A Statistical Agreement-Based Approach for Difference Testing

F. Calle-Alonso, C. J. Pérez

Faculty of Veterinary Medicine, Biostatistics Unit, University of Extremadura, Avda. de la Universidad s/n, 10003 Cáceres, Spain.

Abstract

A statistical approach based on measures of agreement is proposed for use in a sensory analysis context. This approach considers the idea of using statistical agreement to provide information on the homogeneity of the raters' responses, so that this information can then be used to discriminate between products. It can also be used to measure the expertise level of raters. Although the prime focus is on difference testing by the triangle test (ISO 4120:2008), the proposed methodology can also be applied in other contexts such as the paired comparison test (ISO 5495:2009) or the duo-trio test (ISO 10399:2010), among others. The proposed approach is not a substitute for binomial statistical analysis, but rather it can be used as a complement. It is especially useful when few panelists are available and replications are needed. An experiment that evaluates two types of Iberian dry-cured pork loins through the triangle test is performed to illustrate the applicability of the proposed approach.

Keywords: Binomial model, Beta-binomial model, Difference testing, Measures of agreement, Multiple raters, Sensory analysis, Triangle test.

1. Introduction

Sensory analysis can be used to provide subjective information about the acceptance of different products, and is also widely used in determining overall quality. As is well known, the use of a panel is a very important tool in attempting to describe a product's different and complex features. But it also has some drawbacks, subjectivity and low repeatability, for instance. In order to improve the reliability of the results and avoid these problems, some countries have enacted laws which give legal value to sensorial analysis tests and are aimed at homogenizing the results. The International Standardization Organization (ISO) has proposed a standard for sensory analysis to ensure that products and services are safe, reliable, and of good quality (see International Organization for Standarization (TC34/SC12)). This standard has been applied in many different fields: quality control, research and development, market research, protected designation of origin,... For example, the International Organization for Standarization (2004b) norm is generally used by government agencies to regulate all aspects of the triangle test.

As well as the triangle test (International Organization for Standarization (2004b)), others such as the paired comparison test (International Organization for Standarization (2005)) and the duo-trio test (International Organization for Standarization (2004a)) use the binomial distribution to discriminate between products. However, the binomial model is not suitable in situations in which there is overdispersion. An extension of the binomial model – the beta-binomial model – is used to fit overdispersed binomial data (see Ennis and Bi (1998)). The information provided by the binomial model can be complemented with that obtained with statistical measures of agreement.

Agreement among raters is of great importance for researchers and practitioners who describe and evaluate objects and behaviours in a number of fields, including the social and behavioural sciences. Fleiss et al. (1969) and Fleiss Fleiss (1971) were two landmark articles on agreement measures. Since then, statistical agreement has been an active research area whose techniques have been widely used in practice. The most popular measure of agreement is Cohen's kappa. There are, however, many others available, each one with its own particular characteristics that make it interesting to use in different contexts. Agresti (1996) presented several modeling techniques for the analysis of categorical data, in addition to an invaluable summary of the state-of-the-art. Von Eye and Mun (2005) provided a comprehensive reference book that analyses rater agreement from four different perspectives, including log-linear modeling.

Although statistical measures of agreement have been widely used in many fields of knowledge, especially in the biomedical sciences, they have remained almost unexplored in the field of sensory analysis. Nevertheless, a few interesting results can be found in the literature. For example, Wu and Chen (1995) considered the agreement among raters to evaluate the agreement of tea sensory data, and Mounchili et al. (2005) considered agreement in a sensory analysis of milk samples. In the present communication, a statistical agreement-based approach is presented for sensory analysis. This approach proposes the use of an efficient measure of agreement for two or more raters in which the response is given on a qualitative scale. It is shown how these measures can provide information about the process of seeking for differences. The idea is based on measuring the homogeneity of the raters' responses, and then using this information to analyse differences between products. The proposed approach is connected to the standard binomial procedure. Also, measures of agreement can be used to qualify novice raters' aptitudes, and mark when they become experts. Although the prime focus is on difference testing using the triangle test, the methodology can also be applied in other difference or similarity tests (see Bi (2011)).

There are many examples in the literature showing the importance of sensory analysis in terms of designing, testing, launching, and rethinking food products. For example, the characterization of dry-cured shoulder of pork (Lorenzo et al. (2008)), Iberian dry-cured ham (Martín et al. (2010)), pineapple juice (Silva et al. (2010)), and Gamonedo cheese (Ramos-Guajardo and González-Rodríguez (2011)). In the present study, the differences of two Iberian dry-cured pork loins are evaluated through a triangle test by using measures of agreement.

The paper is organized as follows. Section 2 presents a short discussion of the main measures of inter-rater agreement and their application to sensory analysis. Section 3 describes the agreement-based approach and connects it to the standard binomial procedure. Some illustrative examples are also presented. In Section 4, an experiment designed involving a triangle test illustrates the applicability of the proposed approach. Finally, the conclusions are presented in Section 5.

2. Measures of agreement

Currently there is no standard measure of agreement used by the scientific community, although Cohen's kappa has a long history of use as an index of inter-rater agreement. However, Cohen's kappa is not always the best choice (see, e.g., Gwet (2002) and Fletcher et al. (2011)). When two raters are involved, there is a wide range of available measures of agreement in the statistical literature. For more than two raters, the number of available measures is dramatically reduced because of the difficulty of interpreting the results. Also, the levels of agreement tend to decrease as the number of raters grows. In the following paragraphs, we shall describe the main measures of agreement from the perspective of the proposal of the present work.

Sometimes, measures of agreement can be affected by chance. When the raters are not sure about the correct classification of a product, some guessing may occur. Guessing may can be total, if they are not able to distinguish anything at all, or partial, if they are guessing only some of the samples. When two raters make their predictions by chance, they sometimes agree. The question is when such agreements should count towards a statistical index of agreement. Theoretically, if the agreement by chance can be estimated, then this effect could be removed from the total agreement to discover the true agreement among raters. This is what chance-corrected measures try to do, but it is not at all clear that the final conclusion is reliable. Indeed, sometimes these measures yield paradoxical and counter-intuitive results. The choice between chance-corrected or non-chance-corrected measures has been a topic of some debate (see, e.g., Guggenmoos-Holzmann (2006)). However, the best option is to use the measure of agreement that by definition and meaning best fits the nature of the problem being addressed, regardless of whether or not it includes corrections for chance.

Consider m raters and c alternatives on a categorical scale. For the triangle, paired comparison, duo-trio, 2-AFC, and 3-AFC tests, the alternatives are right (positive) or wrong (negative) responses, i.e., $c = 2$. For the sake of simplicity, we shall first consider the notation for two panelists ($m = 2$), to subsequently generalize it to $m \geq 2$. For a qualitative variable X ranging over $1, 2$ (positive and negative ratings, respectively), n_{ij} will denote the observed frequency for rater 1 giving the response $X = i$, and rater 2 giving the response $X = j$. The observed frequencies can be presented in a contingency table of dimension 2×2, or generally $m \times 2$. One has, of course, that $\sum_i \sum_j n_{ij} = n$. The proportion parameters, ρ_{ij}, are estimated from the observed proportions, i.e., $\hat{\rho}_{ij} = n_{ij}/n$.

The simplest non-chance-corrected measure is the proportion of overall agreement, defined as $\sum_i \rho_{ii}$ and estimated as $\sum_i n_{ii}/n$. The estimated value will be 0 when there is no agreement at all, and 1 when the agreement is absolute. This measure has been criticized, because it can be high even with hypothetical raters who randomly guess on each case with probabilities equal to the observed base rate. There are other non-chance-corrected measures, such as the Holley and Guildford G coefficient and the Rogot and Goldberg A_1 and A_2 indices (see Gwet (2002)). These measures were defined for only two raters, but they can be extended to three or more. They include the agreement observed for all the possible rater responses, i.e., for the right and wrong responses jointly. They are unable to distinguish between agreement on right responses and agreement on wrong ones. Since our interest is in discriminating products, it is important to know the agreement for the right and the wrong cases separately, and while G, A_1, and A_2 do provide some information, they are really unsuitable for the present discrimination context.

In order to address the problem of analysing the agreement based on only one specific response, there are some non-chance-corrected measures available for two raters – concordance proportion, Dice index, Goodman and Kruskal λ_r, and the Jaccard measure, among others (see Shoukri (2004)). The most interesting measure in this context is Dice index since it is easier to interpret than the others and leads to more realistic agreement values. Dice's index has been widely used in several fields of knowledge (see, e.g., Ajmone-Marsan et al. (2001) and LaPara et al. (2002)), but has been left practically unexplored in that of sensory analysis. One exception is the work of Mounchili et al. (2005) who applied it to the organoleptic analysis of milk samples.

The proposal that we shall describe in the following section is based on the use of positive and negative Dice indices to discriminate products and assess agreement. In the following paragraphs of this section, we shall present the main results for these indices.

For two raters, the Dice indices are defined as:

$$D_i^{(2)} = \frac{2\rho_{ii}}{\rho_i^{(1)} + \rho_i^{(2)}}, i = 1, 2,$$

where $\rho_i^{(1)}$ and $\rho_i^{(2)}$ are the marginal probabilities for each rater, i.e, $\rho_i^{(1)} = \rho_{i1} + \rho_{i2}$ and $\rho_i^{(2)} = \rho_{1i} + \rho_{2i}$. $D_1^{(2)}$ refers to the positive response and $D_2^{(2)}$ to the negative one. Both values are defined in the interval $[0, 1]$, taking the value 1 when there is total agreement for the i-th alternative, and 0 when there is no agreement at all for that alternative. Graham and Bull (1998) and Mackinnon (2000) used the delta method to derive formulas for the asymptotic standard errors of these specific response measures. Alternatively, the standard errors can be estimated by Jackknife or nonparametric bootstrap (see Severiano et al. (2011)) techniques. We have not found any closed expressions for sampling distributions of Dice's indices for hypothesis testing in the literature. However, simulation-based approaches can be used to estimate confidence intervals and to test hypotheses.

Dice already noted that the similarity measure proposed in an context of ecology in that work could be extended to three or more species. Warrens (2008) discussed this extension further in analysing similarity coefficients for binary data. The Dice index for more than two raters can be defined as:

$$D_i^{(m)} = \frac{m\rho_{ii\ldots i}}{\sum_{j=1}^{m} \rho_i^{(j)}}, \tag{1}$$

where $\rho_{ii\ldots i}$ is the proportion parameter for the case where all the raters have chosen the alternative i, and $\rho_i^{(j)}$ is the marginal probability that is obtained for rater j with alternative i. This generalized index maintains the same properties as the two-rater one. However, there are no longer any closed expressions for asymptotic standard errors such as there were in the two-rater case. Confidence intervals or hypothesis testing must be performed using simulation-based approaches such as Monte Carlo or resampling methods (see, e.g., Manly (1997)).

With respect to the chance-corrected measures, it is worth mentioning that they have traditionally been far more commonly used than the non-chance-corrected ones. They can be presented with the common expression:

$$M(I) = \frac{I_o - I_e}{1 - I_e},$$

where I_o and I_e are the observed and the expected values of the non-chance-corrected index of agreement, respectively.

The most extensively used measures of agreement are Cohen's kappa for two raters (Cohen (1960)) and Fleiss's generalized kappa for several raters (Fleiss (1971)). They have been applied to the estimation of conjoint agreement (agreement for all possible alternatives) in many fields of knowledge. In the specific field of sensory analysis, there are only a few published applications of agreement measures, and most of these use Cohen's kappa. Pons-Sanchez-Cascado et al. (2006) and Baixas-Nogueras et al. (2003) used this index to evaluate the agreement between two rejection methods for anchovies and hake, respectively, and Jenschke et al. (2007) used it to assess the agreement between panelists in a beef tasting experiment. Wu and Chen (1995) used the multi-rater agreement Kappa to evaluate the agreement of tea sensory data. Cohen's Kappa has also been proposed for use in a complementary way (see, e.g., Cicchetti and Feinstein (1990a)). A chance-corrected measure that could be used for discrimination testing is the conditional Kappa, although it does suffer from some drawbacks in that context. In the difference problem, it is unusual for the raters to try to guess – they might do so sometimes, but only in very few cases. Thus, the basic logic behind studying a chance-corrected measure such as the conditional kappa is inappropriate here. In addition, since both Dice indices are used, the overall agreement for the two possible responses is determined from the effect of the marginal proportions that are considered in the conditional kappa, so that there is no need to correct for possible effects of chance as has to be done in the conditional kappa. Note also that the agreement given by the conditional Kappa is in most case underestimated and that, to distinguish fair agreement, the number of agreed responses needs to be very high. Finally, chance-corrected measures may yield misleading values for binary ratings, such as in the problem to be addressed in the present work (see Guggenmoos-Holzmann (2006)). The generalized Dice index does not have these drawbacks, and thus provides a clear and realistic measure of the agreement that may be useful in discriminating products.

In the following section, we shall present the proposed approach and connect it to the standard procedures.

3. An agreement-based approach

3.1. Introduction

The problem of discriminating products is a special case studied in sensory analysis. The three standardized tests defined for this kind of problem are the triangle test, the paired test, and the duo-trio test. Other tests put forward in the literature are demonstrating high potential, for example, Tetrad (see Garcia et al. (2012)), A-Not A, 2-AFC, and 2-AFCR (see Van Hout et al. (2011)). Besides, alternative strategies have been recently considered, like Bayesian methodology used by Bi (2011) and Dubnicka (2013). As a consequence, differentiation tests is a very active research topic.

For discrimination problems, it is usual for there not to be many panelists available, and for a considerable number of observations to be required for the results to be to significant, this latter usually due to the smallness of the differences between the products. Meyners and Brockhoff (2003) showed that one might be able to add to the power of the test while reducing the number of raters by increasing the total number of assessments with replications. In particular, the test must be repeated several times to provide the necessary amount of information, and then results combined. The question of whether it is permissible to combine results from replicated triangular tests has been extensively discussed by various authors. According to Kunert and Meyners (1999), if the experiment is properly randomized and controlled then the assessments are will be independent and will have a binomially distributed success probability. But they also noted that it is difficult for these assumptions to be satisfied when replications are performed, and in such a case the choice probability and a measure of heterogeneity should be estimated. The binomial distribution assumes the existence of only one source of variability – that based on the samples. Therefore, when panelists are rating identically from one sample to another, the variance is completely explained by the binomial distribution. But rating identically for all replications is unusual, although it may sometimes be the case.

A general problem with discrimination testing is the assumption that all panelists have the same probability of discrimination, that there are only two kinds of raters – non-discriminators and perfect discriminators. The former type always guesses, and the latter always discriminates correctly through all the replications. This assumption is unrealistic, and it is evident that panelist variability needs to be taken into account when collecting replicated observations from the same panelists (see Ennis and Jesionka (2011)). To deal with this difficulty, a beta distribution can be used instead of a binomial to model variation in inter-trial choice probabilities.

The beta-binomial model considers the variability among samples as well as the variability among raters (also termed overdispersion), making it possible to combine responses across raters and replications. This increases the power of the test for a small panel size (see Anderson (1988)). The beta-binomial distribution is the natural extension of the binomial. It is based on the binomial with parameter p following a beta distribution with parameters a and b. It is useful to apply the re-parameterization $\mu = a/(a + b)$ and $\gamma = 1/(a + b + 1)$, which are the mean of the binomial parameter p and a scale parameter that measures its variation, respectively (see, e.g, Ennis and Bi (1998)). The scale parameter varies from 0 when there is no overdispersion to 1 when there is total overdispersion. Ennis and Bi (1998) provide hypothesis tests to evaluate whether the parameters differ significantly from the quantities of interest. With the proportion μ, one tests whether or not the differentiation was the result of guessing. In testing the overdispersion parameter, one may study whether the appropriate distribution is the binomial ($H_0 : \gamma = 0$) or the beta-binomial ($H_1 : \gamma \neq 0$). It is more important, however, to correctly estimate and interpret the parameters and their variances that apply to the problem at hand.

The results given by the binomial and beta-binomial models for the same problem are generally sufficiently different for different conclusions to be drawn regarding the products being tested. When the sensorial judgement of these products is fairly easy to do, it is generally advantageous to collect replicated data and analyse it using the beta-binomial model (see Ligget and Delwiche (2005)). Even so, the question of overdispersion should be considered, because the binomial model might be appropriate in some cases. For example, in a test of the sensory quality of cabbage, Radovich et al. (2004) found from their use of the beta-binomial model that overdispersion was not significant in their case, and that the binomial model was better suited to their problem.

The binomial procedure can be complemented with information obtained from an agreement-based approach. Specifically, the positive and negative Dice indices can be used to provide information on the homogeneity of the raters' responses. This can then be used to discriminate between products, and to provide complementary evidence on their differences. As a spin-off, the approach also provides information on the "quality" of the raters involved.

3.2. Using Dice's indices

Firstly in this subsection, we shall consider the relationship between Dice's indices and difference tests. For these latter, let p_0 be the guessing success probability, where $p_0 = 1/2$ for the 2-AFC, paired, and duo-trio methods, and $p_0 = 1/3$ for the 3-AFC and triangle methods. If all raters are guessing then the marginal proportions of success for the m independent raters are the same, i.e., $\rho_1^{(j)} = p_0, j = 1, 2, \ldots, m$. Therefore, the positive response Dice index is

$$D_1^{(m)} = \frac{m p_0^m}{\sum_{j=1}^m p_0} = p_0^{m-1}.$$

Then, the hypothesis tests

$$\begin{aligned} H_0 &: D_1^{(m)} \leq p_0^{m-1} \\ H_1 &: D_1^{(m)} > p_0^{m-1} \end{aligned} \qquad (2)$$

show whether the raters are discriminating the positive response more than would be expected by chance. Analogously, for the negative response Dice index, one will replace p_0 by $1 - p_0$, and test the hypotheses

$$\begin{aligned} H_0 &: D_2^{(m)} \geq (1 - p_0)^{m-1} \\ H_1 &: D_2^{(m)} < (1 - p_0)^{m-1} \end{aligned} \qquad (3)$$

showing whether the raters fail in the discrimination less than expected by chance. For example, in the triangle test with two panelists, when $D_1^{(2)}$ and $D_2^{(2)}$ are significantly greater than 1/3 and less than 2/3, respectively, one can assume that the panelists are not guessing and are actually revealing differences between the products being evaluated.

When both hypothesis tests, (2) and (3), are significant, the raters are indeed discriminating products. In this case, $D_1^{(m)}$ must be large and $D_2^{(m)}$ must be small. This means that the raters are mostly giving correct responses (i.e., a high degree of agreement is attained), and therefore they are noticing differences between the products. Otherwise, there is no evidence that the raters are properly discriminating products. This procedure is also applicable using bilateral hypothesis tests.

Dice indices have several advantages over other measures in studying the agreement for difference tests. Together, positive and negative Dice indices show the consistency of the raters in the two directions, indicating whether the products are different or, on the contrary, are similar. In addition, some other measures can report low overall agreement while the separate agreements for both the positive and negative responses are high. For example, the effect of symmetrically unbalanced marginal totals may lead to a low value of Cohen's kappa (see Cicchetti and Feinstein (1990b)). In this case, the wrong conclusion may be drawn that the products are similar when they are actually different, and the Dice indices could have detected any potential differences. Finally, when just one of the positive or negative agreements is low, most indices tend also to be low because they reward symmetry between agreement and disagreement. in contrast, a low negative dice index with a high positive dice index is indicative of major differentiation.

In order to perform the hypothesis tests of (2) and (3), one must know the statistical distributions under the null hypotheses. Sometimes it is not possible to derive a closed form expression for a given sampling distribution, and indeed this seems to be the case here. No sampling distribution has as yet been obtained in a closed form relating to the Dice index. However, sampling distributions of interest may be estimated by Monte Carlo simulation. Using this technique, it is possible to generate approximations to the true sampling distributions of the test statistics. The precision of the approximation depends strongly on the number of simulations performed. Monte Carlo estimation of sampling distributions is widely used in many practical settings. For example, the IBM SPSS software package offers it as an option when the data does not satisfy the necessary conditions for asymptotic methods to be used or the samples are so large that the computation time required is prohibitive. The procedure provides an unbiased estimate of the exact p-value (see, e.g., Mehta and Patel (2010)).

Once the hypothesis tests have been set, the sampling distribution is estimated according to the number of raters m, the number of replications k, and the number of simulations. Table 1 presents the critical values for several significance levels in different scenarios. These critical values were obtained by Monte Carlo estimating the sampling distributions with 1 000 000 simulations. The estimated sampling distributions for $D_1^{(m)}$ are asymmetrically distributed with right-side tails, whereas those for $D_2^{(m)}$ are approximately normal. Since one-sided hypothesis tests are considered, attention must be paid to the right (left) tail for the sampling distribution of $D_1^{(m)}$ ($D_2^{(m)}$). A reduced table is presented for illustrative purposes.

	$\hat{D}_1^{(m)}$					$\hat{D}_2^{(m)}$			
	$k=10$	$k=20$	$k=30$	$k=40$		$k=10$	$k=20$	$k=30$	$k=40$
$m=2$					$m=2$				
0.01	0.8000	0.6667	0.6316	0.5926	0.99	0.2000	0.3636	0.4324	0.4681
0.025	0.7500	0.6316	0.5833	0.5517	0.975	0.2857	0.4211	0.4706	0.5000
0.05	0.6667	0.5882	0.5455	0.5185	0.95	0.3636	0.4615	0.5128	0.5306
0.1	0.6000	0.5333	0.5000	0.4800	0.9	0.4444	0.5185	0.5455	0.5652
$m=3$					$m=3$				
0.01	0.5455	0.4091	0.3529	0.3243	0.99	0.0000	0.1579	0.1935	0.2308
0.025	0.4615	0.3600	0.3103	0.2857	0.975	0.0000	0.1765	0.2308	0.2647
0.05	0.4000	0.3158	0.2813	0.2553	0.95	0.1579	0.2368	0.2679	0.2917
0.1	0.3333	0.2727	0.2368	0.2195	0.9	0.1765	0.2647	0.3103	0.3288
$m=4$					$m=4$				
0.01	0.3636	0.2667	0.2105	0.1905	0.99	0.0000	0.0000	0.0588	0.1132
0.025	0.3077	0.2105	0.1860	0.1538	0.975	0.0000	0.0784	0.1067	0.1250
0.05	0.2667	0.1667	0.1538	0.1404	0.95	0.0000	0.0851	0.1463	0.1569
0.1	0.2222	0.1429	0.1081	0.0889	0.9	0.0000	0.1509	0.1622	0.1887

Table 1: Critical values for the Monte Carlo estimated sampling distributions.

The rejection regions for the two tests generally grow with increasing number of replications and/or number of panelists. This means that less positive agreement and more negative agreement are necessary to detect significant results, and consequently to reveal product differences. For example, in the triangle test case, rejecting $H_0 : D_1^{(m)} \leq (1/3)^{m-1}$ is difficult when there are only 10 replications and 2 panelists. With $\alpha = 0.01$, a positive agreement value greater than 0.8 is needed, and rejecting the null hypothesis becomes easier as the number of replications increases. For $D_2^{(m)}$, the lowest rejection values are attained with 4 raters and 10 replications. With more replications and fewer panelists it is easier to reject the null hypothesis $H_0 : D_2^{(m)} \geq (2/3)^{m-1}$.

In practice, it is not necessary to use a table of critical values and probabilities, since the approximated sampling distributions are already available, and can be used to obtain the necessary p-values. Monte Carlo simulations can also be used to calculate the one-sided confidence intervals.

It must be remarked that the proposed approach is not valid for overdispersed binomial data. In such cases, the beta-binomial model should be used instead because the parameter p can have great variability. This possibly extreme variability directly affects the estimates of the agreement indices by yielding large positive and negative Dice indices. With overdispersion, even when there is a clear difference between products, both $D_1^{(m)}$ and $D_2^{(m)}$ are large, leading to misinterpretations if the proposed approach is used. Therefore, before applying this proposed approach, it is advisable to perform an overdispersion test. In Subsection 3.4, the effect of overdispersion is illustrated with a simulation based example.

3.3. Pairwise comparisons

Besides the information on discrimination provided by the approach, the problem can also be decomposed into $m(m-1)/2$ two-rater problems, i.e., performing pairwise comparisons for all the raters. These comparisons provide information on the agreement between raters. This can be useful in evaluating the degree of agreement for each panelist relative to the others, and to determine the level of expertise of novice trainee panelists.

These Dice indices can be interpreted in a similar way to that of the general problem. Comparison of all the pairwise results together leads to two possibilities – either all the panelists agree in the same way by pairs or they do not. When all the pairwise comparison hypothesis tests are significant, all the panelists are discriminating products in a similar way.

On the one hand, if all the panelists agree in the same way, i.e., both $D_1^{(2)}$ and $D_2^{(2)}$ are similar for all the pairs then the panelists have the same level of expertise and roughly the same discriminatory reliability. In particular, they all have approximately the same influence on the general agreement and on the differentiation between products. On the other hand, if the panelists agree differently by pairs, one can identify which of them are the sources of the increase or decrease in the general agreement. In Subsection 3.4, we shall present an illustrative example of the interpretation of pairwise comparisons.

The proposed framework for pairwise comparisons is particularly helpful when novice panelists are being trained by an expert. The expert can be taken as the gold standard, with the pairwise comparisons representing an objective form of ranking the panelists by efficacy.

3.4. Illustrative examples

In this subsection, we shall present three examples illustrating some typical scenarios of difference test problems using the triangle test.

Example 1. Binomial data. We first considered two scenarios in the binomial model. In one, the proportion of successes used for the simulation was taken to be $p = 1/3$, corresponding to agreement by chance, so that the products should not be discriminated. In the other, we took $p = 2/3$, corresponding to the raters being able to properly discriminate between products. In each case, 20 000 contingency tables were simulated for 3 raters and 20 replications, fitting the beta-binomial model, and estimating its parameters and the Dice indices, $D_1^{(3)}$ and $D_2^{(3)}$. The p-values were for the hypothesis tests of the beta-binomial model (see Ennis and Bi (1998)) and of the Dice indices (see Subsection 3.2) were also calculated. The averages of the parameter estimates and the p-values over the 20 000 simulations are presented in Table 2.

Param.	$\hat{\mu}$	$\hat{p}$-value	$\hat{\gamma}$	$\hat{p}$-value	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
$\mu = 1/3$	0.3330	0.6238	0.0442	0.7957	0.1042	0.3875	0.4372	0.4997
$\mu = 2/3$	0.6667	0.0006	0.0435	0.7995	0.4363	0.0311	0.1035	0.0314

Table 2: Estimated parameters and p-values for binomial simulated data.

The estimates of the parameter μ agree with the pre-set values used to generate the data with the binomial model, i.e., $p = 1/3$ and $p = 2/3$. The hypothesis test is non-significant for the case generated with $\mu = 1/3$, and significant for the case generated with $\mu = 2/3$. The overdispersion tests are not significant, and the estimates of γ are close to zero in both cases. This validates the simulation process.

When the probability of success is $p = 1/3$, the positive and negative agreement indices are close to their expected values 1/9 and 4/9, respectively. According to the estimated $p - values$, these indices are not significant, and consequently the products can not be considered to have been discriminated.

Finally, when $p = 2/3$, both hypothesis tests are significant, indicating that the agreement is not by chance, and that the raters properly discriminate the products.

Table 3 presents the results for the pairwise comparisons. All the positive Dice indices for the pairs are similar and close to the expected values 1/3 and 2/3, indicating homogeneity among raters for the positive response. The same is the case for the negative response with respect to the expected values 2/3 and 1/3. Thus, the panelists agree (disagree) and differentiate (do not differentiate) in the same way.

$\mu = 1/3$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.3179	0.4966	0.6588	0.4970
Rater 1 vs Rater 3	0.3165	0.4976	0.6592	0.4976
Rater 2 vs Rater 3	0.3138	0.5036	0.6577	0.4961
$\mu = 2/3$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.6576	0.0450	0.3149	0.0432
Rater 1 vs Rater 3	0.6581	0.0441	0.3163	0.0437
Rater 2 vs Rater 3	0.6577	0.0441	0.3156	0.0434

Table 3: Estimated Dice indices and p-values for pairwise comparisons.

Example 2. Consequences of overdispersion.

As previously observed, the proposed approach is not valid for overdispersed binomial data. The following is a simulation-based example to illustrate the effects of overdispersion on the Dice indices. The beta-binomial model is used to generate the data with different levels of overdispersion: low, medium, and high ($\gamma = 0.2$, $\gamma = 0.5$, and $\gamma = 0.8$). Low and high success probabilities were also considered ($\mu = 1/3$ and $\mu = 2/3$). Again 20 000 contingency tables were simulated for 3 raters and 20 replications for the different scenarios. The results are summarized in Table 4.

The estimated values for the parameters of the beta-binomial distribution agree with the ones set beforehand, validating the simulation process. It can be seen that the Dice indices increase as the overdispersion increases. When γ is low, the positive and negative indices are closer to their expected values, whereas, when overdispersion increases, the agreement indices increase too.

Table 5 presents the results for the pairwise comparisons. Again, all the Dice index values increase as the overdispersion increases.

The extreme variability distorts any interpretation of the agreement and

Parameters	$\hat{\mu}$	$\hat{p}$-value	$\hat{\gamma}$	$\hat{p}$-value	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
$\mu = 1/3$, $\gamma = 0.2$	0.3339	0.7516	0.1847	0.3503	0.2426	0.1993	0.5599	0.7618
$\mu = 2/3$, $\gamma = 0.2$	0.6668	0.0008	0.1897	0.3388	0.5637	0.0060	0.2474	0.8239
$\mu = 1/3$, $\gamma = 0.5$	0.3336	0.7571	0.4804	0.0260	0.5015	0.0328	0.7339	0.9668
$\mu = 2/3$, $\gamma = 0.5$	0.6750	0.0001	0.4784	0.0322	0.7407	0.0002	0.4909	0.3769
$\mu = 1/3$, $\gamma = 0.8$	0.3329	0.7662	0.7889	0.0005	0.7905	0.0013	0.8961	0.9994
$\mu = 2/3$, $\gamma = 0.8$	0.6687	0.0000	0.7860	0.0002	0.9018	$1.1 \cdot 10^{-6}$	0.7869	0.9642

Table 4: Estimated parameters and p-values for beta-binomial simulated data.

$\mu = 1/3$, $\gamma = 0.2$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.4461	0.2928	0.7237	0.6694
Rater 1 vs Rater 3	0.4438	0.2954	0.7234	0.6701
Rater 2 vs Rater 3	0.4455	0.2916	0.7249	0.6803
$\mu = 2/3$, $\gamma = 0.2$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.7261	0.0178	0.4467	0.1478
Rater 1 vs Rater 3	0.7263	0.0180	0.4479	0.1500
Rater 2 vs Rater 3	0.7274	0.0174	0.4500	0.2001
$\mu = 1/3$, $\gamma = 0.5$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.6493	0.0782	0.8276	0.8975
Rater 1 vs Rater 3	0.6488	0.0782	0.8273	0.8968
Rater 2 vs Rater 3	0.6495	0.0777	0.8280	0.8972
$\mu = 2/3$, $\gamma = 0.5$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.8354	0.0020	0.6574	0.5331
Rater 1 vs Rater 3	0.8333	0.0028	0.6466	0.4962
Rater 2 vs Rater 3	0.8236	0.0025	0.6307	0.5013
$\mu = 1/3$, $\gamma = 0.8$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.8576	0.0058	0.9311	0.9923
Rater 1 vs Rater 3	0.8573	0.0060	0.9306	0.9923
Rater 2 vs Rater 3	0.8572	0.0062	0.9312	0.9998
$\mu = 2/3$, $\gamma = 0.8$	$\hat{D}_1^{(3)}$	$\hat{p}$-value	$\hat{D}_2^{(3)}$	$\hat{p}$-value
Rater 1 vs Rater 2	0.9298	0.0001	0.8556	0.8985
Rater 1 vs Rater 3	0.9293	0.0001	0.8522	0.8971
Rater 2 vs Rater 3	0.9314	0.0001	0.8570	0.9067

Table 5: Estimated Dice indices and p-values for pairwise comparisons.

the differentiation. Therefore, the proposed approach must only be applied to non-overdispersed binomial data.

Example 3. Detecting non-accurate raters.

The effect of one or more conflictive raters can be analysed in the pairwise comparison framework. Table 6 is the contingency table for a scenario in which one rater disagrees with the other two. Non-overdispersed binomial data are obtained ($\hat{\gamma} = 0.00001$, and $\hat{p} - value = 1$). Using the binomial procedure, the three raters are seen to discriminate between products ($\hat{p} = 0.6480$, $\hat{p} - value = 2.1 \cdot 10^{-9}$), but it can not be seen which panelists differentiate and which do not.

				Rater 3		
				A	F	Total
Rater 1	A	Rater 2	A	4	21	25
			F	2	2	4
			Total	6	23	29
				A	F	Total
	F		A	4	1	5
			F	1	1	2
			Total	5	2	7

Table 6: Contingency table.

Table 7 presents the general Dice indices and the pairwise comparisons. It can be observed that the hypothesis tests for these indices are not simultaneously significant, indicating that the three raters do not properly discriminate between products.

	$\hat{D}_1^{(m)}$	$\hat{p}$-value	$\hat{D}_2^{(m)}$	$\hat{p}$-value
Raters 1 vs 2 vs 3	0.1714	0.2294	0.0789	0.1118
Rater 1 vs Rater 2	0.8475	0.0003	0.3077	0.0037
Rater 1 vs Rater 3	0.3000	0.5538	0.1250	$8 \cdot 10^{-5}$
Rater 2 vs Rater 3	0.3902	0.3526	0.1936	0.0972

Table 7: Estimated Dice indices and p-values.

The hypothesis tests for raters 1 and 2 are significant, denoting that they are indeed able to discriminate. In contrast, the comparisons between rater 3 and the other two (1-3 and 2-3) indicate that the agreement is by chance because the hypothesis tests are not simultaneously significant. It is apparent that rater 3 is the only one with low agreement, but that that rater's failures decisively affect the general agreement. Raters 1 and 2 are able to differentiate the given products, but rater 3 is only guessing.

The following section illustrates the application of the proposed approach in a real context.

4. Application

An experimental study was performed in order to illustrate the potential of the approach in discriminating between two meat products and to evaluate the inter-rater agreement. The triangle test was used with the guidelines defined by the norm International Organization for Standarization (2004b). We shall first describe the experiment.

Two Iberian pork loins of the same quality (Iberian pigs fed partly on fodder and partly on mast) were evaluated. The first is a *Carrefour* house-brand pork loin, and the second is produced by a traditional company, *La Flor Piornalega*. This variety of pork loin is obtained from free-range Iberian pigs fed on cereals and mast (acorns) and sacrificed at 12 months. The two pieces considered were dry-cured at specialist sites in Spain under very similar conditions of humidity and altitude (Guijuelo for the *Carrefour* product, and Piornal for the *La Flor Piornalega* product). The two loins were tasted by three panelists, and the results analysed by the present proposed approach.

The procedure was as follows. A set of three samples was presented simultaneously to each rater, two of them belonging to the same loin. This step was repeated several times with different sets of samples. The raters had to identify which sample was different in each set presented. There were six sessions, and every rater tasted just six sets per session to avoid sensory fatigue. In total, therefore, each rater dealt with 36 sets. All three raters were novices because the objective was to identify whether any differences were noticeable from a regular consumer's point of view.

The samples in the sets were displayed uniformly, and all corresponded to the same two pieces of pork loin. The experiment was performed under the same conditions of temperature and lighting in a standardized tasting room. The patterns followed to display the samples were: CPP, PCC, CCP, PPC, CPC y PCP, with C being *Carrefour* and P *La Flor Piornalega*. To tabulate the results, the two possible responses were A if they found the different sample, and F if they failed. The results are presented in Table 8. Note that the first rater obtained 30 correct responses and only 6 were incorrect, the second rater obtained 28 correct responses and 8 incorrect, and the third rater obtained 27 correct responses and 9 incorrect.

				Rater 3		
				A	F	Total
Rater 1		Rater 2	A	20	4	24
	A		F	4	2	6
			Total	24	6	30
				A	F	Total
			A	2	2	4
	F		F	1	1	2
			Total	3	3	6

Table 8: Contingency table for the experimental results.

First, we shall approach the discrimination problem by following the standard methodology. Depending on the properties of the data, there are two possibilities. If there is no variation among trials then the binomial model considered in the norm International Organization for Standarization (2004b) can be applied. Otherwise, one should use the beta-binomial model (see Ennis and Bi (1998)). In order to choose the model, an overdispersion analysis is applied. The maximum likelihood estimates for the beta-binomial parameters are $\hat{\mu} = 0.8151$ and $\hat{\gamma} = 0.0921$, and the 95% two-sided confidence intervals are $(0.7355, 0.8948)$ and $(0.0000, 0.3222)$, respectively. The scale parameter estimate is close to zero, and the corresponding hypothesis test

$$\begin{aligned} H_0 &: \gamma = 0 \\ H_1 &: \gamma \neq 0, \end{aligned} \tag{4}$$

provides a non-significant result with $p - value = 0.3946$. There is no evidence that γ is different from zero, and hence neither of overdispersion being present. The binomial model can thus be applied.

For the binomial model, p must be estimated using the number of correct differentiations, x_c, and the number of experiments, $k \cdot m = 36 \cdot 3 = 108$. In this experiment the estimated probability of success is $\hat{p} = 0.8148$, with a 95% two-sided confidence interval equal to $(0.7424, 1)$. The hypothesis test

$$\begin{aligned} H_0 &: p = 1/3 \\ H_1 &: p \neq 1/3, \end{aligned} \tag{5}$$

yields $p - value = 2.2 \cdot 10^{-16}$. Thus, the panelists are not guessing, and they are discriminating between products.

The present proposal allows the foregoing information to be complemented with other aspects, such as how intense the agreement is with respect to the differentiation, or whether the panelists differentiate the samples in a similar way.

With respect to the agreement, it can be observed that the most frequent result is the agreement among the three raters (A, A, A), which occurs 20 times out of a total of 36. When there are more than two raters, it becomes more difficult to differentiate all the samples simultaneously. In this case, the three raters found the different sample for the same sets 56% of the time. It is also remarkable that there was only one jointly failed differentiation. The proportion of agreement $\sum_i n_{iii}/n = 21/36 = 0.58$ summarizes this information. This is quite a high proportion result for three raters, but it is interesting to distinguish whether these agreements come from right or from wrong differentiations between the two products.

The generalized Dice index of agreement is used to evaluate the overall conditional agreement for each response. The Dice indices are $\hat{D}_1^{(3)} = 0.71$ for the correct responses, and $\hat{D}_2^{(3)} = 0.13$ for the incorrect ones. The 95% one-sided confidence intervals are (0.3158, 0.8182) and (0,0.5400) respectively. These indices lead to the conclusion that the three raters differentiate the two samples quite well, because the positive agreement among them is high and the negative agreement is low. In order to formalize this result, the proposed one-sided hypothesis tests are applied. Monte Carlo simulations were used to generate the Dice index distributions for 3 raters and 36 replications. Figure 1 shows the distribution of the Dice indices under the null hypotheses $H_0 : D_1^{(3)} \leq 1/9$ and $H_0 : D_2^{(3)} \geq 4/9$, respectively.

The first hypothesis test provides a $p - value = 0$, i.e., none of the 1 000 000 values generated from the statistical distribution is greater than 0.71. The second hypothesis test gives a $p - value = 0.000426$. Both tests are significant, indicating that, simultaneously, the positive agreement is greater than expected by chance and the negative agreement is less than expected by chance. This means that differences between the two products are indeed found. This result reinforces that previously obtained with the binomial model. Moreover, this approach yields information about the degree of discrimination, which, in this case, is high.

A pairwise comparison provides information about whether the raters

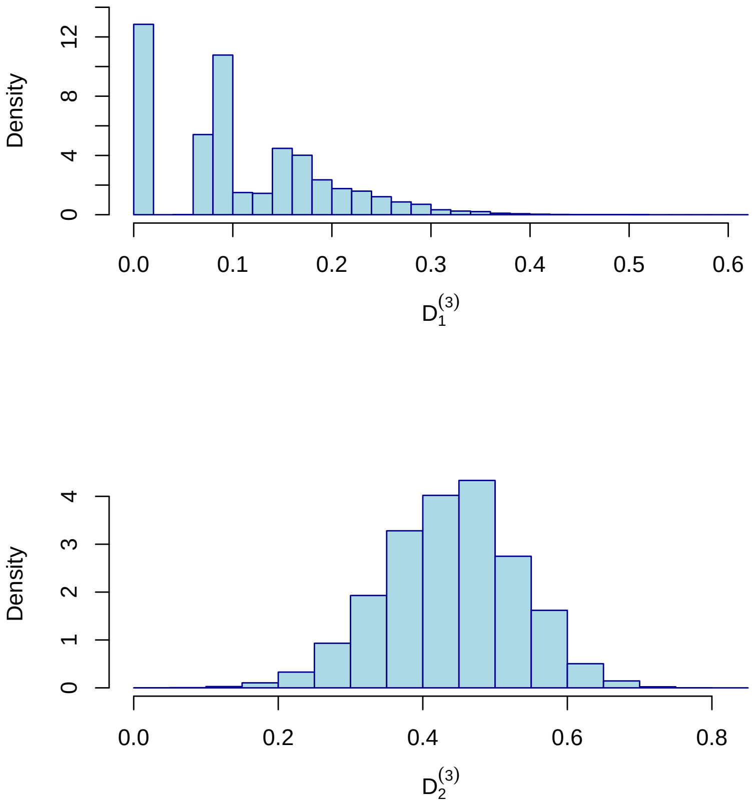

Figure 1: Distribution of Dice indices under the null hypothesis for 3 raters and 36 replications.

are discriminating the samples in a similar way or whether one or more of them are influencing the agreement more. Table 9 presents the pairwise Dice indices for the three raters. It also gives the p-values according to the tests defined in Section 3.

Note that the positive Dice indices are very similar (from 0.80 to 0.84) and are high. The indices for the negative response are also similar (from 0.29 to 0.40). This emphasizes that the raters all behave in a similar way, i.e., there is no rater performing the test in a better or worse way than the others. The three raters seem to have a similar capability to differentiate,

	$D_1^{(2)}$	p-value	$D_2^{(2)}$	p-value
Rater 1 vs Rater 2	0.83	0.0004	0.40	0.0016
Rater 1 vs Rater 3	0.84	0.0002	0.29	0.0149
Rater 2 vs Rater 3	0.80	0.0005	0.35	0.0049

Table 9: Dice indices and p-values for pairwise comparisons.

and they obtained a high degree of differentiation. When performing the hypothesis tests, all the p-values are very small (the largest is 0.0005), rejecting the possibility that the panelists are guessing when comparing them pairwise. According to these results, the panelists found significant differences between the products studied, both all together and individually. The degree of differentiation was very high.

5. Conclusions

A novel approach to sensory analysis discrimination tests has been described. It is based on the generalized positive and negative Dice agreement indices, which are used to develop two hypothesis tests. Monte Carlo simulation is used to obtain the distribution under the null hypothesis and the corresponding p-values. The approach provides information on the discrimination between products and its strength. Pairwise comparison is used to examine the influence of each rater on the discrimination process. This framework can also be used to train novice panelists by comparing them with experts.

The present proposal is not a substitute for the traditional method based on the binomial distribution, but complements it by providing additional information. The applicability of the approach was illustrated by way of some examples, and an experiment was performed using the triangle test scheme to differentiate between two meat products. The results reinforced those given by using the classical binomial model, and showed that the degree of differentiation was quite high. Moreover, all the raters were good at discriminating the products, and none was better or worse than the others in this task.

The proposed approach is especially interesting when the standardized binomial method is not recommended, i.e., when few raters are involved and replications are needed. It is also recommendable when the interest is on rating novice panelists being trained by an expert, because it allows them to be ranked by skill.

Acknowledgments

We thank Julia Calvarro of the Food Hygiene Area (University of Extremadura) for her assistance in the tasting experiment and Jacinto Martín for his insightful comments. We also thank two anonymous referees for comments and suggestions which have improved the content and readability of the paper. This research was partially funded by the *Ministerio de Economía y Competitividad, Spain* (Project MTM2011-28983-C03-02), *Gobierno de Extremadura, Spain* (Project GRU10110), and *European Union* (European Regional Development Funds).

References

AGRESTI, A., 1996. An introduction to categorical data analysis. Wiley.

AJMONE-MARSAN, P., NEGRINI, R., CREPALDI, P., MILANESI, E., GORNI, C., VALENTINI, A., CICOGNA, M., 2001. Assessing genetic diversity in Italian goat populations using AFLP markers. Animal Genetics *32 (5)*, 281–288.

ANDERSON, D. A., 1988. Some models for overdispersed binomial data. Australian Journal of Statistics *30 (2)*, 125–148.

BAIXAS-NOGUERAS, S., BOVER-CID, S., VECIANA-NOGUS, T., NUNES, M., VIDAL-CAROU, M., 2003. Development of a quality index method to evaluate freshness in Mediterranean hake (Merluccius merluccius). Journal of Food Science *68 (3)*, 1067–1071.

BI, J., 2011. Similarity tests using forced-choice methods in terms of Thurstonian discriminal distance, d'. Journal of Sensory Studies *26 (2)*, 151–157.

CICCHETTI, D. V., FEINSTEIN, A. R., 1990a. High agreement but low kappa: I. the problem of two paradoxes. Journal of Clinical Epidemiology *43 (6)*, 551–558.

CICCHETTI, D. V., FEINSTEIN, A. R., 1990b. High agreement but low kappa: II. resolving the paradoxes. Journal of Clinical Epidemiology *43 (6)*, 551–558.

COHEN, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement *20 (1)*, 37–46.

DUBNICKA, S. R., 2013. A bayesian approach to analyzing replicated preference tests. Journal of Sensory Studies.*28 (3)*, 171–187.

ENNIS, D. M., BI, J., 1998. The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. Journal of Sensory Studies *13 (4)*, 389–412.

ENNIS, J., JESIONKA, V., 2011. The power of sensory discrimination methods revisited. Journal of Sensory Studies *26 (5)*, 371–382.

FLEISS, J. L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin *76 (5)*, 378–382.

FLEISS, J. L., COHEN, J., EVERITT, B. S., 1969. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin *72 (5)*, 323–327.

FLETCHER, I., MAZZI, M., NUEBLING, M., 2011. When coders are reliable: The application of three measures to assess inter-rater reliability/agreement with doctor-patient communication data coded with the VR-codes. Patient Education and Counseling *82 (3)*, 341 – 345.

GARCIA, K., ENNIS, J. M., PRINYAWIWATKUL, W., 2012. A large-scale experimental comparison of the Tetrad and triangle tests in children. Journal of Sensory Studies *27 (4)*, 217–222.

GRAHAM, P., BULL, B., 1998. Approximate standard errors and confidence intervals for indices of positive and negative agreement. Journal of Clinical Epidemiology *51 (9)*, 763 – 771.

GUGGENMOOS-HOLZMANN, I., 2006. How reliable are chance-corrected measures of agreement? Statistics in Medicine *12 (23)*, 2191–2205.

GWET, K., 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical Methods For Inter-Rater Reliability Assessment *1*, 1–6.

INTERNATIONAL ORGANIZATION FOR STANDARIZATION. TC34/SC12 - Sensory Analysis.

INTERNATIONAL ORGANIZATION FOR STANDARIZATION, 2004a. ISO 10399:2004. Sensory analysis. Methodology. Duo-Trio test.

INTERNATIONAL ORGANIZATION FOR STANDARIZATION, 2004b. ISO 4120:2004. Sensory analysis. Methodology. Triangular test.

INTERNATIONAL ORGANIZATION FOR STANDARIZATION, 2005. ISO 5495:2005. Sensory analysis. Methodology. Paired comparison test.

JENSCHKE, B. E., HODGEN, J. M., MEISINGER, J. L., HAMLING, A. E., MOSS, D. A., AHNSTRM, M. L., ESKRIDGE, K. M., CALKINS, C. R., 2007. Unsaturated fatty acids and sodium affect the liver-like off-flavor in cooked beef. Journal of Animal Science *85 (11)*, 3072–3078.

KUNERT, J., MEYNERS, M., 1999. On the triangle test with replications. Food Quality and Preference *10 (6)*, 477 – 482.

LAPARA, T. M., NAKATSU, C. H., PANTEA, L. M., ALLEMAN, J. E., 2002. Stability of the bacterial communities supported by a seven-stage biological process treating pharmaceutical wastewater as revealed by PCR-DGGE. Water Research *36 (3)*, 638 – 646.

LIGGET, R. E., DELWICHE, J. F., 2005. The beta-binomial model: variability in overdispersion across methods and over time. Journal of Sensory Studies *20 (1)*, 48–61.

LORENZO, J. M., GARCÍA FONTÁN, M. C., FRANCO, I., CARBALLO, J., 2008. Biochemical characteristics of dry-cured lacón (a Spanish traditional meat product) throughout the manufacture, and sensorial properties of the final product. Effect of some additives. Food Control *19 (12)*, 1148 – 1158.

MACKINNON, A., 2000. A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. Computers in Biology and Medicine *30 (3)*, 127–134.

MANLY, B. F. J., 1997. Randomization, Bootstrap and Monte Carlo Method in Biology. Chapman and Hall.

MARTÍN, A., BENITO, M., ARANDA, E., RUIZ-MOYANO, S., CÓRDOBA, J., CÓRDOBA, M., 2010. Characterization by volatile compounds of microbial deep spoilage in Iberian dry-cured ham. Journal of Food Science *75 (6)*, 360–365.

MEHTA, C. R., PATEL, N. R., 2010. IBM SPSS Exact Tests. IBM.

MEYNERS, M., BROCKHOFF, P., 2003. The design of replicated difference tests. Journal of Sensory Studies *18 (4)*, 291–324.

MOUNCHILI, A., WICHTEL, J., BOSSET, J., DOHOO, I., IMHOF, M., ALTIERI, D., MALLIA, S., STRYHN, H., 2005. HS-SPME gas chromatographic characterization of volatile compounds in milk tainted with off-flavour. International Dairy Journal *15 (12)*, 1203 – 1215.

PONS-SANCHEZ-CASCADO, S., VIDAL-CAROU, M. C., NUNES, M. L., VECIANA-NOGUES, M. T., 2006. Sensory analysis to assess the freshness of Mediterranean anchovies (Engraulis encrasicholus) stored in ice. Food Control *17 (7)*, 564 – 569.

RADOVICH, T. J. K., KLEINHENZ, M. D., DELWICHE, J. F., LIGGETT, R. E., 2004. Triangle tests indicate that irrigation timing affects fresh cabbage sensory quality. Food Quality and Preference *15 (5)*, 471 – 476.

RAMOS-GUAJARDO, A., GONZÁLEZ-RODRÍGUEZ, G., 2011. Hypothesis testing with fuzzy data: An application to quality control of cheese. In: 2011 11th International Conference on Intelligent Systems Design and Applications (ISDA). pp. 1335 –1340.

SEVERIANO, A., CARRICO, J. A., ROBINSON, D. A., RAMIREZ, M., PINTO, F. R., 2011. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. PloS one *6 (5)*, e19539.

SHOUKRI, M., 2004. Measures of interobserver agreement. Chapman and Hall.

SILVA, F., DUARTE, M., CAVALCANTI-MATA, M. E., 2010. Nova metodologia para interpretação de dados de análise sensorial de alimentos. Engenharia Agrícola *30 (5)*, 967–973.

VAN HOUT, D., HAUTUS, M. J., LEE, H.-S., 2011. Investigation of test performance over repeated sessions using signal detection theory: Comparison of three nonattribute-specified difference tests 2-AFCR, A-NOT A and 2-AFC. Journal of Sensory Studies *26 (5)*, 311–321.

VON EYE, A., MUN, E., 2005. Analyzing rater agreement: Manifest variable methods. Lawrence Erlbaum Associates.

WARRENS, M. J., 2008. Similarity coefficients for binary data : properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients. Ph.D. thesis, Leiden University.

WU, H., CHEN, L., 1995. Sensory analysis in quality control–the agreement among raters. Botanical bulletin of Academia Sinica *36*, 121–133.