Food Quality and Preference 31 (2014) 22–27

Contents lists available at SciVerse ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Number of judges necessary for descriptive sensory tests

Rita de Cássia dos Santos Navarro da Silva [a,*], Valéria Paula Rodrigues Minim [a], Alexandre Navarro da Silva [b], Luis Antônio Minim [a]

[a] Departamento de Tecnologia de Alimentos, Universidade Federal de Viçosa (UFV), Zip Code: 36570-000 Viçosa, Minas Gerais, Brazil
[b] Departamento de Engenharia de Produção e Mecânica, Universidade Federal de Viçosa (UFV), Zip Code: 36570-000 Viçosa, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Received 27 May 2013
Received in revised form 18 July 2013
Accepted 20 July 2013
Available online 25 July 2013

Keywords:
Sample size
Power analysis
Panel size
RMSE

ABSTRACT

To determine the number of judges needed in descriptive tests four parameters are necessary: the probability of type I error (α), the probability of type II error (β), the difference in averages which is sought in the experiment (d') and standard deviation of the experimental error (s). Probabilities of the experimental errors and the difference that one desires to detect between means should be stipulated by the researcher. Therefore, only the estimate of the experimental error cannot be previously obtained, which must be obtained experimentally or by means of similar researches previously performed. Because in descriptive sensory analysis the most common approach to data analysis is the analysis of variance, the estimated standard deviation of the experimental error is obtained by the root mean square error (RMSE). Therefore, 574 RMSE values were obtained from previous published studies. The data collected was adjusted to Weibull probability distribution (1.8081, 0.11419), where five percentiles of the distribution are considered in the calculations. Determination of the number of evaluations necessary was performed using the procedure "sample size and power analysis" of the JMP/SAS software. Three probability levels were defined for type I and II errors, four levels of mean difference to be detected in the experiment, and five percentiles of RMSE distribution probabilities. The required numbers of evaluations in descriptive tests, considering these different experimental conditions, were calculated totaling 180 scenarios. Considering the median of the experimental error, a value of alpha (type I error) of 1%, a value of beta (type II error) of 5% and a difference between the average of 10% of the sensory scale, 33 evaluations are needed in the descriptive tests. Further considering that each judge evaluates the samples triplicate, 11 judges are necessary for this specific set of parameters. Other scenarios were also discussed in the paper.

1. Introduction

The descriptive sensory analysis of foods consists of the assessment of sensory characteristics by a team of judges, who identify and quantify the intensity of sensory stimuli present in food using the five human senses (sight, smell, hearing, touch and taste) (Murray, Delahunty, & Baxter, 2001). Descriptive assessment is a valuable tool in the various stages of food processing: development of new products, quality control, storage and shelf-life (Meilgaard, Civille, & Carr, 2006; Stone & Sidel, 2004).

Traditional descriptive evaluation techniques require the teaming of judges with a high degree of training, and the high number of judges may be an obstacle to the application of this valuable tool in industry (Heymann, Machado, Torri, & Robinson, 2012). A recommended ideal number of judges to make up a team is not clear from literature. Different recommendations are found depending on the technique used, for example, six judges for the Flavor Profile (Cairncross & Sjostrom, 1950); ten judges for the Texture Profile (Brandt, Skinner, & Coleman, 1963); and ten to twelve judges for the Quantitative Descriptive Analysis (Stone & Sidel, 1985). However, criteria for determining the number of judges needed are not shown.

On the other hand, teams composed of different sizes are verified in the generic techniques called "Conventional Profile" or "descriptive analysis". In published studies which utilize generic techniques, there is the use of six judges (Lee & Chambers, 2010; Warmund, Elmore, Adhikari, & McGraw, 2011), seven judges (Guàrdia, Aguiar, Claret, Arnau, & Guerrero, 2010; Chueamchaitrakun et al., 2011), eight judges (Perrin & Pagès, 2009; Anyango, De Kock, & Taylor, 2011), nine judges (Westad, Hersleth, Lea, & Martens, 2003; Campo, Ballester, Langlois, Dacremont, & Valentin, 2010; Silva et al., 2012; Tesfaye et al., 2010), ten judges (Leighton, Schonefeldt, & Kruger, 2010; Speaziale, Vásuez-Araujo, Mincione, & Carbonell-Barrachina, 2010; Silva et al., 2013), eleven judges (Bitnes, Ueland, Møller, & Martens, 2009; Plaehn, 2009; Moussaoui & Varela, 2010; Parr, Valentin, Green, & Dacremont,

* Corresponding author. Tel.: +55 31 38993810.
E-mail addresses: rita.navarro@ufv.br (R.C.S.N. Silva), vprm@ufv.br (V.P.R. Minim), alexandre.navarro@ufv.br (A.N. Silva), lminim@ufv.br (L.A. Minim).

0950-3293/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodqual.2013.07.010

2010), twelve judges (Bitnes, Ueland, Møller, and Martens, 2009; Biasoto et al., 2010; Sinesio, Comendador, Peparaio, & Moneta, 2007), fifteen judges (Garcia-Carpintero, Gomez-Gallego, Sanchez-Palomo, & Gonzales Viñas, 2011), seventeen judges (Guinard, Yip, Cubero, & Mazzucchelli, 1999) and twenty judges (Delgado & Guinard, 2011). In general, the lower limit is six judges and the upper limit is twenty judges.

Since the cost associated with sensory evaluation increases with the number of judges participating, it is important to determine the optimal number of evaluators necessary for sensory tests. According to Heymann et al. (2012), it is obvious that training a smaller number of judges requires less time, money and effort, but this may result in a "false savings" due to the possibility of obtaining "poor" data. Thus, the challenge is to determine the optimal number of judges needed in descriptive assessments that allows for reducing the size of the team, but without losses of information on the sensory profile of foods, the description of products and still permit for performing powerful statistical testing.

2. How to determine the ideal number of judges?

Calculating the number of judges for sensory testing has been little explored in literature. Some recent studies have been conducted for affective tests, in which the ideal number of consumers was determined for sensory acceptance tests. In these studies, the authors calculated the size of the sample using data obtained experimentally (Gacula & Rutenbeck, 2006; Mammasse and Schlich, 2014), with data obtained from a literature review (Hough et al., 2006) and by survival analysis (Hough, Calle, Serrat, & Curia, 2007; Libertino, Osornio, & Hough, 2011). For descriptive tests, resampling techniques with experimental data were used by King, Arents, and Moreau (1995), Pages and Périnel (2003), Gacula and Rutenbeck (2006) and Heymann et al. (2012).

In calculating the number of judges, four parameters must be known, using the concept "sample size and power of analysis", described in Kraemer and Thiemann (1987), Montgomery (2001) and Walpole, Myers, Myers, and Ye (2011), which are: (i) level α – probability of type I error, (ii) level β – probability of type II error, also expressed as power of the test $(1 - \beta)$, (iii) d' – difference in average which is sought in the experiment, and (iv) s – standard deviation of the experimental error. The probabilities of experimental errors (α and β) and the difference that one desires to detect between means shall be stipulated by the researcher. Therefore, only the experimental error cannot be previously obtained, which must be obtained experimentally or by means of similar operations previously performed.

3. Identifying the parameters

Knowledge of the definition and interpretation of the four parameters which must be known to calculate the number of judges is of utmost importance in determining the required number of evaluations in descriptive tests. In order to establish the criteria and method for obtaining these parameters, a brief explanation of these concepts will be presented.

3.1. Probability of type I and type II error

The decision errors denominated type I and type II are associated with the hypotheses of the statistical test used to verify the existence of significant differences between treatment means. The null hypothesis (H_0) represents equality between the means, which is tested against the alternative hypothesis (H_a) that opposes the decision that at least one mean differs from the others (Montgomery, 2001). When samples are used and decisions are made based on frequency distributions, there is the probability of committing errors in the decisions. Thus, when detected in a statistical test that the treatment means are different, there is a probability that they are equal considering the population, which characterizes the type I decision error, with a probability of occurrence α. Similarly, when detecting that the treatment means are equal, this conclusion is also subject to a decision error, which is termed as type II error with probability of occurrence β. Table 1 illustrates the decision errors associated with the statistical tests.

The probabilities of committing type I and type II errors should be established prior to conducting the experiment (Stone & Sidel, 2004). These probabilities are also used in the calculations of test power and sample size (Montgomery, 2001; Walpole et al., 2011), according to the research objective.

3.2. Difference in average (d')

The value of d' determines the magnitude of difference that must exist between the treatment means for detecting significance between the same, for given values of α and β (Hough et al., 2006). This magnitude is given in percent of the scale size. Thus, if using a scale of 9 cm and a d' value of 0.02 is stipulated, the treatment means must present a minimum difference 0.18 cm (2% of 9 cm) to be considered different from each other (Fig. 1A). For instance, in this case d' value is 2% on a 9 cm scale (i.e. the minimum difference is of 0.18 cm), does this mean that if the product A is in average at 6.0 and the product B is in average at 6.3, A and B are significantly different since the difference between them is 0.3 cm (>0.18 cm defined).

To avoid misinterpretation regarding the parameter d' used in this work and the value of delta from the Thurstonian model, this theory will be quickly addressed by making a comparison. In difference sensory tests the Thurstonian model is used to estimate the distance (δ) between the intensity means (μ_X and μ_Y) of the stimulus of two products, which is given by the number of standard deviations by which the two distributions are separated (Thurstone, 1927), Fig. 1B. The value of d' in this case is the estimate of the parameter δ. In the Thurstonian theory, it is assumed that the perceived intensity of the sensory stimuli follows a normal distribution and that both products (X and Y) exhibit the same standard deviation (Ennis, 1993; Ishii, Kawaguchi, O'Mahony, & Rousseau, 2007; Jesionka et al., 2013).

Thus, it is clear that the value of d' of the Thurstonian model is a measure of the number of standard deviations, which strictly depends on the distribution and variance of the perception of sensory stimuli. At the same time, the value of d' assumed in this study refers to a percentage value of the scale, i.e., sensory score. In this case no assumption of variance is required.

A value of d' equal to 1 in the Thurstonian theory (comparison tests) represents an equal distance between the means equal to one standard deviation and therefore the two stimuli may be confused (Meilgaard et al., 2006). On the other hand, in descriptive tests the value of d' equal to 1 indicates that the means should present 100% sensory difference in relation to the scale size, so that significant difference is detected. Thus, a treatment is anchored at the extremity "weak" on the scale and the other treatment

Table 1
Decision errors: type I and type II.

		Decision	
		Reject H_0	Do not reject H_0
Truth	H_0 true	Type I error (probability α)	Correct decision
	H_0 false	Correct decision	Type II error (probability β)