^2}. \tag{16}$$

This modified distribution is approximated by the noncentral chi-square distribution $\chi^2_{df_l,\lambda}$[70]. The values are calculated as $\lambda = w^2 N$, 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.[71] 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 P_l, l = 1 or 2, of the corresponding chi-square test:[72,73]

$$P_l = 1 - \beta_l = 1 - \mathrm{CDF}\chi^2_{df_l,\lambda}\,(x_l), \tag{17}$$

where CDF means cumulative distribution function, and β_l denotes the probability of a Type II error. It may be interpreted in the case of factor $X1$ as the β-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 β-risk of a false decision of the factor insignificance, when it is significant.

An example of the power values versus numbers of the factor levels and categories, calculated and plotted in R programming environment,[48,74] is available in Annex A, Example 1. An evaluation of the consensus (the power and β-risk at the set α-risk) of laboratories which participated in a comparison of weld imperfections that are nominal weld characteristics, is in Annex A, Example 2.

3.5 Hypothesis testing for ordinal variables

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

A calculation tool using random MC draws from a multinomial distribution – an Excel spreadsheet with macros [75] – calculates (from the empirical data) the sample vector of relative frequencies $\hat{\boldsymbol{p}} = \left(\hat{p}_{..1}, \hat{p}_{..2}, \ldots, \hat{p}_{..K}\right)$, as well as the variation components $\left(\widehat{C}^{\mathrm{B}}_{X1}, \widehat{C}^{\mathrm{B}}_{X2}, \widehat{V}_{\mathrm{W}}, \widehat{V}_{\mathrm{T}}\right)$ 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 $\widehat{\boldsymbol{p}}$, 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}^{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}^{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 $P_l = 1 - \mathrm{CDF}_{\widehat{SI}_{Xl,\lambda}^{\mathrm{MC}}}\left(SI_{Xl}^{crit}\right)$.

The tool was applied, for example, for evaluation of the consensus of laboratories which participated in a comparison of intensity of odors (ordinal characteristics) of different drinking water samples.[26] This is Example 3 in Annex A.

4 Relationship between categorical and quantitative characteristics

4.1 Sensory evaluation and chemical composition

Categorical characteristics of a substance, material, or object may be correlated with its quantitative characteristics, such as contents of main chemical components and impurities. For example, relationships between sensory evaluation and the chemical composition of meat and meat products have been a subject of research.[76–78] Regression analysis is the known tool for studying and modeling such relationships. However, as in applications of ANOVA, the additivity assumption should not be violated when applying regression analysis.[79] This is possible with multinomial ordered logistic regression (ordered logit), quite commonly applied in medicine,[80] financial activity,[81] in a study of consumer purchasing behavior,[82] and in other fields.

4.2 Multinomial ordered logistic regression

The ordered logit model is based on the following concepts. When $\boldsymbol{Y}$ is an ordinal outcome with K categories, the cumulative probability of the responses of categories q, which are less than or equal to a category k, is $P(q \le k)$. The odds of the responses being less than or equal to a category k, are defined as $P(q \le k)/P(q > k)$, when $k = 1, \ldots, K-1$. For $k = K$, the denominator is zero and the odds cannot be defined. The log odds, called logit, is defined as $\mathrm{logit}\,(P(q \le k)) = \log\,(P(q \le k)/P(q > k))$. The ordinal logistic regression model is parameterized as

$$\mathrm{logit}\,(P(q \le k)) = \gamma_{k0} + \gamma_1 c_1 + \cdots + \gamma_m c_m, \tag{18}$$

where γ_{k0} is the intercept (cutoff point) for category k; c_1 to c_m are the measured component contents (mass fractions), i.e., the observable continuous variables; γ_1 to γ_m are the corresponding regression coefficients (slopes), constant across categories. Note that this model is based on the parallel regression (proportional odds) assumption: the logit dependences on the compositions are parallel hyperplanes for different categories k and, hence, the intercepts are different for each category but the slopes are constant across categories. The odds of being less than or equal to category k are

$$P(q \le k)/P(q > k) = \exp\left(\gamma_{k0} + \gamma_1 c_1 + \cdots + \gamma_m c_m\right). \tag{19}$$

Calculation of the model parameters in R programming environment, including their confidence intervals and goodness-of-fit measures for the model, is described, for example, at the webpage.[83] The following notation is used in R:

$$\text{logit}\left(P\left(q \le k\right)\right) = \gamma_{k0} - \eta_1 c_1 - \cdots - \eta_m c_m, \tag{20}$$

where $\eta_i = -\gamma_i$ for all the regression coefficients of components contents c_1 to c_m. Inverting Eq. (20), the probability of getting a response of a certain category k or below can be obtained[84] as

$$P\left(q \le k\right) = \text{logit}^{-1}\left(P\left(q \le k\right)\right) = 1\Big/\left[1 + 1/\exp\left(\gamma_{k0} - \eta_1 c_1 - \cdots - \eta_m c_m\right)\right]. \tag{21}$$

The R function "polr" of the MASS package[85] is used to fit multinomial ordered logistic models to the experimental data, while the function "predictorEffects" of the Effects package[86] is helpful for calculating and plotting probabilities of getting a response equal to a certain category k.

The goodness-of-fit of a model can be evaluated by calculating several pseudo-R statistics,[87] which estimate the variability in the outcome of the fitted model. For example, McFadden's pseudo-R^2 is defined as

$$\text{pseudo-}R^2 = 1 - \ln L\left(M_{\text{full}}\right)/\ln L\left(M_{\text{intercept}}\right), \tag{22}$$

where M_{full} is a full model with predictors; $M_{\text{intercept}}$ is the model without predictors, i.e., containing only the intercept; and L is the estimated likelihood.

When the M_{full} model does not predict the outcome better than the $M_{\text{intercept}}$ model, its ln $L(M_{\text{full}})$ is not much larger than ln $L(M_{\text{intercept}})$, hence the corresponding ratio is close to 1 and the McFadden's pseudo-R^2 is close to 0: the model has poor predictive value. Conversely, when the M_{full} model is good, its ln $L(M_{\text{full}})$ is close to zero since the likelihood value for each observation is close to 1, and McFadden's pseudo-R^2 is close to 1, indicating successful predictive ability. When comparing two models on the same data, McFadden's pseudo-R^2 would be higher for the model with the greater likelihood.

The *R* function "PseudoR2" of the DescTools package[88] is applicable to the corresponding calculations. Note that correlations between contents of the chemical components may affect the regression coefficients and p values, but they do not influence the predictions, precision of predictions, and the goodness-of-fit statistics.[89]

An example of multinomial ordered logistic regression of sensory responses to the quality of a sausage from different producers versus the chemical composition of this sausage, is available in Annex A, Example 4.

5 Multisensory quality index

A quality index summarizing the responses to different properties of a product is an important measure for assessing the quality of commercial products.[90–92] It can be useful in comparative testing of the same product from different manufacturers[64] and for prediction models of a consumer choice. However, when the responses are ordinal quality characteristics, the problem of the published approaches to the index calculation is again that they use different kinds of mean, while the corresponding algebraic operations with categorical data cannot be performed (Sec. 1.2.2, Note 2). An additional difficulty is caused by multisensory perception[93–95] which leads to possible interactions (correlation) of the responses to different product properties.[28]

5.1 The index for independent responses

The probability mass function P by Eq. (1) of a multinomial random variable $\boldsymbol{Y}$, characterized by a vector $\boldsymbol{p}$ of response probabilities, is the probability that the event $\boldsymbol{Y} = \boldsymbol{n}$ occurs. For example, five such multinomial variables, each corresponding to a product quality property, will be used further with corresponding subscripts from p1 to p5 in the symbols of variables and parameters related to these quality properties.

The probability P_{joint} of the joint (intersection) event, consisting in the expert responses to the five properties simultaneously, is:[96,97]

$$P_{joint} = P\left(\left\{\boldsymbol{Y}_{\text{p1}} = \boldsymbol{n}_{\text{p1}}\right\} \cap \left\{\boldsymbol{Y}_{\text{p2}} = \boldsymbol{n}_{\text{p2}}\right\} \cap \left\{\boldsymbol{Y}_{\text{p3}} = \boldsymbol{n}_{\text{p3}}\right\} \cap \left\{\boldsymbol{Y}_{\text{p4}} = \boldsymbol{n}_{\text{p4}}\right\} \cap \left\{\boldsymbol{Y}_{\text{p5}} = \boldsymbol{n}_{\text{p5}}\right\}\right), \tag{23}$$

where $\cap$ is the symbol of intersection of events.

The Venn diagram of the joint event is shown in Fig. 1.

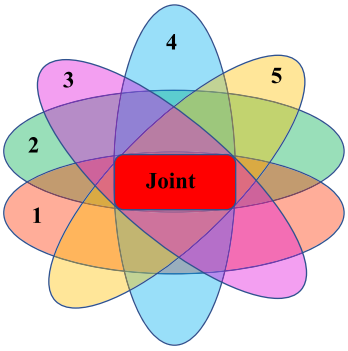

Fig. 1: Venn diagram of the joint event. The event $\boldsymbol{Y}_{p1} = \boldsymbol{n}_{p1}$, when the probabilities of the responses to property 1 by categories are as in the specific vector $\boldsymbol{p}_{p1}$, is shown with a semi-transparent brown ellipse 1; similar color ellipses indicate 2 (green) – the event $\boldsymbol{Y}_{p2} = \boldsymbol{n}_{p2}$ for property 2; 3 (violet) – the event $\boldsymbol{Y}_{p3} = \boldsymbol{n}_{p3}$ for property 3; 4 (blue) – the event $\boldsymbol{Y}_{p4} = \boldsymbol{n}_{p4}$ for property 4; and 5 (yellow) – the event $\boldsymbol{Y}_{p5} = \boldsymbol{n}_{p5}$ for property 5. The joint (intersection) event, consisting of the corresponding responses to the five properties simultaneously, is highlighted by the central red non-transparent shape.[28]

When responses to these five quality properties are independent, the probability of the joint event (joint probability) is the product:

$$P_{\text{joint}} = P_{p1} \cdot P_{p2} \cdot P_{p3} \cdot P_{p4} \cdot P_{p5}. \tag{24}$$

Treating N responses to each quality property as a separate statistical sample, and corresponding frequencies n_k as random variables, the probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ is estimated for the quality property as a vector of relative frequencies $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$, where $\hat{p}_k = n_k/N$. Then, an estimate of P_{p1} by Eq. (1) is $\hat{P}_{p1}$, similarly for other probabilities, while the estimate of the joint probability by Eq. (24) is $\hat{P}_{\text{joint}} = \hat{P}_{p1} \cdot \hat{P}_{p2} \cdot \hat{P}_{p3} \cdot \hat{P}_{p4} \cdot \hat{P}_{p5}$. The estimate $\hat{P}_{\text{joint}}$ is the probability that the product will have these, and no other, quality characteristics, expressed as the sensory property values on their ordinal scales.

The product quality index Q_{index} can be formulated as the negative common logarithm ($-\log_{10}$ or $-$ lg) of the estimate of the joint probability:

$$Q_{\text{index}} = -\lg\left(\hat{P}_{\text{joint}}\right) = -\left[\lg\left(\hat{P}_{p1}\right) + \lg\left(\hat{P}_{p2}\right) + \lg\left(\hat{P}_{p3}\right) + \lg\left(\hat{P}_{p4}\right) + \lg\left(\hat{P}_{p5}\right)\right]. \tag{25}$$

When the probability estimates $\hat{P}$ of the five quality properties tend to 1, the quality index approaches its minimum value $Q_{\text{index}} = 0$. In any other case $Q_{\text{index}} > 0$. A greater Q_{index} value means smaller values of the joint probability P_{joint} and its estimate $\hat{P}_{\text{joint}}$, i.e., a greater chance that the product quality properties will differ from the claimed ones. In this sense, greater Q_{index} values are worse, and the estimation of Q_{index} can be compared with, for example, counting a number of defects. Being the negative logarithm by Eq. (25), the quality index is related to the entropy (a measure of uncertainty[98–100]) of the probability distribution of that product quality properties.

Note that there are no assumptions concerning the contribution of each quality property to the quality index, either being equal or different: the formulated quality index is not a kind of geometric or weighted mean of the property values with probabilities as the weights. Determining the quality characteristics and their categories (the ranges on the ordinal scale), as well as the relevance of these characteristics to consumers, are not the task of this Guide. They are a part of the product specifications in a standard or another regulatory document related to the product.

5.2 The index for responses which might not be independent

When responses to two or more quality properties might not be independent, the probability of the joint (intersection) event P_{joint} can be represented numerically by a Gaussian copula-based procedure.[101,102] This procedure is used for generating samples from a discrete multivariate random variable with the prescribed experimental marginal cumulative distributions $\hat{F}_{k\_p1}, \hat{F}_{k\_p2}, \hat{F}_{k\_p3}, \hat{F}_{k\_p4}, \hat{F}_{k\_p5}$ and an empirical correlation matrix of those quality properties. When a large number of those multivariate samples, each of size N, are generated, an estimate $\hat{P}_{\text{joint}}$ of P_{joint} can be calculated as the relative frequency of realization of the intersection event $(\{\boldsymbol{Y}_{p1} = \boldsymbol{n}_{p1}\} \cap \{\boldsymbol{Y}_{p2} = \boldsymbol{n}_{p2}\} \cap \{\boldsymbol{Y}_{p3} = \boldsymbol{n}_{p3}\} \cap \{\boldsymbol{Y}_{p4} = \boldsymbol{n}_{p4}\} \cap \{\boldsymbol{Y}_{p5} = \boldsymbol{n}_{p5}\})$. Then, the quality index $Q_{\text{index}} = -\lg\left(\hat{P}_{\text{joint}}\right)$ is obtained.

5.3 Variability of the quality index

A representative dataset of responses to each property is necessary for an estimation of the vector of relative frequencies and PMFs by Eq. (1). It may be a dataset containing results of examination of samples from one batch. The quality index characterizes this batch in such a case. When the dataset is accumulated by examination of batches during a specified term, the quality index characterizes the product (and production in the specified term) in general. The variability of the Q_{index} value for $N > 100$ is decreasing and may be considered as negligible at the quality index evaluation.

However, it is important that the Q_{index} value depends on the definition of the joint probability P_{joint} and its estimate $\widehat{P}_{\text{joint}}$. For example, the quality index may also be formulated as the negative common logarithm of the joint probability of the product with some specified (preferred) categories of its quality properties. In particular, in the case of an "ideal" product with the excellent (highest) category of each examined property, the joint probability means the probability that the product will be made at the same production conditions as other products of the same kind during the studied term. Then, the corresponding quality index $Q_{\text{index}}^{\text{exc}}$ is the negative common logarithm of the joint probability of the event when the vector of frequencies of responses of the highest category is obtained for each of the properties.

Such quality indices are discussed in Annex A, Example 5, using the described approach for analysis of a dataset related to a sausage from two producers.[28]

6 Implementation

In practice, the purposes of comparisons and the number of laboratories (producers) able and ready to participate in a comparison may be different, and the number of test items may be small or large. When a consensus is not achieved (the responses are inhomogeneous), outlying laboratory responses should be investigated. The outlier(s) may be removed from the dataset for calculation if the laboratory finds that the conditions of the experiment were violated. If a violation was not detected, removing outliers is not recommended[1,13,16] as this action decreases the power of the test and increases the risks of false decisions.

In any case, a fit-for-purpose algorithm is to be elaborated for correct treatment and interpretation of the obtained data, especially in the complex interlaboratory comparisons of responses correlated with chemical composition of an object or correlated responses to different properties of the same product. Such an algorithm can be represented as a flow chart.

6.1 Algorithms for data treatment

When a property of samples of a product of different producers is examined in one laboratory/institution by a group of experienced experts, and the chemical composition of each sample is known from a certificate provided by its producer, a flow chart of data treatment is shown in Fig. 2. It starts from calculation of the frequencies of expert responses (of different categories), and evaluation of the total variation. The next step is decomposition of the total variation into components with the purpose to assess the effects of two factors influencing the variation – "producer" ($X1$) and "expert" ($X2$).

The components of variation obtained are used for testing the hypotheses on homogeneity of the producers (i.e., the responses to their product quality properties) and homogeneity of the experts (their responses to a property of the same product sample). When responses of different experts are inhomogeneous, and/or the responses to different producers are homogeneous, the analysis is ended. Otherwise, it is assumed that the difference between responses to the product quality of different producers is caused by the differences in the product chemical composition. This hypothesis is tested with multinomial ordered logistic regression analysis. If any of the regression coefficients are statistically significant, probabilities of obtaining a response related to a specific category for different components of the chemical composition are calculated. The last step is plotting such probabilities for visualization of the results and their discussion like in Annex A, Example 4.

Another algorithm is necessary, when the chemical composition of each sample of a product under examination is known and correlations between the responses to different properties of the product are considered. Such an algorithm may start from testing the homogeneity of the datasets of chemical composition that able to

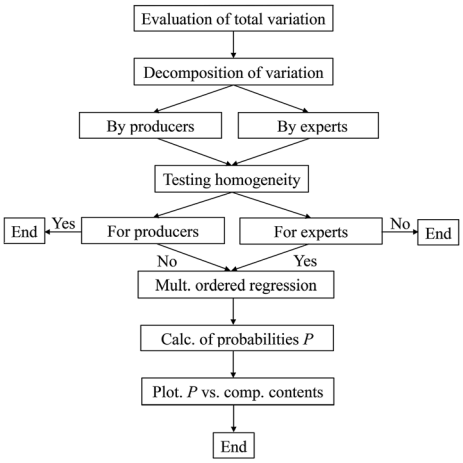

Fig. 2: Flow chart of the data treatment for interlaboratory comparison of categorical characteristics of a substance, material, or object.[27]

influence the responses. For samples, for which the hypothesis on homogeneity of the chemical composition is not rejected, ORDANOVA is implemented. Then only testing correlation of the responses to the different quality properties can be performed. If correlation is not statistically significant, a quality index of the product can be calculated by Eq. (25). Otherwise, the quality index is calculated numerically by a Gaussian copula-based procedure as explained in Sec. 5.2 and Annex A, Example 5.

6.2 Limitations

This Guide discusses the risks of false decisions on consensus as probabilities, not considering the severity of their consequences: quality loss, aesthetic and taste worsening in a product, financial loss, etc. There are also typical limitations of the applied methods of mathematical statistics: the use of any model is a simplified reflection of reality; adequacy of the treatment of a dataset of item-to-item (batch-to-batch) and/or expert-to-expert responses; the goodness-of-fit of experimental and theoretical distributions, etc.

Acknowledgments: The Task Group would like to thank I. Andrić (Croatia), V. N. Naidenko (Russia), M. N. Salikova (Russia) and Y. N. Yariv (Israel) for the help in preparation of the Examples in Annex A of this Guide.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: All other authors state no conflict of interest.
Research funding: This work was prepared under projects 2021-017-2-500 and 2023-016-1-500 of IUPAC (Funder ID: 10.13039/100006987).
Data availability: Not applicable.

Annex A: Examples

Example 1. Calculation of power of the test for nominal variables

A-1-1 Introduction

The aim of this example is to illustrate the dependence of values of power, P_1 and P_2, on the number of laboratories $I = 3$ to 50, and the number of levels of the second factor (condition) $J = 2$ to 10, at the number of

categories of the responses $K = 3$ and 10. Calculations are based on the application of the chi-square distribution in R programming environment.[48] The least number of categories $K = 3$ is set as binary categorical cases ($K = 2$) have already been discussed in previous publications.[33,66] The least number of laboratories is $I = 3$ since bilateral interlaboratory comparisons ($I = 2$) are a specific case in metrology[13,103] not considered here. The range for J starts at $J = 2$, which is usual for testing the influence of a comparison condition on the responses.[48]

A-1-2 Examples of the power calculations for nominal variables

The calculated results for P_1 and P_2 are shown as the yellow and blue transparent surfaces in Figs. 3 and 4, respectively. The calculations are performed at the probability of a Type I error $\alpha = 0.05$ and the medium effect of the sample size $w = 0.3$. Plots (a) and (b) in each figure correspond to $K = 3$ and 10, respectively. Smoothing each surface plot from corresponding discrete values was performed using R.[74]

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 Fig. 3 are from

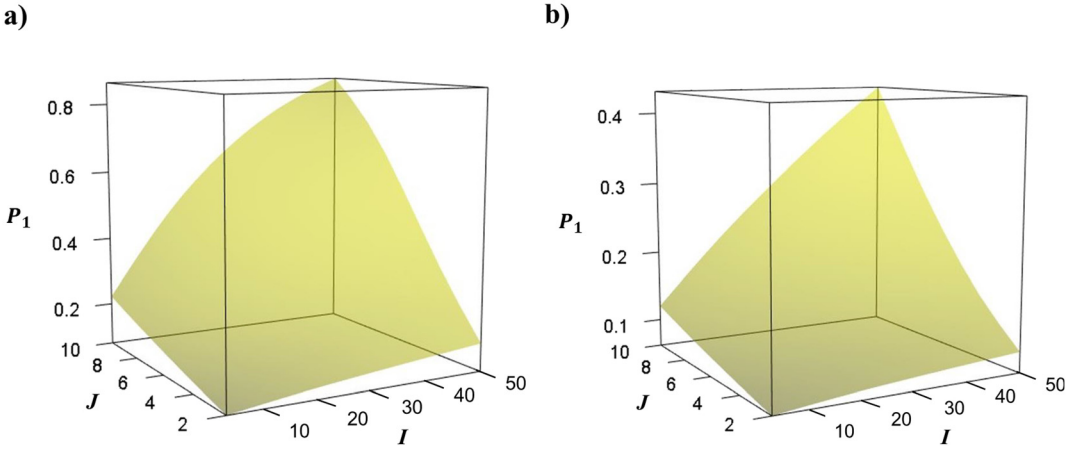

Fig. 3: Power P_1 of the test of the hypothesis on significance of the effect of factor $X1$ in dependence on the number I of laboratories (levels of factor $X1$) and the number J of conditions (levels of factor $X2$). Plots (a) and (b) correspond to the number of response categories $K = 3$ and 10, respectively.[48]

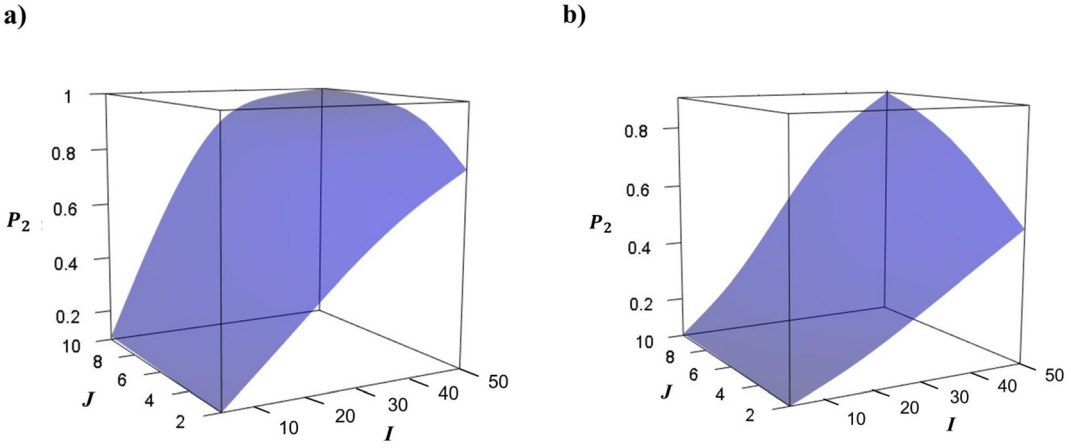

Fig. 4: Power P_2 of the test of the hypothesis on significance of the effect of factor $X2$ in dependence on the number I of laboratories and the number J of conditions. Plots (a) and (b) correspond to the number of response categories $K = 3$ and 10, respectively.[48]

0.08 to 0.86 for $K = 3$ in plot (a), and 0.06 to 0.43 for $K = 10$ in plot (b). In Fig. 4 power P_2 values are from 0.09 to 1.00 for $K = 3$ in plot (a) and 0.07 to 0.90 for $K = 10$ in plot (b). 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 Figs. 3 and 4 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) by Eqs. (15) and (16), 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, to solve a more complex task, a greater N is necessary.

Example 2. A comparison of weld imperfections

A-2-1 Introduction

This example demonstrates the implementation of two-way CATANOVA 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 evaluation of the consensus of the comparison participants when their responses are nominal values.

A-2-2 Experiment

Three accredited laboratories participated in the comparison,[25] $I = 3$, and were asked to recognize and classify weld imperfections according to ISO 6520-1.[104] This standard defines the designation system for macroscopic examination of 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. 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 $N = nIJ = 84$[25].

A-2-3 Implementation of two-way CATANOVA

Numbers of responses by category obtained in each laboratory (technicians 1 and 2) to the 14 imperfections, i.e., frequencies n_{ijk}, are presented in Table 2. The total sample variation of the examination results is $\widehat{V}_T = 0.952$ with $df_T = 83$ by Eq. (4); the between (inter-) laboratory variation is $\widehat{C}_B = 0.047$ with $df_B = 5$ by Eq. (6), and the within (intra-) laboratory variation is $\widehat{V}_W = 0.906$ with $df_W = 78$ by Eq. (7). The $\widehat{C}_B$ decomposition by Eq. (8) and Eq. (10) was performed here considering in addition a possible interaction of the factors.

The individual effect of laboratories as factor $X1$ was $\widehat{C}^B_{X1} = 0.019$ with $df_1 = 8$ degrees of freedom, and the effect of technicians as factor $X2$ was $\widehat{C}^B_{X2} = 0.0149$ with degrees of freedom $df_2 = 4$. Their significance indices were $\widehat{SI}_{X1} = 0.834$ and $\widehat{SI}_{X2} = 1.297$, respectively.

Since $df_1 \widehat{SI}_{X1} = 6.67$ does not exceed the critical value of the chi-square distribution χ^2_8 at 95 % level of confidence $x_1 = 15.51$, and $df_2 \widehat{SI}_{X2} = 5.19 < x_2 = 9.49$, the null hypotheses H_0 on homogeneity of the laboratories and on homogeneity of their technicians were not rejected at 95 % level of confidence. In other words, the influence of both the factors, laboratories $X1$ and technicians $X2$, on the responses was found insignificant at the α-risk, i.e., the probability of a Type I error, $\alpha = 0.05$. The factors' interaction was also tested in a similar way and found insignificant [25].

Table 2: Numbers of the responses by laboratory, technician, and categories, n_{ijk}.

Category, *K*	Laboratory (*X*1), *i*						Total
	1		2		3		
	Technician (*X*2), *j*						
	1	2	1	2	1	2	
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

A-2-4 Evaluation of the consensus

Under hypothesis H_0 for the factor $X1$, the expectation according to Eq. (14) is $E\left[df_1\,\widehat{SI}_{X1}\right] = 8$, the variance is $VAR\left[df_1\,\widehat{SI}_{X1}\right] = 16$, and the standard deviation is $sd_1 = \sqrt{VAR\left[df_1\,\widehat{SI}_{X1}\right]} = 4$. The distribution of $df_1\,\widehat{SI}_{X1}$ is approximated by the chi-square distribution 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 size effect $w = 0.3$, the parameter of non-centrality of the distribution is $\lambda = w^2N = 7.56$. By Eq. (15), $E\left[df_1\,\widehat{SI}_{X1,7.56}\right] = 15.56$, $VAR\left[df_1\,\widehat{SI}_{X1,7.56}\right] = 46.24$, 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 Fig. 5, 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. The vertical black dashed line indicates the critical value x_1. Probabilities of a Type I error are shown by the transparent blue area to the right of the dashed line, and probabilities of a 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, $E\left[df_2\,\widehat{SI}_{X2}\right] = 4$, $VAR\left[df_2\,\widehat{SI}_{X2}\right] = 8$, 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 $\lambda = 7.56$, $E\left[df_2\,\widehat{SI}_{X2,7.56}\right] = 11.56$, $VAR\left[df_2\,\widehat{SI}_{X2,7.56}\right] = 38.24$, 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 probability density functions of the chi-square distributions χ^2_4 and $\chi^2_{4,7.56}$ for factor $X2$ are shown on plot (b) in Fig. 5. The notations are the same as on plot (a) in this figure. The power of the test of insignificance of factor $X2$ is $P_2 = 0.58$.

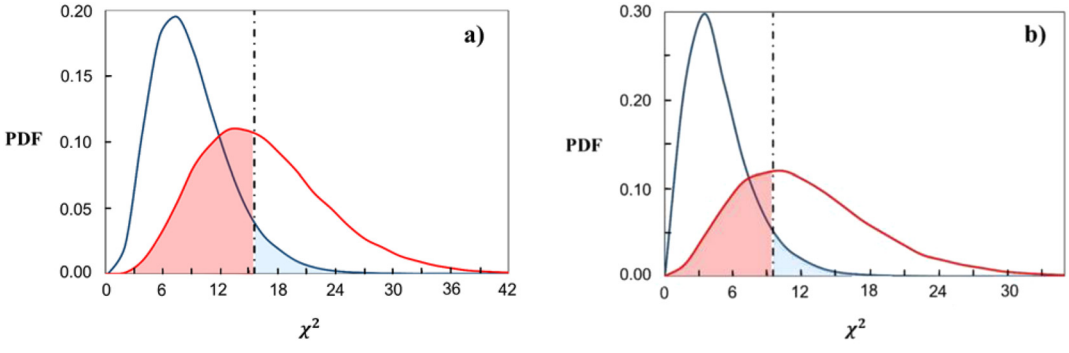

Fig. 5: Probability density functions (PDFs) of the chi-square distributions. Plot (a) is for factor $X1$, and plot (b) is for factor $X2$. The blue lines demonstrate the PDF of the chi-square distribution under hypothesis H_0, the red lines – under hypothesis H_1. The vertical black dashed lines indicate the critical values x_1 and x_2 for 95 % level of confidence, for plot (a) and for plot (b), respectively. Probabilities of a Type I error α are shown as the shaded area of transparent blue color, and probabilities of a Type II error β – by the shaded area of transparent red color.[48]

In other words, a consensus of the laboratories in assessment of the weld imperfections, and an agreement 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 a 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.[47]

Example 3. A comparison of the intensity of odors of drinking water

A-3-1 Introduction

The objective of this example is demonstration of 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, and an evaluation of the consensus of the comparison participants when their responses are ordinal values.

A-3-2 Experiment

Two test items, 1 and 2, were prepared for examination of the intensity of a chlorine and a sulfurous odor, respectively.[26] The components of these items were purchased bottled drinking water (from the same producer and batch), 330 cm^3 in a plastic container for each item, and the initial solutions of the pure reagents in glass vials: 3 cm^3 of sodium hypochlorite, 0.544 g/dm^3, for test item 1 providing a chlorine odor; and 3 cm^3 of sodium sulfide, 0.167 g/dm^3, for test item 2 providing a sulfurous odor.

The solution of sodium hypochlorite was mixed with the drinking water before use by each participating laboratory to obtain the final concentration of sodium hypochlorite in test item 1 equal to 4.9 mg/dm^3. This concentration of sodium hypochlorite corresponds to intensity level 2 of chlorine odor, interpolated between levels 1 and 3 described in the national standard GOST R 57164.[105] The final concentration of sodium sulfide in test item 2 equal to 1.5 mg/dm^3 was obtained by mixing its initial solution with the drinking water, also before use by each participating laboratory. This concentration of sodium sulfide corresponds to intensity level 4 of sulfurous odor, interpolated between levels 3 and 5 by ref. [105].

The assigned categories of the intensity of odor in the prepared items were set according to the preparation procedure.[1] The influence of any lack of chemical homogeneity of the initial solutions on the assigned categories was negligible. The solutions of sodium hypochlorite and sodium sulfide were stable for three weeks, when kept in tightly-closed glassware between temperatures from 4 °C to 20 °C. The stability of the test items 1 and 2 was not relevant, as they were prepared immediately before use.

The components of items 1 and 2 were distributed to the 49 ecological laboratories which participated in the comparison in random order. 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 according to the national standard.[105] There are 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 standard assigns the respective numeric score from 0 to 5. The temperature of a test item was measured and adjusted to (20 ± 2) °C by keeping at room temperature. To adjust a test item temperature to (60 ± 5) °C, the flask with the item was immersed in a water bath for heating.

Finally, 45 laboratories reported. 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 intensity (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.[26]

Table 3: Numbers of the responses by a water sample temperature and categories, n_{jk}.

Category, k	Chlorine odor		Total	Sulfurous odor		Total
	Temperature ($X2$), j			Temperature ($X2$), j		
	1	2		1	2	
0	2	2	4	0	0	0
1	24	14	38	0	0	0
2	10	19	29	0	0	0
3	9	10	19	10	12	22
4	0	0	0	17	11	28
5	0	0	0	18	22	40
Total	45	45	90	45	45	90

A-3-3 Implementation of two-way ORDANOVA without replication

Numbers of the responses (examination results) obtained from all laboratories by categories at the specified temperature of a water sample, i.e., frequencies n_{jk}, are presented in Table 3.

The total sample variation of the responses for the intensity of chlorine odor is $\widehat{V}_{\mathrm{T}} = 0.366$, and for sulfurous odor, it is $\widehat{V}_{\mathrm{T}} = 0.345$ with $df_{\mathrm{T}} = 89$ by Eq. (3). The between-laboratory variation for the intensity of chlorine odor is $\widehat{C}_{\mathrm{B}} = 0.256$, and for sulfurous odor, it is $\widehat{C}_{\mathrm{B}} = 0.250$ with $df_{\mathrm{B}} = 45$ by Eq. (6). The residual variation for the intensity of chlorine odor is $\widehat{V}_{\mathrm{W}} = 0.110$, while for sulfurous odor, it is $\widehat{V}_{\mathrm{W}} = 0.096$ with $df_{\mathrm{W}} = 44$ by Eq. (7).

The significance index of the laboratory factor $\widehat{SI}_{X1} = 1.360$ by Eq. (13) for the chlorine odor intensity exceeds its critical value of 1.185 at 95 % level of confidence; similarly for the sulfurous odor intensity, $\widehat{SI}_{X1} = 1.454$ exceeds its critical value of 1.202. At the same time, the significance index of the temperature factor does not exceed its critical value at 95 % level of confidence for both the chlorine odor intensity ($\widehat{SI}_{X2} = 2.423 < 3.010$) and the sulfurous odor intensity ($\widehat{SI}_{X2} = 0.511 < 3.248$). This means rejecting the null hypothesis H_0 concerning the (zero) difference between laboratories in classifying chlorine or sulfurous odor intensity by categories/levels: this difference is statistically significant at the α-risk, i.e., the probability of a Type I error, $\alpha = 0.05$. The effect of temperature in classifying chlorine or sulfurous odor intensity by categories is not significant as H_0 was not rejected at the same α-risk. Note that this effect might depend on the odorant concentration in water.[26]

A-3-4 Evaluation of the consensus

The proposed algorithm using random MC draws from a multinomial distribution was applied to evaluate the consensus of the laboratories at the given $\alpha = 0.05$[48]. The medium size effect used for the power calculations was assumed equal to $w = 0.3$, hence $\lambda = w^2N = 8.10$. Critical values of the significance indices are $SI_{X1}^{\mathrm{crit}} = 1.18$ and $SI_{X2}^{\mathrm{crit}} = 3.06$ at 95 % level of confidence. The PDFs of the significance index $\widehat{SI}_{Xl}^{\mathrm{MC}}$ under hypothesis H_0, and of the index $\widehat{SI}_{Xl,\lambda}^{\mathrm{MC}}$ modified under hypothesis H_1, for chlorine odor of the water samples are presented in Fig. 6. Plot (a) is related to factor $X1$, and plot (b) – to factor $X2$.

The blue line shows the PDF of $\widehat{SI}_{Xl}^{\mathrm{MC}}$, the red line is the PDF of $\widehat{SI}_{Xl,\lambda}^{\mathrm{MC}}$, the black vertical dashed line indicates the critical value SI_{Xl}^{crit}. Probabilities of a Type I error α and probabilities of a Type II error β are shown as in Fig. 5. 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 $SI_{X1}^{\mathrm{crit}} = 1.20$ and $SI_{X2}^{\mathrm{crit}} = 3.23$ 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 90 % level of confidence ($\alpha = 0.10$) 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

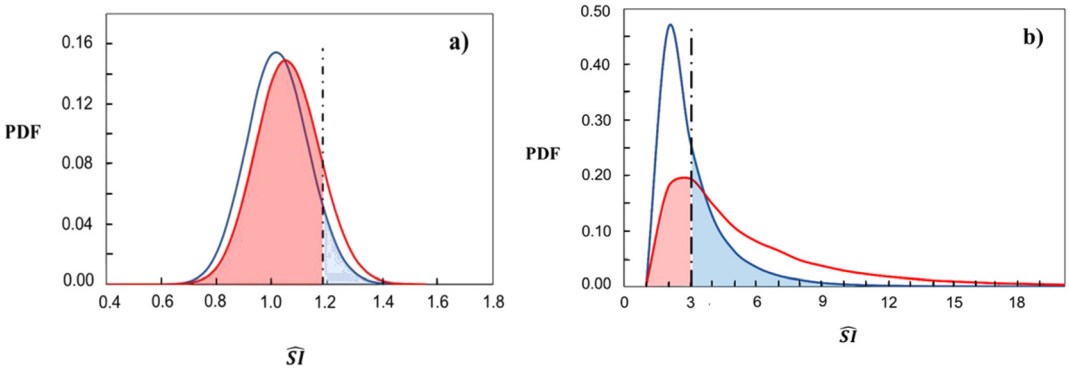

Fig. 6: PDFs of the significance index under hypothesis H_0 and of the index modified under hypothesis H_1 for chlorine odor of the drinking water samples. Plot (a) is related to factor $X1$, and plot (b) – to factor $X2$. Blue line shows the PDF of $\widehat{SI}_{XI}^{MC}$, red line – the PDF of $\widehat{SI}_{XI,\lambda}^{MC}$, black vertical dashed line indicated the critical value SI_{XI}^{crit} at 95 % level of confidence. Probabilities of a Type I error α and probabilities of a Type II error β are depicted as in Fig. 5.[48]

odor, they are $P_1 = 0.17$ and $P_2 = 0.39$. Increasing w also increases the power. Hence, at 90 % level of confidence 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 $P_2 = 0.66$[48].

Example 4. Multinomial ordered logistic regression of sensory responses to the quality of a sausage from different producers versus the chemical composition of the sausage

A-4-1 Introduction

The objective of the present example is 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.

A-4-2 Experimental

Samples of boiled-smoked sausage “Moscowskaya” by the national standard GOST R 55455[106] from $I = 16$ producers were purchased on the market practically simultaneously for the comparative testing of the sausage as a consumer product.[64] Its main chemical components are 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, named further “appearance”; 2) consistency; 3) color and appearance of cut sausage, named further “color”; 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.[27]

A-4-3 Implementation of two-way ORDANOVA without replication

Numbers of responses by experts and categories, i.e., frequencies n_{jk}, are shown in Table 4 for each quality characteristic of the “Moscowskaya” sausage.

Table 4: Numbers of responses to the quality of sausages by experts and categories, n_{jk}.

Category, k	Appearance			Consistency			Color			Taste			Smell		
	Experts ($X2$), j														
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	3	0	0	0	0	0	3	1	2	3	3	4	3	2	3
4	4	3	7	3	3	6	7	4	3	6	4	5	6	5	5
5	9	13	9	13	13	10	6	11	11	7	9	7	7	9	8
Total	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16

Table 5: Results of two-way ORDANOVA in the study of the "Moscowskaya" sausage by different producers.

Property	$\widehat{V}_T$	$\widehat{C}_B$	$\widehat{V}_W$	$X1$ & $X2$	$\widehat{C}^B_{X1}$ & $\widehat{C}^B_{X2}$	$\widehat{SI}_{X1}$ & $\widehat{SI}_{X2}$	df_{X1} & df_{X2}	SI^{crit}_{X1} & SI^{crit}_{X2}
Appearance	0.287	0.184	0.103	Producer	0.162	1.769	15	1.432
				Expert	0.022	1.801	2	2.686
Consistency	0.187	0.112	0.075	Producer	0.104	1.740	15	1.550
				Expert	0.008	0.980	2	2.892
Color	0.352	0.251	0.101	Producer	0.227	2.019	15	1.422
				Expert	0.024	1.621	2	2.554
Taste	0.414	0.293	0.121	Producer	0.289	2.186	15	1.403
				Expert	0.004	0.246	2	2.521
Smell	0.389	0.295	0.094	Producer	0.292	2.351	15	1.408
				Expert	0.004	0.210	2	2.510

Total variation $\widehat{V}_T$ of the responses by Eq. (4), partitioned into the between-producer variation $\widehat{C}_B$ by Eq. (5) and the within-producer residual variation $\widehat{V}_W$ by Eq. (6) are presented in Table 5, which includes the individual effects $\widehat{C}^B_{X1}$ and $\widehat{C}^B_{X2}$ of factors $X1$ and $X2$ (producers and experts, respectively) evaluated using the $\widehat{C}_B$ decomposition by Eqs. (7)–(9). To check the statistical significance of both the factor effects the significance indices $\widehat{SI}_{X1}$ and $\widehat{SI}_{X2}$ were calculated by Eq. (13) at the degrees of freedom $df_{X1} = 15$ and $df_{X2} = 2$. The critical index values SI^{crit}_{Xl} at 95 % level of confidence in Table 5 were obtained using two-way ORDANOVA tool for simulations of the $\widehat{SI}$ distributions.[48]

There is a statistically significant difference at 95 % level of confidence between the producers related to all the quality parameters of the sausage (appearance, consistency, color, taste, and smell). This difference is called the "inhomogeneity" of the producers. At the same time, the significance index values of the expert factor do not exceed its critical value at 95 % level of confidence, i.e., the null hypotheses H_0 on the "homogeneity" of expert responses regarding to each of the five sausage properties were not rejected.

A-4-4 Implementation of the multinomial ordered logistic regression

The multinomial ordered logistic regression model by Eq. (20) was fitted to the component contents in order to predict appearance, color, taste, and smell, assessed by experts according to the three categories shown in Table 4, k = 3, 4, and 5. A logistic regression for dichotomous (binary) outcome variables was used for prediction of consistency, since the corresponding expert responses in Table 4 were only of two categories, k = 4 and 5. Since for each categorical variable, the responses were found to be homogeneous among the three experts at the ORDANOVA study, all their outcomes were taken together, constituting the set of values to be used in the

Table 6: Statistics of the chemical composition of samples of the "Moscowskaya" sausage.

Statistics	Protein, c_1, %	Fat, c_2, %	Moisture, c_3, %	Salt, c_4, %
Minimum	13.7	19.9	53.5	2.2
Maximum	19.5	26.4	59.5	3.5
Mean	15.8	23.0	56.0	2.6
Standard deviation	1.4	4.6	1.9	0.3

regression. Intervals of the sausage main component contents in the certificates of the producers, taken into account in the regression, as well as the means and standard deviations of the contents (mass fraction expressed in %) are shown in Table 6. The calculation results are presented in Table 7, where the estimates for γ_{k0} and η_i coefficients are reported with their standard errors and 95 % confidence intervals (from 2.5 % to 97.5 % quantile). The estimated odds ratios, derived by exponentiating the coefficients, and the McFadden's pseudo-R^2 values by Eq. (22) are also shown in Table 7.[27]

For example, the model by Eq. (20) of category k=3 for appearance is $\text{logit}\,(P\,(q \le 3)) = 108.31 - 1.65c_1 - 0.95c_2 - 1.24c_3 + 2.22c_4$. A one-unit increase in the protein content c_1, for

Table 7: Results of the multinomial ordered logistic regression analysis.

Property	Coefficient	Value	Standard error	2.5 %	97.5 %	Odds ratio	Pseudo-R^2
	γ_{k0} (3|4)	108.31	0.01	108.30	108.32	1.09×10^{47}	
	γ_{k0} (4|5)	110.71	0.61	109.51	111.91	1.21×10^{48}	
Appearance	η_1	1.65	0.30	1.06	2.24	5.20	
	η_2	0.95	0.10	0.76	1.14	2.58	0.13
	η_3	1.24	0.07	1.10	1.38	3.44	
	η_4	−2.22	1.31	−4.78	0.34	0.11	
	γ_{k0}	−113.55	66.07	−243.05	15.95	4.85×10^{-50}	
Consistency	η_1	1.35	0.69	0.00	2.71	3.87	0.11
	η_2	1.08	0.56	−0.01	2.18	2.95	
	η_3	1.22	0.78	−0.30	2.74	3.40	
	γ_{k0} (3|4)	21.76	0.01	21.75	21.78	2.83×10^{9}	
	γ_{k0} (4|5)	23.56	0.44	22.69	24.42	1.70×10^{10}	
Color	η_1	0.56	0.25	0.07	1.05	1.75	
	η_2	−0.03	0.10	−0.23	0.17	0.97	0.09
	η_3	0.31	0.06	0.18	0.43	1.36	
	η_4	−0.51	1.28	−3.02	1.99	0.60	
Color*	γ_{k0} (3|4)	26.80	11.85	3.57	50.02	4.34×10^{11}	0.09
	γ_{k0} (4|5)	28.59	11.93	5.21	51.97	2.60×10^{12}	
	η_1	0.56	0.27	0.07	1.05	1.75	
	η_3	0.36	0.18	0.18	0.43	1.43	
Taste	γ_{k0} (3|4)	202.04	0.00	202.03	202.04	5.55×10^{87}	
	γ_{k0} (4|5)	204.47	0.55	203.39	205.55	6.31×10^{88}	
	η_1	2.87	0.30	2.28	3.46	17.61	0.28
	η_2	1.73	0.09	1.54	1.91	5.62	
	η_3	2.33	0.07	2.19	2.47	10.27	
	η_4	−4.42	1.51	−7.37	−1.47	0.01	
	γ_{k0} (3|4)	218.42	0.00	218.41	218.42	7.19×10^{94}	
Smell	γ_{k0} (4|5)	221.26	0.62	220.04	222.47	1.23×10^{96}	
	η_1	3.25	0.35	2.58	3.93	25.92	
	η_2	1.99	0.10	1.80	2.18	7.33	0.32
	η_3	2.41	0.08	2.26	2.57	11.18	
	η_4	−4.42	1.54	−7.43	−1.40	0.01	

*A shorted model of probability of the color category versus contents of protein and moisture.

example, corresponds to increase in the expected value of logit $(P(q \le 3))$ by 1.65 on the log odds scale, when all the other variables in the model are held constant.

The corresponding odds ratio $\exp(1.65) = 5.2$ indicates that, for every unit increase in the protein content, the odds of the sausage having a better appearance outcome (k = 4 or 5, versus k = 3) is multiplied 5.2 times.

Note, if a confidence interval does not cross zero, the parameter estimate is statistically significant. However, the confidence interval for the estimate η_4 = 2.22 (of the regression coefficient of the salt content c_4) crosses zero and this means that η_4 is statistically not significant here. In other words, the salt content values in the interval shown in Table 6 do not influence the probability of appearance category of a whole sausage.

Probabilities P of obtaining a response of category k by dependence on protein content c_1, calculated at the mean values of contents of other main components (Table 6), are shown on Fig. 7a. In the case of three categories of the observed responses (k = 3, 4, and 5), the probability $P(q = 3) = P(q \le 3), P(q = 4) = P(q \le 4) - P(q = 3)$, and $P(q = 5) = 1 - P(q = 3) - P(q = 4)$, where $P(q \le 3)$ and $P(q \le 4)$ can be evaluated by Eq. (21). The P values for k = 3, 4, and 5 are shown by lines 1, 2, and 3 in Fig. 7a, respectively. The P dependences on fat content c_2 and on moisture content c_3 are shown in Fig. 7b and c, respectively. They were also calculated at contents of other main components equal to their observed mean values in Table 6. The influence of all the component contents on the probability values is very similar, but the probability curves versus contents of protein and moisture are cut at the lower limits of the content intervals (their minimal values observed in the comparison). The full picture is shown in Fig. 7b for the probabilities versus contents of fat, where the probability values of the appearance categories vary from zero to the maximum, or from 1 to 0 and vice versa. Note that increasing c_1, c_2, and c_3 leads to increasing probability of the responses of the highest appearance category k = 5 (excellent quality). The salt content c_4 does not influence probabilities of responses of consistency dichotomous categories k = 4 and 5, hence it may be removed from the list of regressors. The probabilities of responses of different color categories do not depend on contents of fat and salt, c_2 and c_4, in their observed intervals. Model "Color*" without these two variables has the same McFadden's pseudo-R^2 value 0.09 as full model "Color", and practically the same regression coefficients for c_1 and c_3 in Table 7.

The full models for taste and smell are the best fitting models among the qualitative sausage properties: their McFadden's pseudo-R^2 values in Table 7 are about two to three times greater than those for appearance, color, and

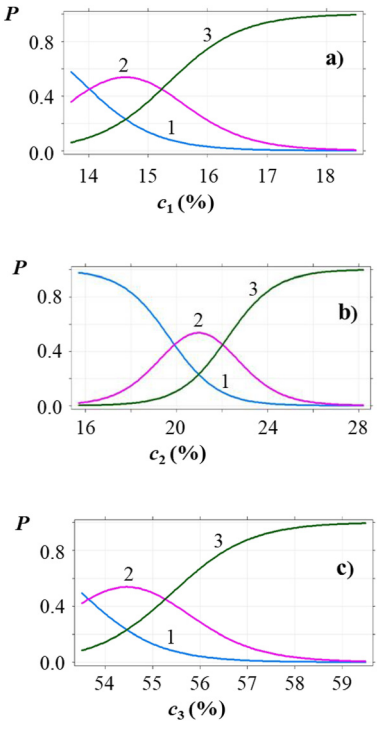

Fig. 7: Probabilities P of responses of different appearance categories in dependence on content of (a) protein c_1, (b) fat c_2, and (c) moisture c_3, mass fractions expressed in %. Each plot is calculated at contents of other main components equal to their observed mean values (Table 6). Line 1 is for category "satisfactory" (k = 3), line 2 for category "good" (k = 4), and line 3 for category "excellent" (k = 5).[27]

consistency. In general, the maximum probability of responses of each category of taste and smell is reached at increasing contents of the influencing main components. Similar effects are also observed in the plots in Fig. 7 for appearance: the first category reaching its maximum probability in the studied ranges of the component contents is 3, then 4, and finally 5, i.e., higher categories are more probable with greater contents of components. However, the salt contents in the interval considered in this study do not significantly influence responses on appearance, color, and consistency. At the same time, taste and smell are influenced by the salt contents in a reverse order than contents of other main components: the greater the salt content, the lower category is the more probable.

The probabilities of responses of the excellent quality category $P(q = 5)$ of both, taste and smell, increase with mass fractions of protein up to c_1 of about 17 %; of fat up to c_2 of about 26 %; and of moisture up to c_3 of about 58 %; while the minimal salt content c_4 = 2.2 % is preferable. These estimates could be helpful for a revision of the specification limits of the sausage composition, necessarily taking into account the mass balance constraint: the sum of actual values of mass fractions of the four main components expressed in % should be equal to or less than 100 %.[107]

Example 5. Comparison of the multisensory quality index values of a sausage of two producers

A-5-1 Introduction

This example demonstrates implementation of two-way ORDANOVA without replication for calculation of the multisensory multinomial quality index of a product, considering possible correlation of the responses to the different quality properties of the same product. It is shown how the index could be used for comparison of quality of this product from two producers.

A-5-2 Experiment

The dataset used here included results of sensory analysis and chemical analysis of the same boiled-smoked sausage "Moscowskaya"[106] as in Example 4. However, this time, the data were accumulated from each of two manufacturers during two years of the production. There were I_1 = 26 batches i = 1, 2, …, 26 of the first sausage manufacturer, named hereafter "producer 1", and I_2 = 54 batches i = 1, 2, …, 54 of the second manufacturer, named "producer 2". Five quality sensory properties of the sausage in a batch were examined without replication at each producer factory by its J = 5 experienced experts j = 1, 2, …, 5: a) appearance and packaging, named "appearance"; b) consistency; c) color and appearance of cut sausage, named "color"; 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, …, 5. A total $N_1 = I_1 \times J$ = 130 responses were obtained for each property, and hence 130 × 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 × 5 = 1,350 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 sausage chemical compositions.[28]

A-5-3 Statistics of the chemical composition

The intervals of mass fractions of the main sausage components expressed in % (protein c_1, fat c_2, moisture c_3, and salt c_4), minimum and maximum measured values, as well as the mean and standard deviations of the mass fractions of I_1 = 26 batches of producer 1 and I_2 = 54 batches of producer 2, are presented in Table 8.

Homogeneity of the two standard deviations (the null hypothesis of equality of variances) was tested using a two-sided Fisher's test at 95 % level of confidence and degrees of freedom for the numerator $I_1 - 1$ = 25, and $I_2 - 1$ = 53 for the denominator. The null hypothesis is rejected for fat, moisture, and salt, and not rejected for protein. Thus, chemical compositions of the sausages of the two producers were, in general, different. Because

Table 8: Statistics of the chemical composition of the sausage from each of two producers.

Producer	Statistic	Protein, c_1, %	Fat, c_2, %	Moisture, c_3, %	Salt, c_4, %
1	Minimum	13.71	27.14	50.11	1.85
	Maximum	15.60	31.02	56.74	2.47
	Mean	14.85	29.35	53.92	2.21
	Standard deviation	0.53	0.93	1.53	0.13
2	Minimum	16.34	20.69	48.45	1.50
	Maximum	20.40	31.24	57.98	2.80
	Mean	17.89	25.10	53.81	2.38
	Standard deviation	0.74	2.33	2.24	0.26

chemical composition may influence expert responses to the sausage quality properties as shown in Example 4, the ordinal data subsets of producer 1 and producer 2 were treated separately.

A-5-4 Implementation of two-way ORDANOVA without replication

Numbers of the responses (frequencies n_{jk}) are shown in Table 9 for each quality property of the sausage. The high categories of the quality properties of the product are to be expected: otherwise, a consumer would not buy the sausage in a store.

Total variation $\widehat{V}_{\mathrm{T}}$ of the responses without replication, partitioned into the between-batch variation $\widehat{C}_{\mathrm{B}}$ and the within-producer residual variation $\widehat{V}_{\mathrm{W}}$ are presented in Table 10, which includes the individual effects $\widehat{C}^{\mathrm{B}}_{X1}$ and $\widehat{C}^{\mathrm{B}}_{X2}$ of factors $X1$ and $X2$ (batches and experts, respectively) evaluated using the $\widehat{C}_{\mathrm{B}}$ decomposition, Eqs. (7)–(9). To check the statistical significance of effects of each factor, the significance indices $\widehat{SI}_{X1}$ and $\widehat{SI}_{X2}$ were calculated

Table 9: Numbers of the responses by experts and categories n_{jk} from each of two producers.

Property	Expert, j	Producer 1, category k					Producer 2, category k				
		1	2	3	4	5	1	2	3	4	5
Appearance	1	0	0	0	4	22	0	0	0	0	54
	2	0	0	0	3	23	0	0	0	5	49
	3	0	0	0	2	24	0	0	0	0	54
	4	0	0	0	0	26	0	0	0	3	51
	5	0	0	0	0	26	0	0	0	1	53
Consistency	1	0	0	1	12	13	0	0	0	2	52
	2	0	0	1	8	17	0	0	0	7	47
	3	0	0	0	10	16	0	0	0	1	53
	4	0	0	0	7	19	0	0	0	3	51
	5	0	0	0	10	16	0	0	0	2	52
Color	1	0	0	0	19	7	0	0	0	3	51
	2	0	0	0	10	16	0	0	0	9	45
	3	0	0	0	13	13	0	0	0	4	50
	4	0	0	0	9	17	0	0	0	3	51
	5	0	0	2	11	13	0	0	0	5	49
Taste	1	0	0	1	11	14	0	0	0	4	50
	2	0	0	1	8	17	0	0	0	8	46
	3	0	0	0	8	18	0	0	0	4	50
	4	0	0	1	5	20	0	0	0	7	47
	5	0	0	0	6	20	0	0	0	7	47
Smell	1	0	0	0	3	23	0	0	0	2	52
	2	0	0	0	2	24	0	0	0	7	47
	3	0	0	0	3	23	0	0	0	2	52
	4	0	0	0	1	25	0	0	0	3	51
	5	0	0	0	3	23	0	0	0	5	49

Table 10: Results of two-way ORDANOVA of the responses to the quality characteristics of the sausage from each of two producers during two years of its production.

Property	$\widehat{V}_T$	$\widehat{C}_B$	$\widehat{V}_W$	X1 & X2	$\widehat{C}^B_{X1}$ & $\widehat{C}^B_{X2}$	$\widehat{SI}_{X1}$ & $\widehat{SI}_{X2}$	df_{X1} & df_{X2}	SI^{crit}_{X1} & SI^{crit}_{X2}
Producer 1								
Appearance	0.064	0.031	0.033	Batch	0.027	2.203	25	1.464
				Expert	0.004	1.895	4	2.192
Consistency	0.250	0.121	0.129	Batch	0.115	2.366	25	1.416
				Expert	0.006	0.763	4	2.268
Color	0.265	0.152	0.113	Batch	0.133	2.585	25	1.429
				Expert	0.019	2.304	4	2.290
Taste	0.238	0.123	0.115	Batch	0.115	2.497	25	1.406
				Expert	0.008	1.040	4	2.189
Smell	0.084	0.039	0.045	Batch	0.038	2.318	25	1.455
				Expert	0.001	0.364	4	2.316
Producer 2								
Appearance	0.032	0.013	0.019	Batch	0.012	1.809	53	1.342
				Expert	0.001	2.691	4	2.293
Consistency	0.053	0.020	0.033	Batch	0.018	1.779	53	1.312
				Expert	0.002	1.934	4	2.283
Color	0.081	0.038	0.043	Batch	0.036	2.290	53	1.302
				Expert	0.002	1.412	4	2.293
Taste	0.099	0.069	0.030	Batch	0.068	3.477	53	1.303
				Expert	0.001	0.654	4	2.321
Smell	0.065	0.046	0.019	Batch	0.045	3.466	53	1.312
				Expert	0.001	1.326	4	2.287

by Eq. (13), with degrees of freedom df_{X1} = 25 for producer 1 and df_{X1} = 53 for producer 2, and df_{X2} = 4 for each of them. The critical index values SI^{crit}_{Xl} at 95 % level of confidence in Table 10 were obtained using two-way ORDANOVA tool for simulations of the $\widehat{SI}$ distributions [48] as in Example 4.

There are significant differences between the studied batches of producer 1 as $\widehat{SI}_{X1}$ are considerably larger than corresponding SI^{crit}_{X1}.

However, there is no statistically significant difference at 95 % level of confidence between the experts' responses related to appearance, consistency, taste, and smell. The $\widehat{SI}_{X2}$ value for color (2.304) exceeded the critical value SI^{crit}_{X2} = 2.290 at 95 % level of confidence, but did not exceed SI^{crit}_{X2} = 3.079 at 99 % level of confidence. Thus, the responses of the five experts of producer 1 are considered as homogeneous/uniform. For producer 2, as it was for producer 1, there is a statistically significant difference between the studied batches related to each of the five quality properties of the sausage at 95 % level of confidence.

The values of the significance index $\widehat{SI}_{X2}$ of the differences of the experts' responses regarding consistency, color, taste, and smell do not exceed their critical values SI^{crit}_{X2} at 95 % level of confidence, i.e., the null hypotheses on the homogeneity of the experts' responses are not rejected. The $\widehat{SI}_{X2}$ value for appearance (2.691) exceeds SI^{crit}_{X2} = 2.293 at 95 % level of confidence but does not exceed SI^{crit}_{X2} = 3.142 at 99 % level of confidence. In other words, the responses of the five experts of producer 2 can be assumed uniform.

The homogeneity of the responses of the five experts of each producer allows the use of the subset of each producer's data for calculation of its sausage multinomial multisensory quality index.

A-5-5 Testing correlation of the responses to the different quality properties

Spearman's rho correlation coefficient, calculated with the IBM SPSS software,[108] is used here as a nonparametric measure of the strength and direction of association that exists between responses to two quality properties as

Table 11: Spearman's rho correlation coefficients for expert evaluation of quality properties of sausages from producer 1 and producer 2.

Properties	Appearance	Consistency	Color	Taste	Smell
Producer 1					
Appearance	1.000	−0.027	0.207	−0.121	0.018
Consistency	−0.027	1.000	0.118	0.088	0.022
Color	0.207	0.118	1.000	0.037	0.052
Taste	−0.121	0.088	0.037	1.000	−0.049
Smell	0.018	0.022	0.052	−0.049	1.000
Producer 2					
Appearance	1.000	0.585	0.522	0.394	0.514
Consistency	0.585	1.000	0.549	0.480	0.439
Color	0.522	0.549	1.000	0.387	0.372
Taste	0.394	0.480	0.387	1.000	0.778
Smell	0.514	0.439	0.372	0.778	1.000

ordinal variables. The association is complete (the variables are strongly correlated), when the coefficient value achieves ±1, and the association is absent when the coefficient value is zero. The matrices of Spearman's rho correlation coefficients for $N_1 = 130$ pairs of responses to quality properties of the sausage of producer 1 (for each pair of the five properties), and similar for $N_2 = 270$ pairs related to the sausage of producer 2, are presented in Table 11.

The SPSS software output contains also the two-tailed significance probability of making the wrong decision on correlation of the ordinal variables when the null hypothesis on their uncorrelation is true (the probability α of a Type I error). From these calculations for producer 1, the null hypothesis was not rejected at 95 % level of confidence ($\alpha = 5$ %) and correlation was not supported for responses to all the quality properties apart of the pair "appearance-color". For this pair, the null hypothesis was not rejected at 99 % level of confidence ($\alpha = 1$ %). For producer 2, the calculated correlation coefficients in Table 11 are considerably larger than for producer 1. The negligible probability of a Type 1 error here ($\alpha = 0$ %) for any pair of tested quality properties indicating a high degree of correlation.

The reasons for the different correlation matrices of responses to the properties of the same sausage of two producers were not studied in this work. However, they are responses of two different groups of experts, each group employed by one producer for the sensory examination of its product.

Note that independent variables are necessarily uncorrelated. Therefore, the multinomial multisensory quality index was calculated for the sausage of producer 1 as a case study of independent responses to the quality properties, while the index for the sausage of producer 2 was evaluated considering that the responses to its quality properties were correlated.

A-5-6 Calculation of the quality index for independent responses

Table 12 gives the vector of frequencies by categories $\boldsymbol{n} = (n_1, n_2, \ldots, n_5)$, $n_k = \sum_{j=1}^{5} n_{jk}$, $k = 1, 2, \ldots, 5$, where frequencies n_{jk} of the homogeneous responses of the five experts of producer 1 are taken from Table 9; and the vector $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$ of relative frequencies $\hat{p}_k = n_k/N_1$ at $N_1 = 130$. The estimate of the probability $\hat{P}_{\text{app}}$ for appearance by Eq. (1), and similarly for the other quality properties, is also shown in Table 12. Hence, the quality index value of sausages from producer 1 by Eq. (25) is $Q_{\text{index1}} = 7.0$.

When the joint probability is considered that the ideal sausage (having category "excellent" for each property) will be produced, then $Q_{\text{index1}}^{\text{exc}} = 95$. This is practically impossible as the joint probability value in this case is $1.28 \cdot 10^{-96}$, i.e., such a perfect-quality level is not achievable under the considered production conditions.

Table 12: Statistics for calculation of the quality index Q_{index1} of the sausage from producer 1.

Statistic	Category, k	Appearance	Consistency	Color	Taste	Smell
n_k	1	0	0	0	0	0
	2	0	0	0	0	0
	3	0	2	2	3	0
	4	9	47	62	38	12
	5	121	81	66	89	118
$\widehat{p}_k$	1	0	0	0	0	0
	2	0	0	0	0	0
	3	0	0.0154	0.0154	0.0231	0
	4	0.0692	0.3615	0.4769	0.2923	0.0923
	5	0.9308	0.6231	0.5077	0.6846	0.9077
$\widehat{p}$		0.1366	0.0199	0.0192	0.0175	0.1200

A-5-7 Calculation of the quality index for correlated responses

The GenOrd package implementing a Gaussian copula-based procedure[109] was applied for simulation of the discrete multivariate random variables with the given correlation matrix and marginal distributions. A large number (10^6) of samples of $N_2 = 270$ occurrences of the multivariate quantities (expert responses to appearance, consistency, color, taste, and smell) were simulated by R function "ordsample", taking into account the experimental marginal distributions for appearance $\widehat{F}_{k\_\text{app}}$, consistency $\widehat{F}_{k\_\text{con}}$, color $\widehat{F}_{k\_\text{col}}$, taste $\widehat{F}_{k\_\text{tas}}$, and smell $\widehat{F}_{k\_\text{sme}}$, and the Spearman correlation matrix from the dataset of producer 2 in Table 11. The calculation code in R programming environment is presented in the paper.[28]

The numerical estimate of the joint probability $\widehat{P}_{\text{joint}}$ was obtained as the number of occurrences, among 10^6 samples, in which the intersection event $(\{\boldsymbol{Y}_{\text{app}} = \boldsymbol{n}_{\text{app}}\} \cap \{\boldsymbol{Y}_{\text{con}} = \boldsymbol{n}_{\text{con}}\} \cap \{\boldsymbol{Y}_{\text{col}} = \boldsymbol{n}_{\text{col}}\} \cap \{\boldsymbol{Y}_{\text{tas}} = \boldsymbol{n}_{\text{tas}}\} \cap \{\boldsymbol{Y}_{\text{sme}} = \boldsymbol{n}_{\text{sme}}\})$ was realized. The negative common logarithm of the joint probability estimate, equal to the multinomial multisensory quality index by its definition, was $Q_{\text{index2}} = 4.6$. When repeating the simulations, the standard deviation of the calculated index due to the simulation variability (reproducibility of the procedure) was about 0.1. To assess the correlation influence, a diagonal correlation matrix was applied to the same dataset of producer 2, i.e. the correlation was ignored. The corresponding quality index value was $Q_{\text{index2}} = 5.0$. Thus, the correlation effect is perceptible here: if correlation is not considered, the product (sausage) quality is assessed worse than it actually is.

Note that the minor correlation detected in the dataset of producer 1 (Table 11) for the pair "appearance-color" was assumed negligible for simplicity. To check if this assumption was sustainable, the original (experimental) correlation matrix was also applied. The corresponding quality index value was $Q_{\text{index1}} = 6.4$. Again, this value is not significantly different from the value $Q_{\text{index1}} = 6.7$ when the diagonal correlation matrix is used, as both values were obtained with the same simulation procedure having the same standard deviation of about 0.1. Thus, the assumption of the independent expert responses on quality properties of the sausage of producer 1 is supported.

In general, the multinomial multisensory quality index value of producer 2 is less than that of producer 1 by about two units, i.e., the probability of the joint event differs by about two orders of magnitude. Therefore, the quality of the sausage manufactured by producer 2 is considered better in the sense that deviations of this sausage quality properties from the claimed ones are less probable.

When the joint probability is considered that the ideal sausage having a category "excellent" for each property will be produced, the quality index $Q^{\text{exc}}_{\text{index2}}$ tends to become infinity (is greater than $Q^{\text{exc}}_{\text{index1}} = 95$), as the joint probability tends to become zero. Anyway, such sausage quality is unlikely for either of the producers under the studied conditions.

Membership of sponsoring bodies

The present membership of the IUPAC Analytical Chemistry Division (V) is

***President*:** D. Craston (UK); ***Past President*:** D. Shaw (USA); ***Secretary*:** L. Torsi (Italy); ***Titular Members*:** J. Barek (Czech), R. Burks (USA), F. Emmerling (Germany), H. Li (China), V. Peterson (Australia), S. Ražić (Serbia), A. Tintaru (France); ***Associate Members*:** E.M.M. Flores (Brazil), I. Leito (Estonia), F. Pitcitelli (Italy), T. Pradeep (India), M. Ramalingam (Malaysia), T. Takeuchi (Japan); ***National Representatives*:** R. Apak (Turkey), V. Baranovskaya (Russia), P. Forbes (South Africa), Gábor Galbács (Hungary), N. Galic (Croatia), P. Jarujamrus (Thailand), I. Kuselman (Israel), M. Piston (Uruguay), D. van Oevelen (Netherlands), S.K. Wiedmer (Finland); ***Emeritus Fellows*:** D.B. Hibbert (Australia), J. Labuda (Slovakia), M.C.F. Magalhães (Portugal).

The present membership of the IUPAC Subcommittee on Metrology in Chemistry is

Chair*:** D.B. Hibbert (Australia); ***Members: E.M.M. Flores (Brazil), J. Meija (Canada), Z. Mester (Canada), H. Li (China), I. Leito (Estonia), S.K. Wiedmer (Finland), S.K. Aggrawal (India), F.R. Pennecchi (Italy), I. Kuselman (Israel), M. Ramalingam (Malaysia), M.F. Camões (Portugal), R.J.N.B. da Silva (Portugal), V. Baranovskaya (Russia), A. Botha (South Africa), D. Craston (UK), S.L.R. Ellison (UK), D. Shaw (USA).

The present membership of the Cooperation of International Traceability in Analytical Chemistry (CITAC) is

***Chair*:** Z. Mester (Canada); ***Past Chair*:** B. Guetler (Germany); ***Vice-Chair*:** R.J.N.B. da Silva (Portugal); ***Secretary*:** F.R. Lourenço (Brazil); ***Members*:** A. Squirrell (Australia), M. Horsky (Austria), W. Wegscheider, deceased (Austria), O. P. de Oliveira Junior (Brazil), V. Ponçano (Brazil), T.R.L. Dadamos (Brazil), J.E.S. Sarkis (Brazil), H. Li (China), T. Näykki (Finland), P. Fisicaro (France), S.G. Walch (Germany), I. Papadakis (Greece), D.W.M. Sin (Hong Kong, P. R. China), P. K. Gupta (India), M. Nabi (Iran), M. Walsh (Ireland), I. Kuselman (Israel), F.R. Pennecchi (Italy), M. Sega (Italy), T. Fujimoto (Japan), R.B. Khousam (Lebanon), O. Zakaria (Malaysia), Y.M. Nakanishi (Mexico), L. Samuel (New Zealand), V. Baranovskaya (Russia), N. Oganyan (Russia), T.L. Teo (Singapore), A. Botha (South Africa), M. Obkircher (Switzerland), S. Wunderli (Switzerland), R. Kaarls (the Netherlands), R.J.C. Brown (UK), S.L.R. Ellison (UK), V. Iyengar (USA), J. D. Messman (USA).

The membership of the Task Group is

Chair*:** I. Kuselman (Israel); ***Members: P.S. Cheow (Singapore, project 2021-017-2-500), A. Botha (South Africa, project 2023-016-1-500), T. Gadrich (Israel), D. B. Hibbert (Australia), F.R. Pennecchi (Italy), A.A. Semenova (Russia).

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. ISO 5725-1:2023. *Accuracy (Trueness and Precision) of Measurement Methods and Results — Part 1: General Principles and Definitions*; The International Organization for Standardization: Geneva, 2023.
4. Magnusson, B.; Örnemark, U., Eds. In *The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics*, EURACHEM Guide, 2nd ed., 2014. Available from: http://www.eurachem.org.
5. International Bureau of Weights and Measures (BIPM). *The BIPM Key Comparison Database (KCDB)*: Sèvres, France, 2024. Available from: https://www.bipm.org/kcdb/.
6. Consultative Committee for Amount of Substance: Metrology in Chemistry and Biology (CCQM). *Estimation of a Consensus KCRV and Associated Degrees of Equivalence*. CCQM Guidance note; International Bureau of Weights and Measures (BIPM): Sèvres, France, 2013. Available from: https://www.bipm.org/documents/20126/28430045/working-document-ID-5794/49d366bc-295f-18ca-c4d3-d68aa54077b5.
7. Ellison, S. L. R. Consistency Plots: A Simple Graphical Tool for Investigating Agreement in Key Comparisons. *Accredit. Qual. Assur.* **2022**, *27*, 341–348. https://doi.org/10.1007/s00769-022-01520-z.
8. Koepke, A.; Lafarge, T.; Possolo, A.; Toman, B. Consensus Building for Interlaboratory Studies, Key Comparisons, and Meta-analysis. *Metrologia* **2017**, *54*, S34–S62. https://doi.org/10.1088/1681-7575/aa6c0e.
9. Possolo, A. Interlaboratory Consensus Building Challenge. *Anal. Bioanal. Chem.* **2020**, *412*, 3955–3956. https://doi.org/10.1007/s00216-020-02695-5.

10. Possolo, A. Solution to Interlaboratory Consensus Building Challenge. *Anal. Bioanal. Chem.* **2021**, *413*, 3–5. https://doi.org/10.1007/s00216-020-03053-1.
11. Tutmez, B. Relative Uncertainty-based Bayesian Interlaboratory Consensus Building. *Sci. Total Environ.* **2023**, *870*, 161977. https://doi.org/10.1016/j.scitotenv.2023.161977.
12. Bodnar, O.; Bodnar, T. Bayesian Estimation in Multivariate Inter-laboratory Studies with Unknown Covariance Matrices. *Metrologia* **2023**, *60*, 054003. https://doi.org/10.1088/1681-7575/acee03.
13. ISO 13528:2022. *Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparison*; International Organization for Standardization: Geneva, 2022.
14. Thompson, M.; Ellison, S. L. R. Dark Uncertainty. *Accredit. Qual. Assur.* **2011**, *16*, 483–487. https://doi.org/10.1007/s00769-011-0803-0.
15. Thompson, M. 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. Geoanalytical. Res.* **2017**, *42*, 12195–96. https://doi.org/10.1111/ggr.12195.
16. ISO Guide 33405:2024. *Reference Materials – Approaches for Characterization and Assessment of Homogeneity and Stability*; International Organization for Standardization: Geneva, 2024.
17. ISO 5725-2:2019. *Accuracy (Trueness and Precision) of Measurement Methods and Results — Part 2: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method*; International Organization for Standardization: Geneva, 2019.
18. Merkatas, C.; Toman, B.; Possolo, A.; Schlamminger, S. Shades of Dark Uncertainty and Consensus Value for the Newtonian Constant of Gravitation. *Metrologia* **2019**, *56*, 054001. https://doi.org/10.1088/1681-7575/ab3365.
19. Hodges, J. T.; Viallon, J.; Brewer, P. J.; Drouin, B. J.; Gorshelev, V.; Janssen, C.; Lee, S.; Possolo, A.; Smith, M. A. H.; Walden, J.; Wielgosz, R. I. Recommendation of a Consensus Value of the Ozone Absorption Cross-section at 253.65 nm Based on a Literature Review. *Metrologia* **2019**, *56*, 034001. https://doi.org/10.1088/1681-7575/ab0bdd.
20. Jackson, D.; Bowden, J.; Baker, R. 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* **2010**, *140*, 961–970. https://doi.org/10.1016/j.jspi.2009.09.017.
21. Hoffman, J. I. E. In *Basic Biostatistics for Medical and Biomedical Practitioners*; Academic Press, 2019, 2nd ed.; pp. 621–629. Chapter 36 – Meta-Analysis.
22. Agresti, A. *Categorical Data Analysis*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, 2013.
23. JCGM 200:2012. *International Vocabulary of Metrology — Basic and General Concepts and Associated Terms (VIM)*, 3rd ed.; Join Committee for Guides in Metrology (JCGM), International Bureau of Weights and Measures (BIPM): Sèvres, France, 2012. Available from: http://www.bipm.org/en/publications/guides/.
24. Hibbert, D. B.; Korte, E.-H.; Örnemark, U. Fundamental and Metrological Concepts in Analytical Chemistry (IUPAC Recommendations 2021). *Pure Appl. Chem.* **2021**, *93*, 997. https://doi.org/10.1515/pac-2019-0819.
25. Gadrich, T.; Kuselman, I.; Andrić, I. Macroscopic Examination of Welds: Interlaboratory Comparison of Nominal Data. *SN Appl. Sci.* **2020**, *2*, 2168. https://doi.org/10.1007/s42452-020-03907-4.
26. Gadrich, T.; Kuselman, I.; Pennecchi, F. R.; Hibbert, D. B.; Semenova, A. A.; Cheow, P. S.; Naidenko, V. N. Interlaboratory Comparison of the Intensity of Drinking Water Odor and Taste by Two-way Ordinal Analysis of Variation Without Replication. *J. Water. Health* **2022**, *20*, 1005–1016. https://doi.org/10.2166/wh.2022.060.
27. Gadrich, T.; Pennecchi, F. R.; Kuselman, I.; Hibbert, D. B.; Semenova, A. A.; Cheow, P. S. 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**, *2022*, 4181460–12. https://doi.org/10.1155/2022/4181460.
28. Gadrich, T.; Pennecchi, F. R.; Kuselman, I.; Hibbert, D. B.; Semenova, A. A.; Salikova, M. A Novel Multisensory Quality Index of a Food Product: An Analysis of a Sausage Properties. *Chemometr. Intell. Lab. Syst.* **2023**, *237C*, 104815. https://doi.org/10.1016/j.chemolab.2023.104815.
29. da Silva, R. B.; Ellison, S. L. R., Eds. In *Assessment of Performance and Uncertainty in Qualitative Chemical Analysis*; Eurachem/CITAC Guide, 2021. Available from: https://www.eurachem.org.
30. ISO 33406:2024. *Approaches for the Production of Reference Materials With Qualitative Properties*; International Organization for Standardization: Geneva, 2024.
31. ISO 6658:2017. *Sensory Analysis — Methodology — General Guidance*; International Organization for Standardization: Geneva, 2017.
32. Hibbert, D. B. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*; Elsevier, 2019; pp. 149–192. Chapter 4.07 – Chemometric Analysis of Sensory Data.
33. FisherJr, W. P.; Pendrill, L., Eds. In *Models, Measurement, and Metrology Extending the SI*; De Gruyter, 2024.
34. Bashkansky, E.; Gadrich, T.; Kuselman, I. Interlaboratory Comparison of Test Results of an Ordinal or Nominal Binary Property: Analysis of Variation. *Accredit. Qual. Assur.* **2012**, *17*, 239–243. https://doi.org/10.1007/s00769-011-0856-0.
35. Tiikkainen, U.; Ciaralli, L.; Laurent, C.; Obkircher, M.; Patriarca, M.; Robouch, P.; Sarkany, E. Is Harmonization of Performance Assessment in Non-quantitative Proficiency Testing Possible/Necessary? *Accredit. Qual. Assur.* **2022**, *27*, 1–8. https://doi.org/10.1007/s00769-021-01492-6.
36. Leik, R. K. A Measure of Ordinal Consensus. *Pac. Sociol. Rev* **1966**, *9*, 85–90. https://doi.org/10.2307/1388242.
37. Keyton, J.; Springston, J. Redefining Cohesiveness in Groups. *Small Group Res.* **1990**, *21*, 234–254. https://doi.org/10.1177/1046496490212006.

38. Alcalde-Unzu1, J.; Vorsatz, M. Do we Agree? Measuring the Cohesiveness of Preferences. *Theory Decis.* **2016**, *80*, 313–339. https://doi.org/10.1007/s11238-015-9494-z.
39. Tastle, W. J.; Wierman, M. J.; Dumdum, U. R. Ranking Ordinal Scales Using the Consensus Measure. *Issues Inf. Syst.* **2005**, *6*, 96. https://doi.org/10.48009/2_iis_2005_96-102.
40. Tastle, W. J.; Wierman, M. J. Consensus and Dissention: A Measure of Ordinal Dispersion. *Int. J. Approx. Reason.* **2007**, *45*, 531–545. https://doi.org/10.1016/j.ijar.2006.06.024.
41. Chiclana, F.; Tapia García, J. M.; del Moral, M. J.; Herrera-Viedma, E. A Statistical Comparative Study of Different Similarity Measures of Consensus in Group Decision Making. *J. Inf. Sci.* **2013**, *221*, 110. https://doi.org/10.1016/j.ins.2012.09.014.
42. Colley, R.; Grandi, U.; Hidalgo, C.; Macedo, M.; Navarrete, C. In *Proc. of the 32nd International Joint Conference on Artificial Intelligence Main Track*; Measuring and Controlling Divisiveness in Rank Aggregation, 2023; pp. 2616–2623.
43. Perez, I. J.; Cabrerizo, F. J.; Alonso, S.; Herrera-Viedma, E. A new Consensus Model for Group Decision Making Problems with Non-homogeneous Experts. *IEEE Trans. Syst. Man Cybern. Syst.* **2014**, *44*, 494–498. https://doi.org/10.1109/tsmc.2013.2259155.
44. Jakobsson, U.; Westergren, A. Statistical Methods for Assessing Agreement for Ordinal Data. *Scand. J. Caring Sci.* **2005**, *19*, 427–431. https://doi.org/10.1111/j.1471-6712.2005.00368.x.
45. Vituri, D. W.; Évora, Y. D. M. Reliability of Indicators of Nursing Care Quality: Testing Inter-examiner Agreement and Reliability. *Rev. Latino-Am. Enfermagem.* **2014**, *22*, 234–240. https://doi.org/10.1590/0104-1169.3262.2407.
46. Schnuerch, M.; Haaf, J. M.; Sarafoglou, A.; Rouder, J. N. Meaningful Comparisons with Ordinal-scale Items. *Collabra: Psychol.* **2022**, *8*, 11. https://doi.org/10.1525/collabra.38594.
47. Gadrich, T.; Marmor, Y. N. Two-way ORDANOVA: Analyzing Ordinal Variation in a Cross-balanced Design. *J. Stat. Plan. Inference.* **2021**, *215*, 330–343. https://doi.org/10.1016/j.jspi.2021.04.005.
48. Gadrich, T.; Marmor, Y. N.; Pennecchi, F. R.; Hibbert, D. B.; Semenova, A. A.; Kuselman, I. Power of a Test for Assessing Interlaboratory Consensus of Nominal and Ordinal Characteristics of a Substance, Material, or Object. *Metrologia* **2024**, *61*, 045004. https://doi.org/10.1088/1681-7575/ad5846.
49. Mittag, H.-J.; Rinne, H. *Statistical Methods of Quality Assurance*; Charman & Hall: London, 1993.
50. Hollebecq, L.-J. β-risk in Proficiency Testing in Relation to the Number of Participants. *Acta IMEKO* **2023**, *12* (1), 1–9. identifier: IMEKO-ACTA-12 (2023)-03-11; https://doi.org/10.21014/actaimeko.v12i3.1433.
51. Kuselman, I.; Fajgelj, A. IUPAC/CITAC Guide: Selection and Use of Proficiency Testing Schemes for a Limited Number of Participants – Chemical Analytical Laboratories. *Pure Appl. Chem.* **2010**, *82*, 1099–1135. https://doi.org/10.1351/PAC-REP-09-08-15.
52. Stepanov, A. V.; Chunovkina, A. G. On Testing of the Homogeneity of Variances for Two-side Power Distribution Family. *Accredit. Qual. Assur.* **2023**, *28*, 129–137. https://doi.org/10.1007/s00769-022-01525-8.
53. Jiménez-Gamero, I.; Analla, M. The Importance of Type II Error in Hypothesis Testing. *Int. J. Stat. Probab.* **2023**, *12*, 42. https://doi.org/10.5539/ijsp.v12n2p42.
54. ISO 3534-1:2006. *Statistics. Vocabulary and Symbols. Part 1: General Statistical Terms and Terms Used in Probability*; International Standard Organization: Geneva, 2006.
55. ISO 3534-2:2006. *Statistics. Vocabulary and Symbols. Part 2: Applied Statistics*; International Standard Organization: Geneva, 2006.
56. ISO 3534-3:2013. *Statistics. Vocabulary and Symbols. Part 3: Design of Experiments*; International Standard Organization: Geneva, 2013.
57. ISO 9000:2015. *Quality Measurement Systems – Fundamentals and Vocabulary*; International Standard Organization: Geneva, 2015.
58. Hibbert, D. B., Ed. In *Compendium of Terminology in Analytical Chemistry (IUPAC Orange Book)*; RSC, CPI Group Ltd: Croydon, UK, 2023.
59. ISO 704:2022. *Terminology Work – Principles and Methods*; International Organization for Standardization: Geneva, 2022.
60. Bertin, E. P. In *Introduction to X-Ray Spectrometric Analysis*; Springer: Boston, MA, 1978; pp. 255–278. Chapter – Qualitative and Semiquantitative Analysis.
61. Kreisberger, G.; Himmelsbach, M.; Buchberger, W.; Klampfl, C. W. Identification and Semi-quantitative Determination of Anti-oxidants in Lubricants Employing Thin-layer Chromatography-spray Mass Spectrometry. *J. Cromatogr. A.* **2015**, *1383*, 169–174. https://doi.org/10.1016/j.chroma.2015.01.048.
62. Zhai, K.; Zhang, B.; Zhu, L. A New Proposed Semi-quantitative Method for the Organic Additives Analysis in Traditional Lime Mortar. *J. Cult. Herrit.* **2023**, *62*, 284–292. https://doi.org/10.1016/j.culher.2023.06.003.
63. ISO 8586:2023. *Sensory Analysis – Selection and Training of Sensory Assessors*; International Organization for Standardization: Geneva, 2023.
64. ISO/IEC Guide 46:2017. *Comparative Testing of Consumer Products and Relative Services – General Principles*; International Organization for Standardization: Geneva, 2017.
65. NIST/SEMATECH. *e-Handbook of Statistical Methods*. Chapter – Multinomial PDF. Available from: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/multpdf.htm.
66. Gadrich, T.; Marmor, Y. N.; Bashkansky, E. Accuracy of Categorical Measurements: Nominal Scale. *Measurement* **2025**, *250*, 117044. https://doi.org/10.1016/j.measurement.2025.117044.
67. Anderson, R. J.; Landis, J. R. CATANOVA for Multidimensional Contingency Tables: Nominal-scale Response. *Comm. Statist. Theory Methods* **1980**, *9*, 1191. https://doi.org/10.1080/03610928008827952.
68. Gadrich, T.; Bashkansky, E.; Kuselman, I. Comparison of Biased and Unbiased Estimators of Variances of Qualitative and Semi-quantitative Results of Testing. *Accredit. Qual. Assur.* **2013**, *18*, 85–90. https://doi.org/10.1007/s00769-012-0939-6.

69. NIST/SEMATECH. In *e-Handbook of Statistical Methods*. Chapter – Chi-Square test for the variance. Available from: https://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm.
70. Owen, D. B. *Handbook of Statistical Tables*; Addison-Wesley: London, 1962.
71. Zaiontz, C. *Real Statistics Using Excel*. Chapter – Effect size for Chi-square test. Available from: https://real-statistics.com/chi-square-and-f-distributions/effect-size-chi-square/.
72. Zaiontz, C. *Real Statistics Using Excel*. Chapter – Power of Chi-square tests. Available from: https://real-statistics.com/chi-square-and-f-distributions/power-chi-square-tests/.
73. Champely, S. *The Comprehensive R Archive Network (CRAN)*. DescTools: Tools for Descriptive Statistics. R package documentation – Power calculations for ChiSquared tests, 2024. Available from: https://rdrr.io/cran/DescTools/man/power.chisq.test.html.
74. Murdoch, D.; Adler, D.; Nenadic, O.; Urbanek, S.; Chen, M.; Gebhardt, A.; Bolker, B.; Csardi, G.; Strzelecki, A.; Senger, A.; et al. *The Comprehensive R Archive Network (CRAN)*. Package rgl: 3D Visualization Using OpenGL, 2024. Available from: https://cran.r-project.org/web/packages/rgl/index.html.
75. Marmor, Y. N. *Research Areas – Factor Analysis Calculator Tool for Categorical Data*. Available from: https://w3.braude.ac.il/lecturer/dr-yariv-n-marmor/.
76. Vivar-Vera, M. d. A.; Pérez-Silva, A.; Ruiz-López, I. I.; Hernández-Cázares, A. S.; Solano-Barrera, S.; Ruiz-Espinosa, H.; Bernardino-Nicanor, A.; González-Cruz, L. Chemical, Physical and Sensory Properties of Vienna Sausages Formulated with a Starfruit Dietary Fiber Concentrate. *J. Food Sci. Technol.* **2018**, *55*, 3303–3313. https://doi.org/10.1007/s13197-018-3265-0.
77. Lazic, I. B.; Jovanovic, J.; Simunovic, S.; Raseta, M.; Trbovic, D.; Baltic, T.; Ciric, J. Evaluation of Sensory and Chemical Parameters of Fermented Sausages. *Meat Technol.* **2019**, *60*, 84. https://doi.org/10.18485/meattech.2019.60.2.2.
78. Yarali, E. Sensory Analysis in Meat and Meat Products. *Int. J. Agric. Sci.* **2023**, *8*, 27.
79. Schroeder, L. D.; Sjoquist, D. L.; Stephan, P. E. *Understanding Regression Analysis: An Introductory Guide*, 2nd ed.; SAGE Publications: Los Angeles, 2017.
80. Hastie, T. J.; Botha, J. L.; Schnitzler, C. M. Regression with an Ordered Categorical Response. *Stat. Med.* **1989**, *8*, 785–794. https://doi.org/10.1002/sim.4780080703.
81. Frees, E. W. *Regression Modelling with Actuarial and Financial Applications*; Cambridge University Press: New York, 2010.
82. Kaygisiz, F.; Bolat, B. A.; Bulut, D. Determining Factors Affecting Consumer's Decision to Purchase Organic Chicken Meat. *Braz. J. Poul. Sci.* **2019**, *12*, 1. https://doi.org/10.1590/1806-9061-2019-1060.
83. University of California, Los Angeles (UCLA). *Advanced Research Computing – Statistical Methods and Data Analysis*. Chapter – Ordinal logistic regression, 2022. Available from: https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/.
84. Brown, M. K. W. *Evaluating an Ordinal Output Using Data Modeling, Algorithmic Modeling, and Numerical Analysis*. Murray State Theses and Dissertations, Vol. 168, 2020. Available from: https://digitalcommons.murraystate.edu/etd/168/.
85. Venables, W. N.; Ripley, B. D. *MASS: Modern Applied Statistics with S*, 4th ed.; Springer: New York, 2002. Available from: http://www.stats.ox.ac.uk/pub/MASS4.
86. Fox, J.; Weisberg, S. *An R Companion to Applied Regression*, 3rd ed.; Sage: Los Angeles, 2019.
87. University of California, Los Angeles (UCLA). *Advanced Research Computing – Statistical Methods and Data Analysis*. Chapter – FAQ: What are pseudo R-squareds? 2022. Available from: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/.
88. Signorell, A.; Alfons, A.; Anderegg, N.; Aragon, T.; Arachchige, C.; Arppe, A.; Baddeley, A.; Barton, K. Bolker, B.; Borchers, H. W.; et al. *The Comprehensive R Archive Network (CRAN)*. DescTools: Tools for Descriptive Statistics. R package version 0.99.54, 2024. Available from: https://cran.r-project.org/package=DescTools.
89. Hilbe, J. M. *Logistic Regression Models*; Chapman & Hall/CRC Press: New York, 2009.
90. Hyldig, G.; Green-Petersen, D. M. B. Quality Index Method – An Objective Tool for Determination of Sensory Quality. *J. Aquat. Food. Prod. Technol.* **2004**, *13*, 71–80. https://doi.org/10.1300/J030v13n04_06.
91. Salami, S. A.; O'Grady, M. N.; Luciano, G.; Priolo, A.; McGee, M.; Moloney, A. P.; Kerry, J. P. Quality Indices and Sensory Attributes of Beef from Steers Offered Grass Silage and a Concentrate Supplement with Dried Citrus Pulp. *Meat Sci.* **2020**, *168*, 108181. https://doi.org/10.1016/j.meatsci.2020.108181.
92. Imm, B.-Y.; Lee, J. H.; Lee, S. H. Sensory Quality Index (SQI) for Commercial Food Products. *Food Qual. Prefer.* **2011**, *22*, 748–752. https://doi.org/10.1016/j.foodqual.2011.05.007.
93. Shi, Z.; Müller, H. J. Multisensory Perception and Action: Development, Decision-making, and Neural Mechanisms. *Front. Integr. Neurosci.* **2013**, *7*, 1. https://doi.org/10.3389/fnint.2013.00081.
94. Spence, C. Multisensory Perception. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience. Sensation, Perception and Attention*; Wiley, Vol. 2, 2018, 4th ed.; pp. 1–56. Chapter 14 – Multisensory perception.
95. Prescott, J. Multisensory Processes in Flavour Perception and Their Influence on Food Choice. *Curr. Opin. Food Sci.* **2015**, *3*, 47–52. https://doi.org/10.1016/j.cofs.2015.02.007.
96. Rumsey, D. *Probability for Dummies*; Wiley, 2006.
97. Pishro-Nik, H. Introduction to Probability. In *Statistics and Random Processes*; Kappa Research LLC, 2014.
98. Wang, Q. A. Probability Distribution and Entropy as a Measure of Uncertainty. *J. Phys. A: Math. Theor.* **2008**, *41*, 065004. https://doi.org/0.1088/1751-8113/41/6/065004.

99. Namdari, A.; Li, Z. (S.) A Review of Entropy Measures for Uncertainty Quantification of Stochastic Processes. *Adv. Mech. Eng.* **2019**, *11*, 1. https://doi.org/10.1177/1687814019857350.
100. Garanin, D. A.; Lukashevich, N. S.; Efimenko, S. V.; Chernorutsky, I. G.; Barykin, S. E.; Kazaryan, R.; Buniak, V.; Parfenov, A. Reduction of Uncertainty Using Adaptive Modeling Under Stochastic Criteria of Information Content. *Front. Appl. Math. Stat.* **2023**, *8*, 1. https://doi.org/10.3389/fams.2022.1092156.
101. Panagiotelis, A.; Czado, C.; Joe, H. Pair Copula Constructions for Multivariate Discrete Data. *J. Amer. Stat. Assoc.* **2012**, *107*, 1063–1072. https://doi.org/10.1080/01621459.2012.682850.
102. Ferrari, P. A.; Barbiero, A. Simulating Ordinal Data. *Multivariate Behav. Res.* **2012**, *47*, 566–589. https://doi.org/10.1080/00273171.2012.692630.
103. EURAMET Guide No. 4. *Guide on Comparisons*, Ver. 2.0. European Association of National Metrology Institutes, 2021. Available from: https://www.euramet.org/publications-media-centre/euramet-guides.
104. ISO 6520-1:2007. *Welding and Allied Processes – Classification of Geometric Imperfections in Metallic Materials. Part 1 – Fusion Welding*; International Organization for Standardization: Geneva, 2007.
105. GOST R 57164:2016. *Drinking Water. Methods for Determination of Odor, Taste, and Turbidity*; Russian Federal Agency for Technical Regulation and Metrology: Moscow, 2016. Available from: https://runorm.com/catalog/1004/876961/.
106. GOST R 55455:2013. *Boiled-smoked Meat Sausages. Specifications*; Russian Federal Agency for Technical Regulation and Metrology: Moscow, 2013. Available from: https://gostperevod.com/catalogsearch/result/?q=gost+55455-2013.
107. Pennecchi, F. R.; Kuselman, I.; Hibbert, D. B. IUPAC/CITAC Guide: Evaluation of Risks of False Decisions in Conformity Assessment of a Substance or Material with Mass Balance Constraint (IUPAC Technical Report). *Pure Appl. Chem.* **2023**, *95*, 1217–1254. https://doi.org/10.1515/pac-2022-0801.
108. International Business Machines Corporation. *IBM SPSS Software*, 2024. Available from: https://www.ibm.com/analytics/spss-statistics-software.
109. Barbiero, A.; Ferrari, P. A. Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions. *The Comprehensive R Archive Network (CRAN). Package "GenOrd"* **2015**. Available from: https://cran.r-project.org/web/packages/GenOrd/GenOrd.pdf.