László Sipos[1], Márta Ladányi[2], Zoltán Kókai[1], Attila Gere[1]

Received: 2016. August – Accepted: 2016. November

Revision of the performance evaluation methods of sensory panelists performing descriptive analysis

Keywords: performance evaluation of panelists, correlation, regression, regression diagnostics, standard

1. Summary

Sensory tests form the basis for sensory science. Sensory science uses human senses as measurement tools. During sensory tests, the properties of a product are evaluated by sensory panelists and by a sensory panel consisting of them. Decisions made after sensory tests are fundamentally determined by the quality of the data experienced, therefore, the quality of sensory data is determined by the trained and expert sensory sensory panel and its members. In our work, revision of the correlation and regression methods recommended by the standard titled „MSZ ISO 11132:2013 Sensory analysis. Methodology. Guidelines for monitoring the performance of a quantitative sensory panel" are described, and corrections are suggested.

2. Introduction and literature review

According to Kermit és Lengard **[1]**, a good sensory sensory panel should provide precise, discriminative and accurate results. An ideal group performance can be achieved if the products are differentiated by each member of the group ("great product variety"), and the same results are obtained several times ("small variation for the individual panelists"). However, there should be agreement with the other panel members regarding the sensory property, within a given tolerance (small variation among panelists) **[2],[3],[4]**. The task of the sensory panel leader is to collect the necessary information about members of the panel during sensory tests. By monitoring and following up performance it can be ensured that panel members and panel are capable of distinction, their results are constant, repeatable and free of error **[5]**.

Based on their training, sensory panelists are classified into three categories by the literature: naive assessors (consumers), trained panelists, expert panelists. Panelists of different training levels are required for different tasks **[6], [7], [8]**.

It is characteristic of naive assessors that they do not analyze sensations, they experience them, during judging they rely on their own experience and project their own preferences onto the products being panelistd. Therefore, when untrained assessors are asked, it should be focused on liking or preference: "*Which product do you like most? How much do you like certain products? What is the ideal intensity of a property in a product? Which one would you choose, which one would you buy?"* During analyses generally called consumer tests, a large number of queries are performed (at least 60 persons), representing the underlying population (regarding gender, age, place of residence, educational level, net earnings, etc.,), based on a sampling plan. Panelists typically do not have prior knowledge of the products, they only evaluate a few products, with the help of simplified scales and easy to understand short questionnaires. In these cases, personal, subjective tastes are tested. Preference tests can be applied widely, for example, to compare competitor products, for product optimization, to monitor changes in formulation, or to perform brand studies or packaging tests. Preference tests are intended to determine whether there is an perceivable difference between the products tested,

[1] Szent István Egyetem, Élelmiszertudományi Kar, Árukezelési és Érzékszervi Minősítési Tanszék, H-1118 Budapest, Villányi út 35-43.
[2] Szent István Egyetem, Kertészettudományi Kar, Biometria és Agrárinformatika Tanszék, H-1118, Budapest, Villányi út 29-43.

and if there is such a difference, which products differ from each other and to what extent **[9]**.

Trained panelists receive specialized knowledge regarding the planning and execution of experiments used in the area of sensory science and good practices of experimental conditions and testing. In addition to learning about the different sensory methodologies (difference testing, ranking tests, general tests), they undergo multi-stage panelist selection tests, where the measurement limits and the accuracy of their senses are tested. These are helped by domestic (MSZ), international (ISO) and adopted (MSZ ISO) standards: color recognition, color intensity test, flavor intensity test, flavor recognition test, odor recognition test, odor intensity test.

Trained panelists perform objective qualification in the sensory panel, their tasks include carrying out routine tests in the sensory sensory panel (sensory panel), reception of raw materials, finished product inspection, conformity assessment. Accordingly, the method of questioning is of an analytical nature: what are the intensities of the samples from the point of view of a specific, objectively definable characteristic, is there a difference between the samples, what is the nature of the difference, what characteristics are associated with the sample. During sensory tests, the emphasis is on measuring the intensity of the perceived characteristic, results are typically obtained by the statistical analysis of the values given by the members of the sensory panel **[7], [8]**.

Expert panelists are panelists of special sensitivity, experience and talent selected from among trained panelists (called "noses" in certain fields). Compared to trained panelists, they receive special, product specific training, lasting several months or even years, where they learn about the recognition, intensity values and errors of the sensory properties of products. For these trainings, it is advantageous to use specialized tests, different aromatic substances, fragrance trainings – *Le nez du vin, Le nez du Café* – reference samples, or an flavour wheel. They also have great experience in the software part of the methods. Both trained and expert panelists make decisions involving great responsibility.

The testing and development of trained and expert panelists, as well as the measurement of their performance are typically realized through a standardized, multi-stage system based on feedback, implemented under standardized conditions, preferably with software support.

3. Objectives

Among other things, the standard titled „MSZ ISO 11132:2013 Sensory analysis. Methodology. Guidelines for monitoring the performance of a quantitative sensory panel" describes the methods for measuring the performance of individual panelists. With regard to the performance evaluation of trained and expert panelists, the objective of this research is refining, supplementing and revising the correlation and regression methods recommended by the above standard.

4. Materials and methods

The Annex (A4.2) to the standard titled MSZ ISO 11132:2013 Sensory analysis. Methodology. Guidelines for monitoring the performance of a quantitative sensory panel describes in detail the evaluation of individual panelists using correlation and regression analysis. The three key indicators of the performance of an individual panelist are the correlation coefficient, the intercept and the slope. The correlation coefficient shows how similarly the panelists use the rating scale when measuring a given property. An intercept significantly different from zero indicates panel inhomogeneity, i.e., that one or more panelists do not agree with the other members of the panel. A low slope shows that the range of scale used by the given panelist is not as wide as it is in the case of other panelists. The perfect line of the ideal sensory panel is one where the panel average and the average of the panelist overlap, the slope and the correlation coefficient are both 1.0, and the line intersects the coordinate axes at zero **[5]**.

The initial matrix contains the values given by 4 panelists to 6 products ***(Table 1)***. The annex to the standard contains only the results, therefore, detailed calculations belonging to the results are shown in the example of the first panelist.

For its examination between the individual and the panel, the average of the entire panel is used as a reference by the ISO standard. Determination of the correlation, slope and intercept of the regression line fitted to the individual panelists' points is also recommended. The formula of the Pearson correlation coefficient (r), where $\bar{x}$ is the average of the x_i values and $\bar{y}$ is the average of the y_i values:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

The correlation coefficient is a measure of the closeness of the linear relationship between the panel member and the panel average, and its value is independent of measurement units. It indicates the size and direction of the linear relationship between the two values. The further away it is from zero, the stronger the relationship is, and its value is between -1 (perfect negative correlation) and +1 (perfect positive correlation). In panel studies it is typically a significantly positive value, close to +1 (its value is zero if there is no relationship between the two rankings, they are random relative to each other, uncorrelated). Calculability conditions for the linear Pearson correlation coefficient are summarized according to Fidy and Makara [10]. Variables *Y* and *X* have to be

quantitative (continuous) variables, and their distributions must follow a normal distribution. All covariances must be linear. For each *X* value, there should be a corresponding *Y* value. *X* and *Y* values have to be measured independently of each other. Sample selection has to be by random sampling.

The essence of regression analysis is that a function-like relationship is sought between one or more "independent" (explanatory) and a dependent variable. The linear model fitted to the data can be described as follows: $Y=b_0 + b_1X + e$, where value b_0 is the theoretical regression constant (intercept), value b_1 is the theoretical regression coefficient, and value e is the random error (residual), about which it is assumed that it can be given, in the case of different values of *X*, by independent, normal distribution probability variables with an expected value of 0, that have the same standard deviation. The estimation is based on the least squares principle, according to which the so-called residual/error sum of squares is minimized:

$$F: (b_0, b_1) \mapsto \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$ **[11].**

Unfortunately, the standard does not mention the diagnostics of the regression model, therefore, in the section discussing the results, this is also presented step by step. After calculating the linear correlation, the goodness of the model selection can be determined based on the F value obtained during the variance analysis performed for the regression model. For the parameter estimation to be accepted as correct or good, the t values for the panel average and the intercept have to be evaluated. The values given by panelist no. 1 are estimated with the regression function obtained, using the entire panel average. The residuals are given by the differences between the estimated and the actual values. Residuals are plotted against the panel average (independent variable). If the figure is an irregular cloud of points, the independence of the residuals and the panel average is accepted. An important step of the diagnostics is the testing of the normality of the residuals. If the absolute values of skewness and kurtosis are less than 1, then the normality is acceptable. If the absolute values of the ratios (kurtosis)/(kurtosis error) and (skewness)/(skewness error) are less than 2, then again the normality is acceptable **[12], [13]**.

5. Results

Details of the evaluation of the individual using correlation and regression is described by standard ISO 11132:2012 on panel performance through a sample (A 4.2). The annex to the standard contains only the results, therefore, detailed calculations belonging to the results are shown in the example of the first panelist. Unfortunately, the standard does not mention the diagnostics of the regression model either, therefore, it is also described in detail in the following. In a linear case, the coefficient of determination of the model is the square of the empirical correlation coefficient (r), R^2=0.98, which can be given as the ratio of the model and the total variance, and its interpretation is that the standard deviation of the average values of the first panelist can be explained by the model in 98% ***(Table 2)***. In the case of the coefficient of determination, there is a specific connection, and it shows, how one variable can be predicted based on the other variable. On the other hand, the correlation coefficient is symmetrical, there is a two-way connection, and even in the case of a significant correlation it does not indicate a cause-and-effect relationship.

During variance analysis of variance of the regression model (ANOVA), $F=248.1307$ was obtained, its value is high, therefore, selecting a linear model proved to be good ***(Table 3)***. The F-value of the ANOVA was found to be significant $p=9.49 \cdot 10^{-5}$.

The t-value for the panel average is significant, and quite large ($t_{panel\ average}=15.75$; $p=9.49 \cdot 10^{-5}$) for the parameter estimation to be accepted as correct or good. The t-value for the intercept ($t_{interception}=1.03$) is not significant ($p=0.36$), therefore, the model intersects the axis at 0, i.e., the performance of the panelist, cosindering the panel consensus, is adequate ***(Table 4)***.

The value given by panelist no. 1 is estimated by the regression function obtained, using the entire panel average. Residuals are given by the differences between the estimated and the actual values. Residuals are plotted against the panel average (independent variable). Since the figure is an irregular cloud of points, the independence of the residuals and the panel average is accepted ***(Table 5, Table 6, Figure 1***).

An important step of diagnostics is the testing of the normality of the residuals. If the absolute values of skewness and kurtosis are less than 1, then the normality is acceptable. If the absolute values of the ratios (kurtosis)/(kurtosis error) and (skewness)/(skewness error) are less than 2, then again the normality is acceptable **[12], [13]** (***Table 7***).

If the kurtosis and/or the skewness is higher than 1, then the D'Agostino test is applied: $K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2)$, where, $Z^2(\sqrt{b_1})$ is the normal approximation of the skewness, and $Z^2(b_2)$ is the normal approximation of the kurtosis. If the sum of square is smaller than the critical value, and the significance level is above 0.05, then the normality is acceptable [14]. In our case, because of the kurtosis, the D'Agostino test is applied, which can be carried out in two ways. If the calculated p-value (0.690825) > 0.05, then the series of residuals follows a normal distribution. If the tabular critical value (5.991465) more than the calculated value (0.739738), then the series of residuals follows a normal distribution ***(Table 8)***.

The three key indicators of the performance of an individual panelist are the correlation coefficient, the

value of which is $r=0.992036$ (R^2=0.9841), the intercept b_0= –0.4254, and the slope b_1= 1.180552 (***Figure 2***). (It should be noted that if the intercept is not significant, then it is worth running the linear model in a way where there is no intercept, because in our case this means a consensus between the panelist and the sensory panel.)

Since the panel average also includes the dataset of the panelist with whom the comparison (correlation) is performed, the use of the panel average recommended by ISO distorts the results. Admittedly, in all cases, the correlation coefficient will have an aditional part that comes solely from the fact that the value of the individual is compared to values (the entire panel average) that also includes his/her data.

To eliminate this, it is advisable to introduce a correction – reflecting reality better – where the comparison (correlation) is based, instead of the entire average, on the average without the value of the given panelist. This is especially important because one of the conditions for the calculability of the Pearson coefficient, i.e, the independence of the variables, can only be ensured this way. Of course, the consequence of this method is that the values of all three performance characteristics – correlation coefficient, intercept, slope – will change.

As a result of the correction, of course, the correlation coefficient decreases in each case, because the dataset of the given panelist is taken out of the panel average. Due to the correction, the values of the intercept and the slope can either improve or deteriorate, depending on the panelist. The only exception is the very unlikely, special case, where the individual products are classified by all of the panelists completely identically, i.e., the panel average and the average of the panelist overlap. The regression line starts from point zero with a slope of 1, and the points are located on the line. In this case, these characteristics do not change.

The performance characteristics of the panelists were calculated both according to ISO, and according to the procedure modified by the correction. Based on the results it can be stated that all three performance characteristics of all of the panelists change, and the correlation coefficient was reduced in all cases. For the sake of clarity, symbols and color codes have been introduced. Directions of the changes in the values are indicated by arrows (increase ↑, decrease ↓). The evolution of the value of the panelist is shown by the different colors (green – improvement, red - deterioration). In the ideal case, the value of the correlation coefficient is 1, the intercept is 0, and the slope is 1 (**Table 9**).

For Panelist 1 described above, due to the correction, the value of the correlation coefficient decreased, while both the intercept and the slope approached those of the panel. A similar trend was observed in the case of Panelist 2 as well, who differed from Panelist 1 only marginally, based on the correction method. According to ISO, the best panelist was Panelist 4. It is worth highlighting Panelist 3, for whom all three parameters deteriorated. Of the sensory panel members, his/her correlation coefficient decreased to the greatest extent, so the results of the panel were distorted mainly by the values of this panelist. This is supported by the deteriorating intercept and slope values also. The slope values, lowest for this panelist, also indicate that the range of scale used by this panelist was not as wide as it was in the case of other panelists. It is important to emphasize that the value of the correlation coefficient (r) can only be interpreted at a certain significance level. If the correlation is not significant, a linear relationship is not proven, and so to reveal the connections, further studies are needed. The Pearson correlation matrix calculated with the correction (α=0.05) shows that the correlation coefficient of Panelist 3 was not proven to be significant (***Table 10***).

It should be emphasized that, at a preset significance level, the critical value is determined by the number of elements. Further calculations were carried out which showed that if 8 products had been tested by Panelist 3 instead of 6, then the calculated value would have been greater than the critical value, and so at a given significance level it is acceptable that there is a linear relationship between the corrected panel average and the values of the given panel member (r=0.7349; p=0.038; α=0.05). Of course, the averages were substituted for products 7 and 8, and so the value of r did not change. When shown as a scatter plot, in the case of 6 samples there are 6 points, while in the case of 8 samples there are 8 points, but it looks like there are only 7, because the two average points overlap. The value of the correlation coefficient is strongly influenced by outlier values. This is supported by the fact that, in the case of 8 samples, the value of the lowest point can only be balanced by two average values in order for the correlation coefficient to be significant (α=0.05) (**Figure 3**). It is advisable to perform further studies to determine the reasons for outliers.

With the help of the critical values of the Pearson correlation table, the effect of setting the significance level on the results can be demonstrated as well. Based on the results of Panelist 3 (r=0.7349), it would be significant in the case of 6 products if α=0.1, in the case of 8 products if α=0.05, and in the case of 12 products if α=0.01. For the correlation coefficient to be significant is also important, because this way the linear regression will be significant as well ***(Table 11)*** **[15]**.

In summary, it can be stated that the use of the panel average recommended by ISO can distort the results. Instead of this, it is advisable to take the value of the panelist to be evaluated out, for a better approximation of reality using the correction method. Both the corrected panel average and the value of the correlation coefficient calculated from it are

significantly influenced by the outlying value(s). Critical values of the Pearson correlation are influenced by the number of elements and the significance level. When performing the regression, professionally informed decisions can only be made after carrying out the diagnostics for the regression model. As a result of the correction, the correlation coefficient almost always decreases, but the change in the intercept and the slope depends on the panelist's evaluation.

A trained sensory panel usually consists of 10-12 members, which is a low number of elements (n<13) in a statistical sense, and the normality requirement regarding the variables may be breached. Consequently, the conditions for the parametric tests are not met, and so less efficient but distribution-independent, non-parametric methods should be selected. It is important to emphasize that, in addition to being distribution-independent, certain conditions are required by these tests. When analyzing the points given by the individual and the panel using the Pearson correlation, distorted results may be obtained.

Instead, with the non-parametric, robust counterpart of the correlation, the Spearman rho value should be calculated, which is not susceptible to either damages to the normality condition, or to differences in sample distribution **[16]**. The monotonicity of two variables, i.e., the closeness of changing together is measured by the Spearman rank correlation procedure. During the Spearman rank correlation, after putting sample data in order, a ranking number conversion is performed, i.e., ranks are assigned to the elements of the ordered samples (instead of $X_i \rightarrow Rank(X_i)$=ranking number). Then, the calculation of the Pearson correlation is performed for the ranking numbers **[17], [18]**.

Due to the required independence of the variables, carrying out the correction is recommended in this case as well. Here, the results of the Spearman correlation matrix calculated with the correction (α=0.05) were similar to those of the correlation calculated with the Pearson correction. The correlation coefficient of Panelist 3 was not significant here either and, again, 8 instead of 6 products had to be tested for the calculated to be greater than the critical value, in order for the result to be significant (r=0.6957; p=0.1361; α=0.05) *(**Table 12**)*.

However, based on the critical values of the Spearman correlation table, taking into account the results of Panelist 3 (r=0.6957), the results change, because they would be significant in the case of 8 products for α=0.1, in the case of 10 products for α=0.05, and in the case of 14 products for α=0.01. Calculating with the Spearman correlation, more products are required to obtain a significant result *(**Table 13**)* **[19]**. And, instead of linear regression, it is advisable to use robust regression, with the testing of the slope **[17]**.

6. Conclusions

For its analysis between the individual and the sensory panel, the average of the entire panel is used as a reference by standard MSZ ISO 11132:2013, and it recommends the determination of the correlation, slope and intercept of the regression line fitted to the individual panelist's points. Since the panel average also contains the dataset of the panelist for whom the comparison (correlation) is performed, the use of the panel average recommended by ISO distorts the results. It can be seen that the correlation coefficient will always have an additional part which is solely due to the fact that the values of the individual are compared to values (the entire panel average) which also include this panelist. To correct this, a solution is proposed in the results section.

Our calculations proved that, since the panel average also includes the dataset of the panelist for whom the comparison (correlation) is performed, the use of the panel average recommended by the standard distorts the results, therefore, for a correct calculation, the value of the panelist to be examined has to be excluded when calculating the average. Due to the proposed correction method, the correlation coefficient almost always decreases, while the values of the intercept and the slope change as a function of the panelist's evaluation.

Our calculations also proved that the corrected panel average – and so the correlation coefficient that comes from it – is significantly influenced by outlying value(s). It was highlighted that it is not well-founded to interpret the level of significance as the closeness of the relationship, because it is only related to the reliability of the decision rejecting the null hypothesis **[18]**. If the correlation is not significant, then the linear relationship is not proven, and so further studies are needed to explore the connections. Our calculations confirmed that a seemingly high correlation coefficient does not necessarily mean a significant difference in the case of a small number of elements, however, in the case of a large number of elements, a seemingly low corrrelation can still be significant, based on the table of bilateral critical values **[15]**.

In our work we proved that if the conditions for parametric tests are not met then, instead of the Pearson correlation, it is advisable to calculate the Spearman rho value using its non-parametric, distribution-independent robust counterpart. It was pointed out that, due to the required independence of the variables, carrying out the correction is also recommended in this case. Of course, the critical value should be determined based on the table of bilateral critical values for the Spearman correlation coefficient **[19]**.

In summary, based on the above, revision and modification of the standard "MSZ ISO 11132:2013 Sensory analysis. Methodology. Guidelines for monitoring the performance of a quantitative sensory panel" is recommended.

7. Acknowledgements

Supported BY the ÚNKP-16-4 New National Excellence Program of the Ministry of Human Capacities.

8. References

[1] Kermit, M., Lengard, V. (2005): Assessing the performance of a sensory panel–panellist monitoring and tracking. *Journal of Chemometrics,* 19, 154–161.

[2] Bi, J., Kuesten, C. (2012): Intraclass Correlation Coefficient (ICC): A Framework for Monitoring and Assessing Performance of Trained Sensory Panels and Panelists. *Journal of Sensory Studies,* 27, 5, 352–364.

[3] Derndorfer, E., Baierl, A., Nimmervoll, E., Sinkovits, E. (2005). A panel performance procedure implemented in R. *Journal of Sensory Studies,* 20, 217–227.

[4] Carbonell, L., Izquierdo, L., Carbonell, I. (2007). Sensory analysis of Spanish mandarin juices. Selection of attributes and panel performance. *Food Quality and Preference,* 18, 329–341.

[5] MSZ ISO 11132:2013 Érzékszervi vizsgálatok. Módszertan. Általános irányelvek a leíró vizsgálatot végző bírálóbizottság teljesítményének mérésére

[6] Molnár, P. (1991): Élelmiszerek érzékszervi vizsgálata. Budapest: Akadémiai Kiadó, 11–204.

[7] Kókai, Z. (2003): Az almafajták érzékszervi bírálata. Doktori értekezés. Budapest: Budapesti Közgazdaságtudományi és Államigazgatási Egyetem, 35-59.

[8] MSZ ISO 6658:2007 Érzékszervi vizsgálat. Módszertan. Általános útmutató.

[9] ISO 11136:2014 Sensory analysis -- Methodology -- General guidance for conducting hedonic tests with consumers in a controlled area

[10] Fidy, J. Makara, G. (2005): Biostatisztika. InforMed 2002 Kft.

[11] Harnos Zs., Ladányi, M. (2005): Biometria agrártudományi alkalmazásokkal. Budapest, Aula. 274-284, 307.

[12] Tabachnick, B. G., Fidell, L. S. (2003). Using Multivariate Statistics , 6th ed. Boston: Allyn and Bacon.

[13] Tabachnick, G. G., Fidell, L. S. (2007): Experimental Designs Using ANOVA. Belmont, CA: Duxbury.

[14] D'Agostino, R. B., Belanger, A., D'Agostino, R.B. Jr. (1990): "A suggestion for using powerful and informative tests of normality". *The American Statistician,* 44 (4) 316–321.

[15] Bevington, P.R. (1969): Data Reduction and Error Analysis for Physical Sciences. McGraw-Hill Book, New York.

[16] Sajtos, L., Mitev, A. (2007): SPSS kutatási és adatelemzési kézikönyv. Budapest: Alinea Kiadó. 163-244.

[17] Bard, Y. (1974): Nonlinear parameter estimation. New York, Academic Press.

[18] Vargha, A. (2008): Matematikai statisztika pszichológiai, biológiai és nyelvészeti alkalmazásokkal. Budapest, Pólya. 265-330.

[19] Weathington, B. L., Cunningham, C. J. L., Pittenger, D. J. (2012): Understanding Business Research. New Jersey, John Wiley & Sons. 454.