Food Research International 55 (2014) 137–149

Contents lists available at ScienceDirect

Food Research International

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

Review

Observations on the use of statistical methods in Food Science and Technology

Daniel Granato [a,*], Verônica Maria de Araújo Calado [b], Basil Jarvis [c]

[a] State University of Ponta Grossa, Food Science and Technology Graduate Programme, Av. Carlos Cavalcanti, 4748, Uvaranas Campus, 84030-900 Ponta Grossa, Brazil
[b] Federal University of Rio de Janeiro, School of Chemistry, Rio de Janeiro, Brazil
[c] The University of Reading, Department of Food and Nutrition Sciences, School of Chemistry, Food and Pharmacy, Whiteknights, Reading, Berkshire RG6 6AP, United Kingdom

ARTICLE INFO

Article history:
Received 4 August 2013
Accepted 17 October 2013

Keywords:
Correlation
ANOVA
Statistics
Regression
Normality
Multiple comparison tests

ABSTRACT

Statistical methods are important aids to detect trends, explore relationships and draw conclusions from experimental data. However, it is not uncommon to find that many researchers apply statistical tests without first checking whether they are appropriate for the intended application. The aim of this paper is to present some of the more important univariate and bivariate parametric and non-parametric statistical techniques and to highlight their uses based on practical examples in Food Science and Technology. The underlying requirements for use of particular statistical tests, together with their advantages and disadvantages in practical applications are also discussed, such as the need to check for normality and homogeneity of variances prior to the comparison of two or more sample sets in inference tests, correlation and regression analysis.

Contents

1. Introduction

Statistics is essentially a branch of mathematics applied to analysis of data. In Food Science, statistical procedures are required in the planning, analysis and interpretation of experimental work. Such work may include surveys of the chemical, physical (e.g. rheological), sensory and microbiological composition of food and beverages during development and manufacture, including changes to these properties as a consequence of process optimization. Other studies may look at the association between variables that require analysis of data to aid interpretation and presentation of the results. Appropriate statistical methods need to be used to assess and make

* Corresponding author.
E-mail addresses: granatod@gmail.com (D. Granato), calado@eq.ufrj.br (V.M. de Araújo Calado), basil.jarvis@btconnect.com (B. Jarvis).

0963-9969/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodres.2013.10.024

inferences about the factors that influence the responses; for example: evaluation of the effect of adding increasing concentrations of a fruit extract on the acidity and sensory acceptance of a product; or the assessment of the effects on the biochemical markers (inflammation, oxidative stress, etc.) in experimental animals treated with different doses of a food extract or ingredient.

In this sense, the use of statistical tools in food research and development is important both in academia and in industrial research in the food, chemical, and biotechnological industries. However, experience shows that many workers frequently select the wrong tests, or use the correct tests in wrong situations. For instance, many researchers often fail to pay attention to important concepts prior to comparing mean values. This may arise for one or more of the following reasons: a lack of interest in performing calculations, misinterpretation of statistical results or misuse of statistical software, among others. The rapid increase in computing power has had an important impact on the practice and application of statistics. Today, many software packages are available that facilitate statistical analysis of data; when used properly they provide a valuable tool to enable different types of statistical and mathematical analyses to be done rapidly. Such software packages take seconds to generate linear/non-linear models, draw graphs or resolve complex numerical algorithms that used to take a considerable amount of time using manual procedures.

The importance of proper application of statistics in Food Research cannot be ignored; it is essential if one is to understand data and make decisions that take account of the statistical variability of measurement and process control systems, summarize experimental results, etc. The objectives of this paper are: 1) to explain some concepts regarding data analysis in Food Science and Technology; 2) to provide some statistical information and 3) to discuss and present some published examples of mathematical modeling.

2. Concepts of statistics applied in Food Science

Use of the correct statistical tools is essential since the researcher needs to extract as much information as possible from experimental results. When work is published in a journal sufficient detail must be provided to permit the reader to understand fully the aims and outcome of the research and, should it be appropriate, for the work to be repeated. However, we observe that many published articles contain insufficient detail regarding the statistical tests used to interpret and discuss the published results. The reported analysis of results is often restricted to descriptive statistics (mean, median, minimum, maximum values, standard deviation and/or coefficient of variation). These, and other statistical tests such as correlation, regression, and comparison of mean values, are often based on the slavish use of 'statistical packages' that may, or may not, be appropriate for the purpose. It is essential that the researcher should take into consideration the basis of inferential statistical tests, prior to their application. Indeed, the researcher needs to understand the possible choices for relevant data analysis in order to plan experimental work appropriately and then to understand the results within a comprehensive data structure and draw conclusions based on the work.

Regardless of the type of experimental design a researcher uses, it is essential to test the statistical quality of the data prior to their further evaluation. If the quality of the data is poor, analysis of experimental data will often lead to misleading conclusions. Data may be of poor quality if, for example, insufficient samples have been tested; the samples have not been drawn randomly from the test population(s); the measurement uncertainty of the analytical method(s) used is large; the person doing the analysis is inadequately trained; or if the analytical results include 'censored' values. All of these considerations should be addressed prior to setting up an experimental plan and all are generally within the control of the researcher. Sometimes experimental results may fall outside the limits of an analytical method; for instance, the level of an analyte in a sample may be below the lowest limit or, more rarely, above the highest limit of detection or quantification of a method. Such results are referred to as left- or right-censored values, respectively. How should such results be handled? This is a subject much under discussion in many fields, including (food) chemistry, microbiology and toxicology and several questions still need to be addressed with respect to the suitability of the procedure used to handle censored data (Baert et al., 2007; Bergstrand & Karlsson, 2009).

Some workers merely record that results are less than (or more than) the limit value — in which case they cannot be included in a statistical analysis of data; some replace censored data by the corresponding limit of detection (LOD) (Govaerts, Beck, Lecoutre, le Bailly, & Vanden Eeckaut, 2005) and others choose to record the values as half the limit value (for left-censored data) (Granato, Caruso, Nagato, & Alaburda, in press; Tressou, Leblanc, Feinberg, & Bertail, 2004). Omission or *ad hoc* adjustment of such data can result in serious bias in analysis of the other results. Another widely used method is based on the replacement of censored data by random samples from a uniform distribution with zero as minimum and LOD as maximum (Govaerts et al., 2005). A procedure, known as the Tobit regression, for evaluation of censored data in food microbiology has been described by Lorrimer and Kiermeier (2007) — the concepts are equally applicable in other areas of Food Science.

Two characteristics of data sets must be considered prior to the application of any inferential tests:

1. Do the data conform to the principles of 'normality', i.e. to a 'normal' distribution (ND)?
2. Do the data satisfy an assumption of homoscedasticity, i.e. uniformity of variance?

What do we mean by a 'normal' distribution (ND)? A population ND can be described as a bell-shaped curve (Fig. 1) under which approximately 95% of values lie within the range mean (μ) $\pm$ 2 standard deviations (σ) and approximately 99% lie within the range $\mu \pm 3\sigma$. The standard deviation is a measure of the dispersion of values around the mean value and is determined as the square root of the variance, i.e. $\sigma = \sqrt{\sigma^2}$. The mean value ($\bar{x}$) and standard deviation (s) of a set of data obtained by analysis of random samples provide estimates of the population statistics.

If a number of random samples from a 'lot' or 'batch' of food, or indeed of other test matrix, is analyzed for some particular attribute (e.g. sugar content, acidity, pH level) it would be unrealistic to assume that the analytical results will be absolutely identical between the different samples, or even between subsamples of the same product. The reasons relate to the measurement uncertainty of the analytical method used for the test and the intrinsic variation in composition that occurs both within and between samples. We would therefore expect to obtain a range of values from the analyses. If only a few samples are analyzed, the results may appear to be randomly distributed between the lowest and highest levels (Fig. 2A); but if we were able to examine at least 20 samples, we would expect to obtain a distribution of results that conform reasonably well to a ND (Fig. 2B) with an even spread of results on either side of the mean value. However, in some cases, the distribution will not be 'normal' and may show considerable skewness (Fig. 2C) — such results would be expected, for instance, in the case of microbiological colony counts.

Since, for a ND, approximately 95% of results would be expected to lie within the range $\bar{x} \pm 2s$ we describe the lower and upper bounds of this range as the 95% Confidence Limits (CL) of the results; similarly, we describe the bounds of the 99% CL as $\bar{x} \pm 3s$. What this means is that 19 of 20 results of an analysis would be expected to lie within the bounds of the 95% CL, but by definition one result might occur outside this limits; similarly, one in 100 results might be expected to lie outside the 99% CL bounds. Results that do fall outside the CLs are often referred to as 'outliers' — whether such results

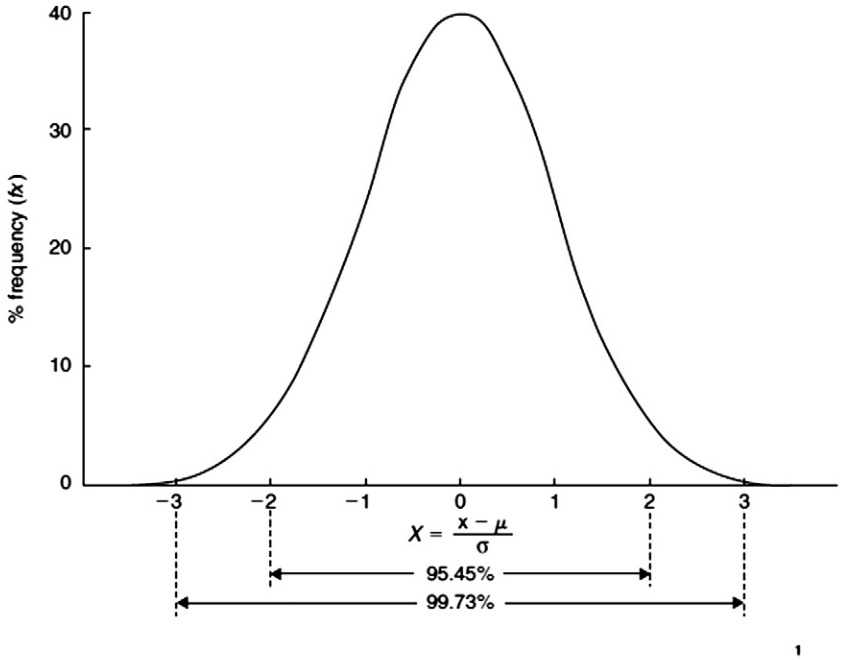


Fig. 1. A population normal distribution (ND) curve showing that approximately 95.45% of all results lie within ±2 standard deviations (s) of the mean and 99.73% lie within ±3s. Modified from Jarvis (2008).

are true values or occur because of faults in analytical technique can never be known, but it is essential to assess the frequency with which outliers occur. Various techniques exist for estimating the likelihood of outlier results (assuming ND) of which the most useful are those described by Youden and Steiner (1975), Anonymous (1994) and Horwitz (1995).

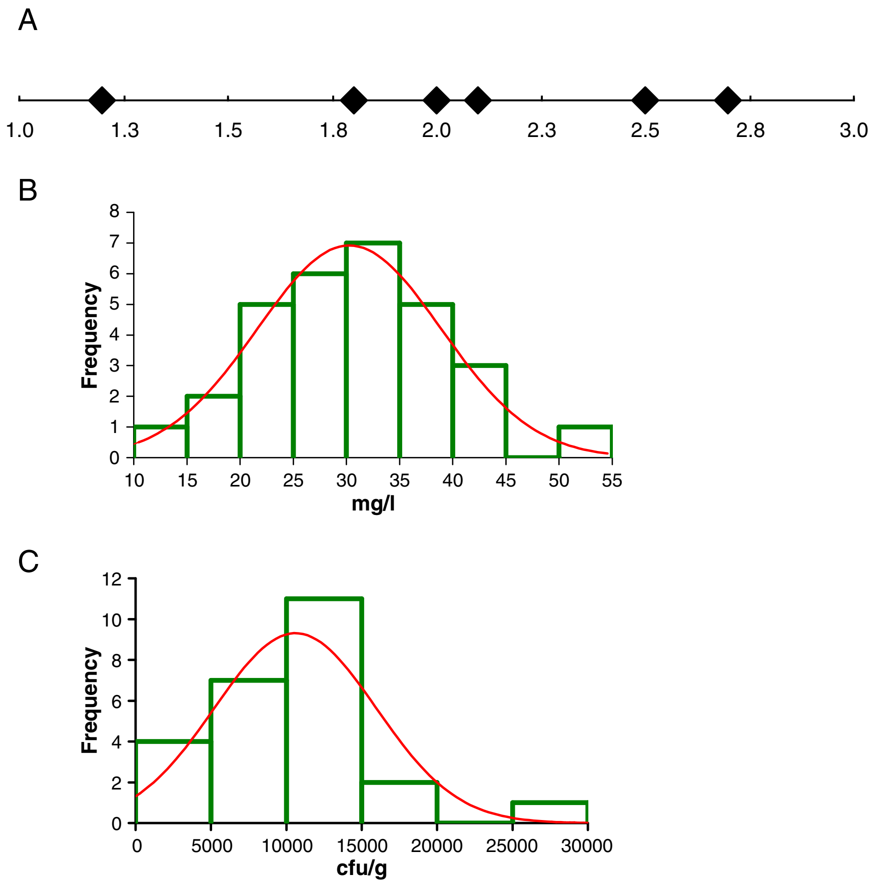


Fig. 2. (A) Plot of six analytical values, mean 2.05, SD = 0.532; range 1.20 to 2.70; (B) plot of analytical values from 30 replicate samples, overlaid with a ND curve for$\bar{x}$ = 30.3, s = 8.6. The data values are very slightly skewed but otherwise conform well to a ND; (C) plot of microbiological colony counts on 25 samples (as colony forming units/g) overlaid with a ND plot for $\bar{x}$ = 10,600 and s = 5360. Note that the data distribution shows a marked left-hand skew and kurtosis. The data do not conform to a ND.

Uniformity of variances is important in comparing results from two or more different sets of samples. If the variance of one set of results is much larger than that of a second set then it is not possible to use many standard parametric statistical tests to compare the mean values in a meaningful way. Fig. 3 shows distributions for two sets of ND data. Both sample sets have the same mean value of 10 g/l and but the standard deviation of sample set A ($s_A = \pm 0.25$) is only half that of sample set B ($s_B = 0.50$). Thus the variance of set B is four times greater than that of set A and 25% of the values under curve B fall outside the bounds of the 95% CL of set A. Such differences in variance show that the two distributions are very significantly different.

Often the researcher will wish to compare results from different samples; such tests may include comparison of mean or median values, determination of correlation and regression parameters, etc. Preliminary questions regarding the population(s) from which the samples were drawn need to be established in order to ensure that the correct analysis is chosen.

Experimental data can assume various forms: the distribution of data values for a measured variable following replicate analysis of samples may be either *continuous*, i.e. it can assume any value within a given range, or *discrete*, i.e. it can assume only whole number (integer) values. The latter generally applies to counts rather than measurements as in qualitative microbiological tests. In some situations, experimental variables may be '*nominal*', e.g. male or female gender selection for taste trials, '*categorical*', e.g. values can be sorted according to defined categories such as good, average or poor for qualitative taste tests, or '*ordinal*', e.g. values are ranked on a semi-quantitative basis using a predetermined scale such as hedonic taste panel scores. It is essential to understand data set designations in order to carry out appropriate analyses. Special nonparametric procedures are required for analysis of nominal, categorical and ordinal data sets.

In the next sections, we will focus on those statistical parameters that need to be determined and the underlying requirements for each test; and provide examples of the use of particular tests. From a practical standpoint, the analyst may choose to use statistical packages, either those that are free of cost (e.g. *R*, Action, Chemoface) or commercial software such as SAS (*Statistical Analysis Software*), Microsoft Excel, SPSS (*Statistical Package for Social Science*), Statistica, Statgraphics, Minitab, Design-Expert, and Prisma in order to design and analyze experimental data. However, an understanding of the underlying principles is vital to ensure that the correct tests are done.

2.1. Normality and homoscedasticity

2.1.1. Normality of data: is it truly important?

The normality of experimental results is an important premise for the use of parametric statistical tests, such as analysis of variance (ANOVA), correlation analysis, simple and multiple regression and *t*-tests. If the assumption of normality is not confirmed by relevant tests, interpretation and inference from any statistical test may not be reliable or valid (Shapiro & Wilk, 1965).

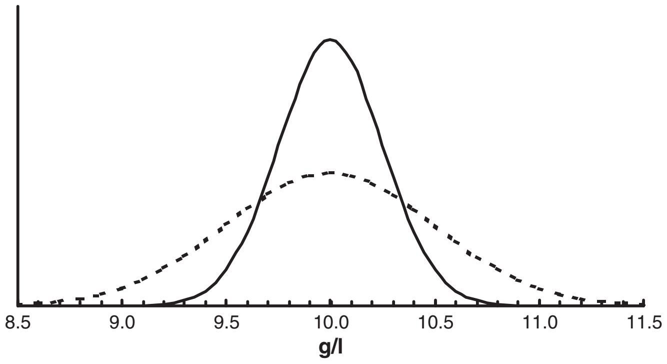


Fig. 3. Comparison of two ND curves both having $\bar{x} = 10$ g/l; curve A (——) has $s = 0.25$ and B has $s = 0.5$ (⋯⋯). Note that more than 25% of the data values for curve B fall outside the 95% CLs (9.5, 10.5) of the data in curve A.

Normality tests assess the likelihood that the given data set $\{x_1, \ldots, x_n\}$ conforms to a ND. Typically, the null hypothesis H_0 is that the observations are distributed normally, with population mean μ and population variance σ^2; the alternative hypothesis H_a is that the distribution is not normal. It is essential that the analyst identify the statistical distribution of the data. Most chemical constituents and contaminants conform well, or reasonably well, to a ND, but it is generally recognized that microbiological data do not. Whilst microbial colony counts generally conform to a lognormal distribution, the numbers of cells in dilute suspensions generally approximate to a Poisson distribution. The prevalence of very low levels of specific organisms, especially pathogenic organisms such as *Cronobacter* spp. and *Salmonella* spp., in infant feeds and other dried foods show evidence of over-dispersion that is best described by a negative-binomial or a beta-Poisson distribution (Jongenburger, 2012). Data from microbiological studies therefore require a mathematical transformation before statistical analysis is done (Jarvis, 2008). It is usual to transform microbial colony counts by using the $\log_{10}$ transformation although the natural logarithmic transformation (ln) is strictly the more accurate. Data conforming to a Poisson distribution is transformed to the square root of the count value. Other more complex transformations are required for negative binomial and beta-Poisson distributions (Jarvis, 2008).

In practice, there are two ways to check experimental results for conformance to a ND: graphically or by using numerical methods. The graphical method, usually displayed by normal quantile–quantile plots, histograms or box plots, is the simplest and easiest way to assess the normality of data; however, this method should not be used for small data sets due to lack of sufficient quantitative information (Razali & Wah, 2011). Numerical approaches are the best way to test for the normality of data, including determination of kurtosis and skewness; for example, tests such as those attributed to Anderson–Darling (AD), Kolmogorov–Smirnov (KS), Shapiro–Wilk (SW), Lilliefors (LF), and Cramér von Mises (CM). Frequently, people use histograms or probability plot graphs to test for normality (when they do!), but it can be risky since it does not provide quantitative proof that data follow ND. The shape of the graph depends on the number of samples examined and the number of bins used. Due to the small number of values the data shown in Fig. 4 do not appear to follow a normal distribution but the hypothesis of normality is not rejected by tests.

Razali and Wah (2011) studied the power and efficiency of four tests (AD, KS, SW, and LF) using Monte Carlo simulation and concluded that SW is the most powerful test for all types of distribution and sample

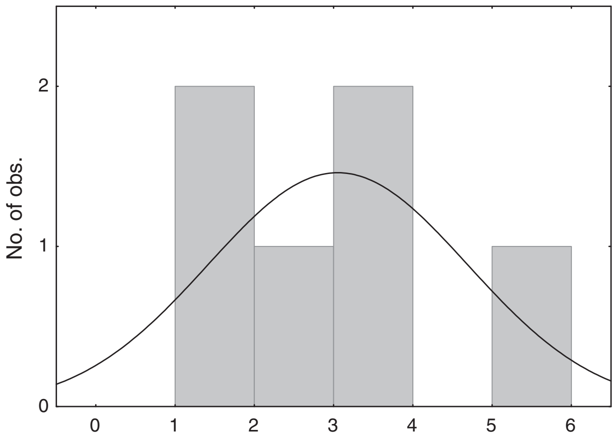


Fig. 4. Histogram of data values overlaid with a ND plot. Although the data do not appear to conform to a ND, tests for normality do not reject the null hypothesis due to the small number of data: Kolmogorov–Smirnov: $p_{KS} > 0.20$; Lillifors: $p_{Lillifors} > 0.10$, and Shapiro–Wilk: $p_{SW} = 0.68666$.

sizes, whereas KS is the least accurate test. They also confirmed that AD is almost comparable with SW and that LF always outperforms KS. Our experience in analyzing different types of experimental data (sensory, chemical, physicochemical, microbiological), suggests the use of SW to check for normality of data, regardless of the sample size. Ideally, *only one* test should be used to determine whether a data set conforms, or not, to normality and the conclusion must be based on the critical p-value of the test. If the test shows a $p < 0.05$ then the null hypothesis, that the data conform to normality, must be rejected; conversely, if $p \geq 0.05$ then the hypothesis of normality is not rejected. Using the example of Fig. 5, the SW test gives $p = 0.02355 < 0.05$, so the null hypothesis is rejected and the alternative hypothesis, that the data do not follow a normal distribution, is accepted but if the KS test had been used $p = 0.217 > 0.05$ so the hypothesis of normality is not rejected. The moral is to choose your test with care and to understand its limitations.

In sensory and microbiological studies, for example, it is very common to obtain results that do not follow a ND (Granato, Ribeiro, Castro, & Masson, 2010). Fig. 6 shows 50 observations of scores (0–100%) that describe the extent to which panelists find the flavor of a new food product acceptable. The distribution is slightly asymmetric and do not conform to a ND ($p < 0.05$ using the SW test) and the researcher would have to decide either to apply non-parametric statistics or to transform the results in order to use parametric tests. By using the square root or ln transformations and subjecting the transformed values to Shapiro–Wilk test, the new p-values would be 0.091 and 0.083, respectively. Therefore, either transformation would be suitable to transform these asymmetric values into normally distributed data.

2.1.2. Equality of variances: importance and concepts

It has been observed that some researchers pay no attention to the application of appropriate statistical techniques to validate experimental data. For all types of measurements, variance homogeneity, called homoscedasticity, should be assessed graphically and by a numerical test (Montgomery, 2009). This procedure is necessary in order to guarantee the correct application of tests for comparison of mean values and the user should always display the probability (p-value) of the test in text, tables or figures.

Model-based approaches usually assume that the variance is constant; however it is necessary to prove this by using an appropriate test. Distributions having a mean value (μ) = 0 and variance (σ^2) = 1 are called standard normal distributions and are often used to describe, at least approximately, any variable that tends to be distributed equally around the mean (Gabriel, 1964).

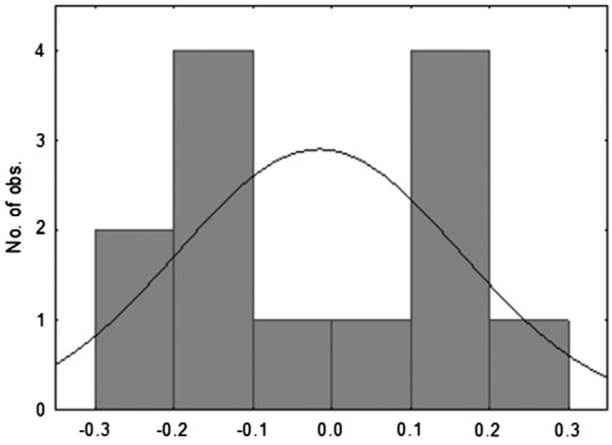


Fig. 5. Histogram of data values overlaid with a ND curve; the Shapiro–Wilk (SW) rejects the hypothesis for normality ($p = 0.0236$) but the Kolmogorov–Smirnov (KS) test ($p > 0.20$) does not reject the null hypothesis. The result from the Lilliefors tests ($p < 0.10$) is indeterminate.

The tests to check for homogeneity of variances require the following hypotheses: H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_\mathrm{k}^2$ and H_a: $\sigma_\mathrm{k}^2 \neq \sigma_\mathrm{l}^2$ for at least one pair (k, l). Assumption that the variance of data is homoscedastic when it is not causes serious violation of the operational requirements of many statistical tests and results, for instance, in overestimates of the goodness of fit as measured by the Pearson coefficient of regression analysis. Incorrect assumption of normality of variances may result in the misuse of parametric tests to compare mean values; for such data a suitable non-parametric test may be more appropriate.

Several tests are available to check the equality of variances on data from three or more samples and include those of Cochran, Bartlett, Brown–Forsythe and Levene; the F-test is used to check for homogeneity of two variances. In the example given above (Fig. 3) the ratio of the variances for data sets A and B, each comprising 1000 values, is four. Assuming that the 'degrees of freedom' in each case is infinity, the F-test value from standard statistical tables should not exceed a value of one. So since $F_\mathrm{A/B} = 4 \gg F_{\infty,\infty} = 1$, the difference in the variances is statistically significant at $p < 0.001$.

Levene's test for homogeneity of variances (Levene, 1960) is robust and is typically used to examine the plausibility of homoscedasticity for data from three or more samples. It is less sensitive to departures from normality than that of Bartlett's test (Snedecor & Cochrane, 1989), which should be used only if there is convincing evidence that the experimental results come from a normal distribution We strongly recommend the use of the Levene test to check for homogeneity of variances for sets of analytical, as well as for sensory (e.g. hedonic tests), data.

Thus, the researcher has two distinct situations: the variances are essentially equal or they are not. If the test shows that variances are heterogeneous, two possibilities exist: to use a non-parametric test or to transform the dependent variable in order to obtain a constant variance of the residues. Different types of transformations can be used, such as the logarithmic, Box–Cox, square root or inverse function transformations, depending on the distribution of the data. This approach can be used when the analytes do not follow a normal distribution (as in the case of microbiological and sensory data). Data transformation may normalize the distribution, stabilize the variances or/and stabilize a trend (Rasmussen & Dunlap, 1991). Parametric analysis of transformed data provides a better strategy than non-parametric analysis because the former is more powerful and accurate than the latter (Gibbons, 1993). However, it is important to keep in mind that the point of the transformation is to ensure the validity of the analysis (ND, equal standard deviations) and not to ensure a certain type of result (Rasmussen & Dunlap, 1991). It is worth noting that transformation should be avoided if possible since the transformed variable loses its absolute identity.

Another important issue that needs to be considered is the use of statistical software to check for homogeneity of variances. Take into consideration the following example: a researcher measures the content of a certain phenolic compound in a star fruit using different solvents (methanol, water, or ethyl acetate) by means of high-performance liquid chromatography (HPLC) and obtains the following results (expressed as mg/100 g of fruit pulp): methanol: 22.36; 22.45; 22.50; water: 13.30; 13.40; 13.55; ethyl acetate: 11.12; 11.22; 11.18. By applying the Bartlett's test using both Statistica and Action software, the p-value was 0.4970. On the contrary, when Levene's test was applied, p-values of 0.5393 and 0.3695 were obtained using Action and Statistica software, respectively. This is because of differences in the method of calculation: while the Levene's test implemented in Statistica performs an ANOVA on the deviations from the mean, Action software carries out the analysis on the deviations from the group medians (this is also known as the Brown–Forsythe test). Thus, the researcher needs to understand how the p-values are obtained in each statistical software rather to care about the number itself.

More details about tests to check for homogeneity of variances can be found in Levene (1960), Brown and Forsythe (1974), Lim and Loh (1996), Keselman and Wilcox (1999) and Gastwirth, Gel, and Miao (2009).

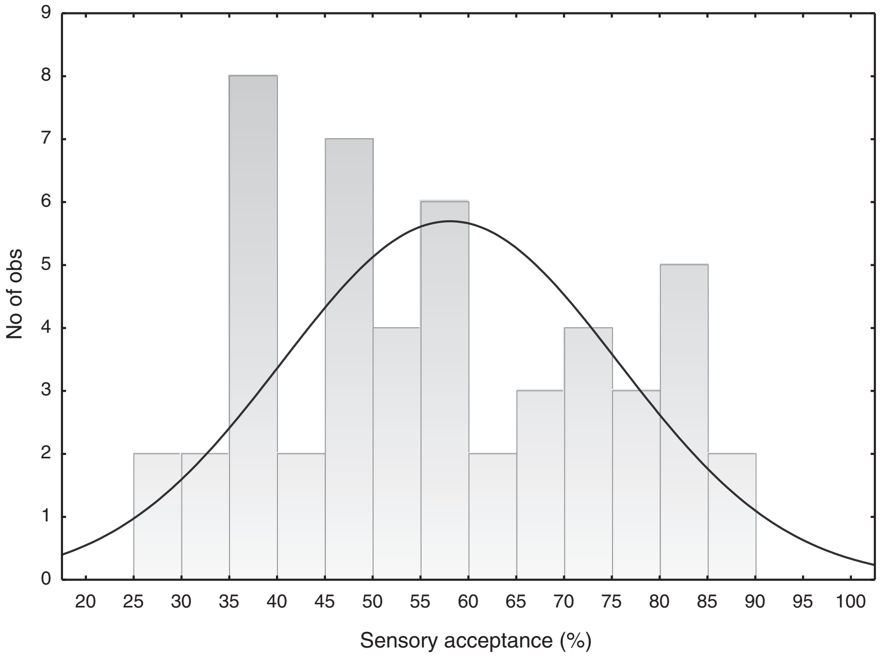


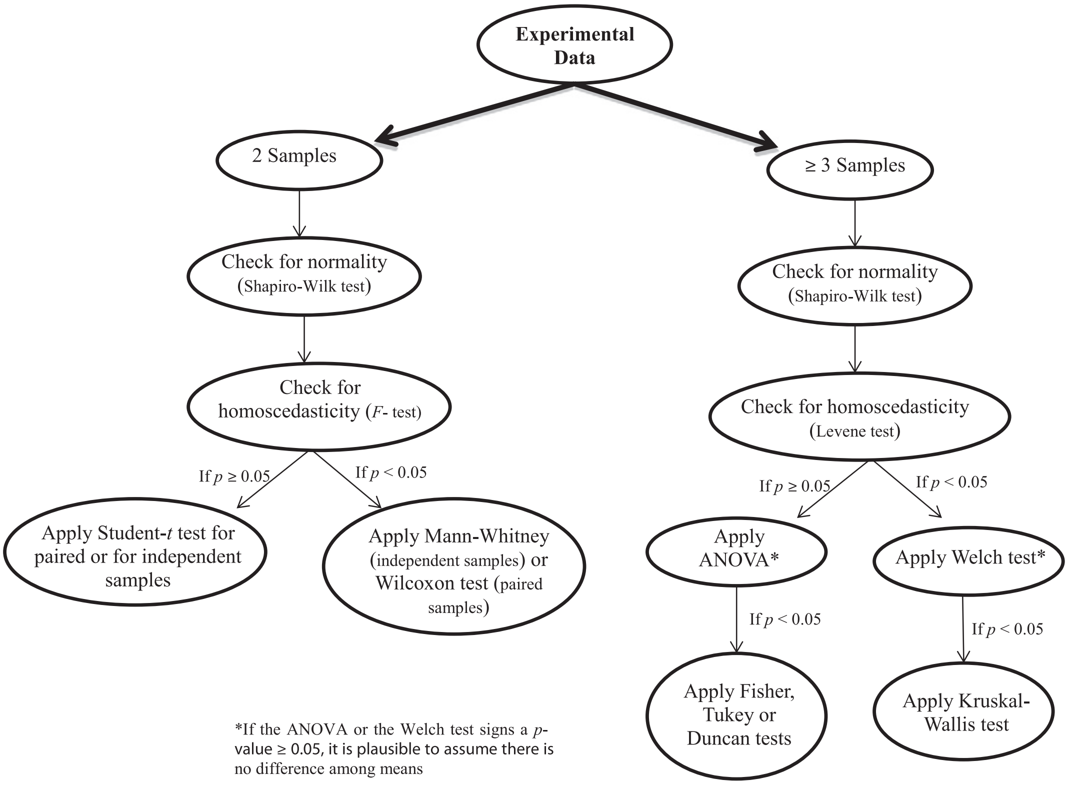
Fig. 6. Simulated scores for a sensory analysis of a new product development based on a scale ranging from 0 to 100% of acceptance, overlaid with a ND curve for $\bar{x} = 57.88$ and $s = 17.51$. The SW test for normality gives $p = 0.045 < 0.050$; thus the hypothesis of normality of the data is rejected.

2.2. Parametric statistics in Food Science

Depending on the statistical distribution of data, sample size, and homoscedasticity, samples and treatments can be compared using parametric or non-parametric tests. Parametric tests should be used when data are normally distributed and there is homogeneity of variances, as shown by the Shapiro–Wilk and Levene (or F) tests, respectively. Then, a Student's t-test is used to check for differences between two mean values or an ANOVA is used when three or more mean values need to be compared (Fig. 7).

2.2.1. Comparing two samples/treatments

When the mean values for a specific characteristic in two data sets are to be compared and both data sets are normally distributed and


Experimental Data
2 Samples
≥ 3 Samples
Check for normality (Shapiro-Wilk test)
Check for normality (Shapiro-Wilk test)
Check for homoscedasticity (F- test)
Check for homoscedasticity (Levene test)
If p ≥ 0.05
If p < 0.05
If p ≥ 0.05
If p < 0.05
Apply Student-t test for paired or for independent samples
Apply Mann-Whitney (independent samples) or Wilcoxon test (paired samples)
Apply ANOVA*
Apply Welch test*
If p < 0.05
If p < 0.05
Apply Fisher, Tukey or Duncan tests
Apply Kruskal-Wallis test

*If the ANOVA or the Welch test signs a p-value ≥ 0.05, it is plausible to assume there is no difference among means

Fig. 7. Statistical steps and tests to compare two or more samples in relation to a quantitative response variable.

have similar variance, a Student's *t*-test should be used. The null hypothesis (H_0) is that the mean values do not differ; the alternative (H_a) is that they do differ. However, if there are more than two data sets it is not correct to test each pair using a *t*-test (one of us recently received a paper in which 21 *t*-tests had been applied to compare individual pairs for seven sets of data!). The approach to be taken depends on whether the data are paired or independent and it is sometimes difficult to choose the correct version of the test. Samples are considered to be independent when they differ in nature and do not depend on one another. For example, Student's *t*-test for independent samples should be used to compare the ascorbic acid content of two cultivars of strawberries, or the production of an enzyme by two bacterial strains. A paired sample *t*-test is used if each of several samples are analyzed in parallel for the same characteristic by two different methods or if tests (e.g. blood sugar levels) are done on samples taken from subjects both before and after ingestion of a specific food ingredient. If the variances are not strictly equal, a correction factor (Welch's test) should be included in the statistical analysis. If the data do not conform to ND, non-parametric tests should be used.

2.2.2. Analysis of variances for three or more data sets

Analysis of variances (ANOVA) is a parametric statistical tool that partitions the observed variance into components that arise from different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are all equal. In this sense, the null hypothesis, H_0, says there are no differences among results from different treatments or sample sets; the alternative hypothesis (H_a) is that the results do differ. If the null hypothesis is rejected then the alternative hypothesis, H_a, is accepted, i.e. at least one set of results differs from the others. The ANOVA procedure should be used to compare the mean values of three or more data sets. One practical example of application of analysis of variance is provided by Oroian, Amariei, Escriche, and Gutt (2013): authors investigated the rheological behavior of honeys from Spain under different temperatures (25 °C, 30 °C, 35 °C, 40 °C, 45 °C, and 50 °C) and concentrations and compared the samples using one-way ANOVA followed by a test of multiple comparison of means.

Three alternative models can be used in an ANOVA: *fixed effects*, *random effects* or *mixed effects models*. The fixed effects model is appropriate when the levels of the independent variables (factors) are set by the experimental design. The random effects model, which is often of greatest interest to a researcher, assumes that the levels of the effects are randomly selected from an infinite population of possible levels (Calado & Montgomery, 2003). Independent variables may occur at several levels and it may be necessary to choose randomly only some levels; for instance, when samples are obtained randomly from four out of ten retailers of a particular product, or when three out of a possible seven brands of product are selected for analysis, provided always that the selection is done randomly. In certain circumstance some independent variables are assumed to be *random effects* and others to be *fixed effects*; here a *mixed model* should be used. In most experimental work a random effects model approach is often the most appropriate.

Depending on the number of factors to be analyzed, we can have:

1) A one-way ANOVA in which only one factor is assessed. This is the case for relatively simple comparisons of physicochemical, colorimetric, chemical and microbiological analytes (Alezandro, Granato, Lajolo, & Genovese, 2011; Corry, Jarvis, Passmore, & Hedges, 2007; Granato & Masson, 2010; Oroian, 2012). For example, if five samples of apple are analyzed for catechin content, the “apples” are the independent variable and the “catechin content” is the dependent response variable. Another important application of one-way ANOVA is when different groups of test animals that are treated with an extract/drug and compared to a control group (Macedo et al., 2013).
2) A 2-way ANOVA is used for two factors in which only the main effects are analyzed. The 2-way ANOVA determines the differences and possible interactions when response variables are from two or more categories. The use of 2-way ANOVA enables comparison and contrast of variables resulting from independent or joint actions (MacFarland, 2012). This type of ANOVA can be employed in sensory evaluation when both panelists and samples are sources of variation (Granato, Ribeiro, & Masson, 2012) or when the consistency of the panelists needs to be assessed;
3) A factorial ANOVA for *n* factors, that analyzes the main and the interaction effects is the most usual approach for many experiments, such as in a descriptive sensory or microbiological evaluation of foods and beverages (Ellendersen, Granato, Guergoletto, & Wosiacki, 2012; Jarvis, 2008; Mon & Li-Chan, 2007);
4) A repeated-measures (RM) ANOVA is used to analyze designs in which responses on two or more dependent variables correspond to measurements at different levels of one or more varying conditions. Benincá, Granato, Castro, Masson, and Wiecheteck (2011) used a RM-ANOVA to examine results from assessments of different instrumental color attributes for a mixture of juices from yacón (Peruvian ground apple) tubers and yellow passion fruit as a function of storage time.

The ANOVA approach provides a global analysis of the overall variance and assesses whether or not the variance of one or more, data sets differs significantly from the others but the output does not identify which variables differ. Thus, post-hoc tests need to be performed in order to specify exactly which pairs of means differ statistically. Various parametric and non-parametric tests can be used to compare the means of response variables, based on a normal or non-normal distribution of means, respectively. The choice of post-hoc tests to be used should be decided in advance so that bias is not attributed to any one set of data.

In recent times, so-called ‘robust’ ANOVA methods have been developed that are not affected by outlier data and can be used when data do not conform strictly to ND. They were developed following a need to analyze inter-laboratory studies during validation of analytical methods for use in chemistry and microbiology and are important also in determination of measurement uncertainty estimates that are nowadays required as part of laboratory accreditation (Elison, Rosslein, & Williams, 2000). Two approaches have been described: the first (Anonymous, 2001a) is based on a recursive application of an M-type estimator (Barnet & Lewis, 1978) and the second uses the Median Absolute Paired Deviation (MAPD) described originally by Rousseuux and Croux (1993). For explanation of these procedures refer to Hedges and Jarvis (2006) and Hedges (2008). Software for the MAD procedure can be downloaded from the Anonymous (2001b).

2.2.3. Post-hoc tests to compare three or more samples/treatments

Post-hoc tests are used for investigation of statistically significant differences ($p_{\mathrm{ANOVA}} < 0.05$) identified in an analysis of variance. When the mean values of three or more samples have homogeneous variances, the performance of such post-hoc tests in terms of Type I error (accepting equality of means when they are actually different) and Type II error (rejecting equality when they are not different) has been evaluated by many workers including Gabriel (1964), Boardman and Moffitt (1971), O'Neill and Wetheril (1971), Bernardson (1975), Conagini, Barbin, and Demétrio (2008), but there are still many unanswered questions regarding suitability. In practice, the choice of the best test to compare mean values depends on the investigator's experience. We recommend the use of Duncan's multiple range test (MRT) or Fisher's Least Significant Difference (LSD) test because of their high power to detect significant differences in mean values (Fig. 4) or Tukey's Honest Significant Difference (HSD) test.

The Tukey HSD is a single-step multiple comparison generally used in conjunction with an ANOVA to identify if one mean value differs significantly from another. It compares all possible pairs of means and is the most useful test for multiple comparisons. However, the method is not statistically robust, being sensitive to the requirement that the means need to follow a normal distribution. Even when there is a significant difference between a pair of means, this test often does not pinpoint it (Calado & Montgomery, 2003). However, many researchers claim that the Tukey test is the procedure of choice since it avoids Type II errors.

Fisher's LSD test is a statistical significance test used where sample sizes are small and when the distribution of the residues is normal ($p_{\text{Levene}} \geq 0.05$). The test is much more robust than Tukey but it is the most sensitive to Type I errors; yet it provides an important tool for comparing means after an ANOVA procedure (Carmer & Swanson, 1973). Duncan's MRT is not restricted to data conforming strictly to ND and does not require a significant overall 'between-treatments' F test but it is also likely to give Type I errors. The Tukey–Kramer HSD single-step multiple comparison procedure compares mean values using a 'Studentized Range', which is more conservative than the original Tukey HSD test, and can be used with unequal group sizes. It is much stricter than many other tests but is less likely to give Type I errors (Keppel & Wickens, 2004).

2.3. Non-parametric statistics in Food Science

Non-parametric procedures use ranked data rather than actual data values. The data are ranked from the lowest to the highest and each value is assigned, in order, the integer values from 1 to n (where n = total sample size) (Hollander & Wolfe, 1973). Non-parametric methods provide an objective approach when there is no reliable (universally recognized) underlying scale for the original data and when there is concern that the results of standard parametric techniques would be criticized for their dependence on an artificial metric (Siegel, 1956). Such tests have the obvious advantage of not requiring any assumption of normality or homogeneity of variance. Because they compare medians rather than means the comparison is not influenced by outlier values. The major disadvantage of non-parametric techniques is the lack of defined parameters and it is more difficult to make quantitative statements about the actual difference between populations. Ranking for non-parametric procedures preserves information about the order of the data but discards the actual values. Because information is discarded, non-parametric procedures can never be as powerful (i.e. less able to detect differences) as their parametric counterparts (Hollander & Wolfe, 1973).

Non-parametric tests are also used for nominal, categorical and ordinal data or when data have been assigned values on an arbitrary scale for which no definitive numerical interpretation exists, such as when evaluating preferences in sensory evaluation. Every non-parametric procedure has its peculiar sensitivities and blind spots. If possible, it is always advisable to run different nonparametric tests; if there are discrepancies in the results, one should try to understand why certain tests give different results.

2.3.1. Comparing two samples/treatments

Non-parametric tests can be used to examine data in a manner that is analogous to the use of paired and independent *t*-tests. Independent tests are evaluated using the Mann–Whitney test (Mann & Whitney, 1947) and paired tests by the Wilcoxon signed rank tests. Both require ranking of data as the first stage of the evaluation.

In what way are data unsuitable for parametric testing? Choice of non-parametric tests is predicated on examination of data that do not conform to a ND, especially data from samples drawn from different populations. For example, consider the comparison of a new simple method (B) with a standard method (A) for estimation of patulin in samples of an apple puree manufactured at different times from a number of individual ingredient sources. Since all ingredients come from different populations and the products are prepared individually, samples taken from any one 'lot' would come from a unique population but across 'lots' each population would be different. Since the same samples are tested by the two methods it is appropriate to consider the tests to be paired but as each sub-set of analyses is made on a different population of samples it is not appropriate to use a Student's *t*-test. The method of choice is the Wilcoxon Signed Ranks test (Wilcoxon, 1945) where the difference between the results for the two methods (A, B) is determined and a sign (+, −) is used to define whether method A or B is the greater. Values are then ranked in sequence from the smallest value upwards but without reference to the sign; the sign is then reinstated against the rank value and the sums of the + and − ranks are determined (Table 1). The lesser rank total is then compared with tabulated values for the Wilcoxon signed rank statistic (U) and the statistical significance is determined.

Another example might be where two independent methods have been used to swab chicken skins for levels of bacteria. Each area will be independent of all other areas and the two swabbing procedures are also completely independent. Once again we are dealing with non-normal populations from discrete work areas so we cannot use parametric tests to compare the results. In this case the procedure of choice is the Mann–Whitney U test. If we assume that there are n pairs of tests, ranks are allocated from 1 to $2n$ without regard for the data set from which they come (Table 2). The total rank scores for each method are then determined and a U statistic is calculated for each data set. The smaller value of U is then compared with tabulated values to determine the significance.

Full details of these tests, together with tables of significant values, are found in standard texts such a Snedecor & Cochrane (1989).

2.3.2. Comparing three or more samples or treatments

It will have been noted above that the key requirement for a parametric comparison of three or more variables using ANOVA is that the data must conform to, or be transformable, to a ND and the variances should be relatively homogeneous. If this is not possible, two non-parametric approaches can be used: the Kruskal–Wallis and the Friedman tests. Both tests rely on ranking of results but whilst the Kruskal–Wallis test ranks all results (with ties getting the average rank) before summating the rank values according to the treatment, in the Friedman test ranks are determined for each individual treatment.

The Kruskal and Wallis (1952) is a non-parametric multiple range test of differences in central tendency (median) that essentially provides a one-way analysis of variance for three or more independent samples based on ranked data. It is most often used for analysis of

Table 1
Use of the Wilcoxon Signed Ranks test to determine the level of patulin in lots of an apple compote (legal limit 25 μg/kg) using two analytical methods (A & B).

Production lot	Patulin (μg/kg)		Difference	Sign	Rank (R)	Signed R
	A	B	A − B			
1	12.5	10.5	2.0	+	8.5	+8.5
2	11.5	10.8	0.7	+	6	+6
3	12.5	13.0	−0.5	−	4	−4
4	12.0	12.0	0		−	−
5	14.0	12.0	2.0	+	8.5	+8.5
6	12.5	12.4	1.0	+	1	+1
7	14.0	12.3	1.7	+	7	+7
8	12.5	12.7	−0.2	−	−2	−2
9	14.0	13.5	0.5	+	4	+4
10	13.0	12.5	0.5	+	4	+4
Mean	12.85	12.17	Sum R+			+39
			Sum R−			−6

Tabulate the results for methods A & B, then determine the difference (A − B).
Ignoring the sign and any zero value allocate a rank to each difference, using an average rank if results are identical. Then allocate the relevant sign to each rank.
Add the rank scores for R+ and R− and, for the number of pairs (in this case n = 9), compare the smaller rank total with the tabulated value in tables of Wilcoxon's signed ranks. If, as in this case, the smaller rank total is ≤ published value then the difference is statistically significant at $p = 0.05$. Hence the null hypothesis, that results from both methods are equal, should be rejected as method A gives higher results. Whether the differences are of practical importance is another matter!

Table 2
Comparison of bacterial numbers on cotton and plastic sponge swabs taken from chicken neck skins immediately after evisceration.

Number of bacteria (CFU × 10^{-4}/25 cm^2)			
Cotton swab (A)		Sponge swab (B)	
Count	Rank	Count	Rank
110	14.5	20	7
16	6	200	20
24	8	5	2
105	13	89	11
155	18	125	16
2	1	140	17
104	12	180	19
10	4	49	10
7	3	15	5
28	9	110	14.5

Allocate ranks (1–20) across both sets of data; average the rank for identical counts (in this case, counts of 110).
Calculate the rank totals: $R_A = 88.5$; $R_B = 121.5$.
Calculate the U_A statistic for data set A: $U_A = [n_A(n_A + 1) / 2 + n_A n_B - R_A] = 66.5$.
Similarly, calculate the U_B statistic for data set B: $U_B = 33.5$.
Take the smaller value of U as the test statistic and compare it with the Mann–Whitney tabulated value for $p = 0.05$ with $n_A = n_B = 10$.
The calculated value of U_B (33.5) > U_{critical} (23) so the null hypothesis that both methods give equal results is not rejected — the 2-tailed probability is $p = 0.25$.

data having one 'nominal' variable and one 'continuous' variable. An example of its use might be the case of a food engineering study in which three different brands of filter pads are tested to assess their effectiveness in filtering beer. Replicate filters of each brand are assessed using one batch of raw beer to determine the period of time before each filter becomes blocked (as judged by e.g. the filter pressure); the work is then repeated on further batches of beer. The objective is to assess the cost-effectiveness of the different brands. The filter brands provide the 'nominal' values and the time to blocking is the continuous measurement. The test can be used for data with non-homogeneous variances or for data that do not conform to ND; it can be used when the groups are of unequal size, but the test assumes that the shape and distribution of data for each group is similar. The test is essentially an extension of the Mann–Whitney U test (*qv*), which can be applied to pairs of data as a post-hoc test of significance between pairs of results.

The Friedman (1937, 1939) test is used to determine differences in central location (median) for analysis of trials with one-way repeated measures having three or more dependent samples and is also based on ranking of data. Dependent samples might include, for instance, the assessment scores of k taste panelists who are judging n independent samples of wine — are the scores given by each panelist consistent or is there a significant difference between panelists and, if so, does this differ between the wine samples? The data consists of a matrix of n rows (the 'blocks') representing the wine samples and k columns (treatments) representing the panelists. Five trainee wine tasters have assessed the overall quality of each of six wines, using hedonic scores from 1 (excellent) through 3 (average) to 5 (awful). The results are summarized in Table 3. The data for each 'block' are ranked from lowest to highest, tied values each being given the average rank. The ranked data are then analyzed to determine the χ^2 value. If the null hypothesis that the mean scores do not differ is rejected it is then necessary to carry out post-hoc tests to determine the source(s) of the differences.

Recognizing that some variation in response is to be expected, the instructor wishes to know whether there is overall agreement between the tasters. The approach is to use the Friedman procedure that ranks the performance of tasters for each wine. If the taste score between two or more individuals is identical each receives the average rank in Table 4.

Table 3
Non-parametric analysis of the scores from a wine tasting.

Wine	Score for taster no.				
	1	2	3	4	5
A	1	1	2	1	3
B	5	5	4	5	5
C	2	3	3	1	3
D	1	3	1	3	2
E	1	2	1	2	2
F	2	2	2	4	2

The scores for each taster are totaled (R_k) and the square of the totals is determined (R_k^2). The number of tasters = $k = 5$; the number of samples = $n = 6$. We determine a value M using the equation: $M = \frac{12}{nk(k+1)} \sum R_k^2 - 3n(k+1)$.

For these data, M = 4.233 with $\upsilon = k - 1 = 4$ degrees of freedom. From the Tables, we find that the critical value for $\chi^2_{(p\,=\,0.05,\ \upsilon\,=\,4)}$ is 9.49 which is greater than M and therefore we do not reject the null hypothesis that the tasters have scored the samples uniformly.

Full details of these, and other nonparametric and parametric tests, are given in standard works including Sheskin (2011).

2.4. Bivariate correlation analysis

Correlation is a method of analysis used to study the possible association between two continuous variables. The correlation coefficient (r) is a measure that shows the degree of association between both variables (Granato, Calado, Oliveira, & Ares, 2013). This parametric test requires both data sets to consist of independent random variables that conform to ND. The correlation coefficient measures the degree of linear association between the two sets of data (A and B), and its value lies between −1 and +1. The closer the absolute value, $|r|$, is to 1, the stronger the correlation between the data values (Ellison, Barwick, & Farrant, 2009). The correlation between two variables is positive (Fig. 8A) if high values for one variable are associated with high values for the other variable and negative (Fig. 8B) if one variable is low when the other is high. A correlation close to r = 0 (Fig. 8C) indicates that there is no linear relation, or at best a very weak correlation, between the two variables. However, a low r-value does not necessarily imply that there is no relationship between the responses; a low value can be due to the existence of a non-linear correlation between these variables, but the presence of outliers in one or both data sets may also affect the r-value (Altman, 1999).

Table 4
Average rank for the wine tasters.

Wine (n)	Ranked score for taster (k)				
	1	2	3	4	5
A	2	2	4	2	5
B	3.5	3.5	1	3.5	3.5
C	2	4	4	1	4
D	1.5	4.5	1.5	4.5	3
E	1.5	4	1.5	4	4
F	2.5	2.5	2.5	5	2.5
Total (R_k)	13	20.5	14.5	20	22
R_k^2	169.0	420.3	210.3	400.0	484.0

Many workers calculate the Pearson linear correlation coefficient in order to seek to determine the strength of association between data sets. However, when more than five variables are analyzed, the analysis is compromised because correlation coefficients do not assess simultaneous association among results for all variables, which makes it difficult to understand and interpret the structure and patterns of the data. For example, if one considers five sets of response variables (A, B, C, D and E), it is necessary to calculate the correlation coefficients, and their significance, for each data set pair, i.e. AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE. It is easy to understand and interpret up to three correlations coefficients but, in order to better understand multiple responses, a more sophisticated multivariate statistical approach, such as principal component analysis, clustering techniques, linear discriminant analysis should be used (Besten et al., 2012, 2013; Granato et al., in press; Zielinski et al., in press).

When large sets of results ($\geq$30) are analyzed, data should be formally checked for normality. If the data do not follow a normal distribution a non-parametric approach, such as the Spearman's rank correlation coefficient, should be used to analyze for any correlation between the responses. Fig. 9 shows the steps to follow when two data sets (each with $n \geq 8$) are to be analyzed with respect to correlation. Spearman's correlation coefficient (ρ) should be used when either or both data sets do not conform to ND, when the sample size is small, or when the variables are measured as ordinals i.e. first, fifth, eighth, etc. in a sequence of values. The Spearman correlation coefficient does not require the assumption that the relationship between variables is linear.

One good example to compare both Pearson and Spearman correlation coefficients can be obtained by analyzing the data sets below:

A: 12.56; 14.46; 16.65; 25.68; 16.80; 28.95; 32.25; 30.33; 32.81; 28.29; 29.98; 30.32; 33.57

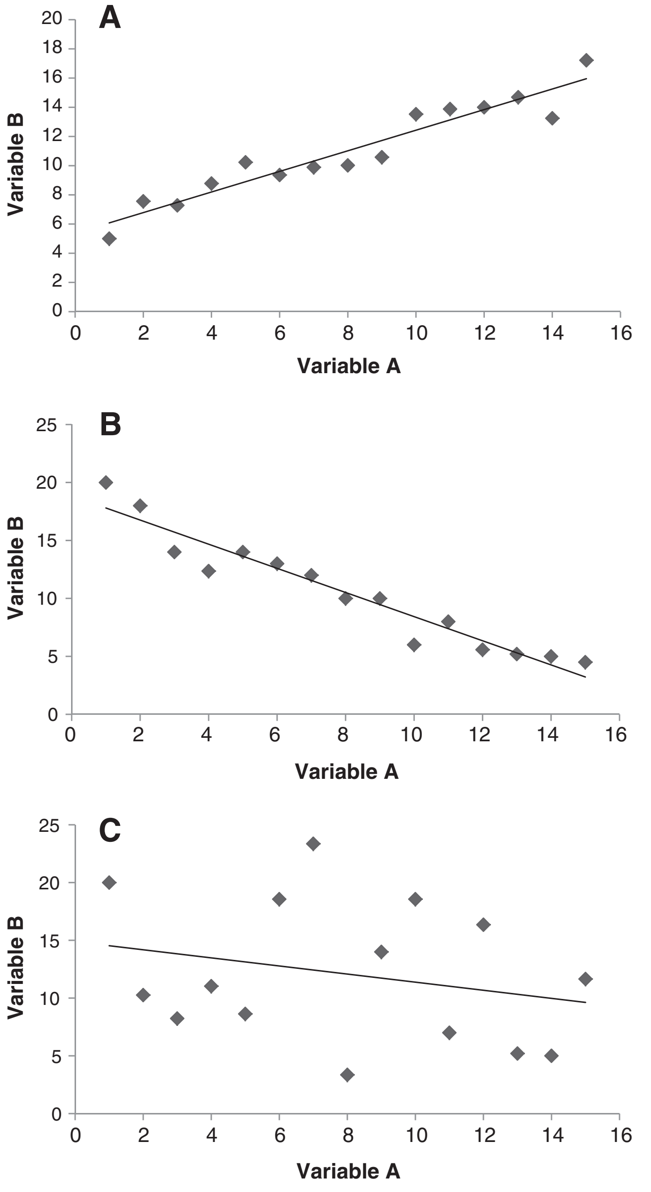


Fig. 8. (A) Positive correlation, (B) negative correlation, and (C) almost null correlation.

B: 53.25; 65.33; 53.68; 62.74; 61.53; 64.89; 66.40; 60.99; 68.50; 56.30; 66.36; 68.25; 73.89.

The normality of both data sets was assessed using the Shapiro–Wilk test and *p*-values of 0.017 and 0.605 were obtained for A and B, respectively. Since one data set does not follow a normal distribution, Spearman correlation rank coefficient ($\rho = 0.753$, $p = 0.004$) should be used rather than the Pearson correlation coefficient ($r = 0.660$, $p = 0.014$). As observed in this simple example, it is possible to assume that depending on the method employed to assess correlation between two response variables, the coefficient magnitude and its significance may be highly different, and the conclusion (inferences) of the study may be misleading or even wrong.

There is no scientific consensus about the qualitative assessment of correlation coefficients, that is, whether a correlation coefficient is truly strong, moderate or weak. Granato, Castro, Ellendersen, and Masson (2010) established an arbitrary scale for the strength of correlations between variables using the following criteria: perfect ($|r| = 1.0$), strong ($0.80 \leq |r| < 1.0$), moderate ($0.50 \leq |r| < 0.80$), weak ($0.10 \leq |r| < 0.50$), and very weak (almost none) correlation ($0.10 \leq |r|$).

There are two major concerns regarding correlation tests: the significance of the correlation and the interpretation of results. Firstly, to assume a statistically significant association between variables the *p*-value of the correlation coefficient should be <0.05. Granato, Katayama, and Castro (2011) showed that with large data sets, the correlation coefficient is often statistically significant even at a moderate or low *r* value. On the other hand, when the data set is small (Granato, Freitas, & Masson, 2010), high values of *r* may be observed but the statistical probability of the correlation is not significant ($p > 0.05$). Secondly, when a correlation coefficient is calculated, it is not always possible to assume causation because

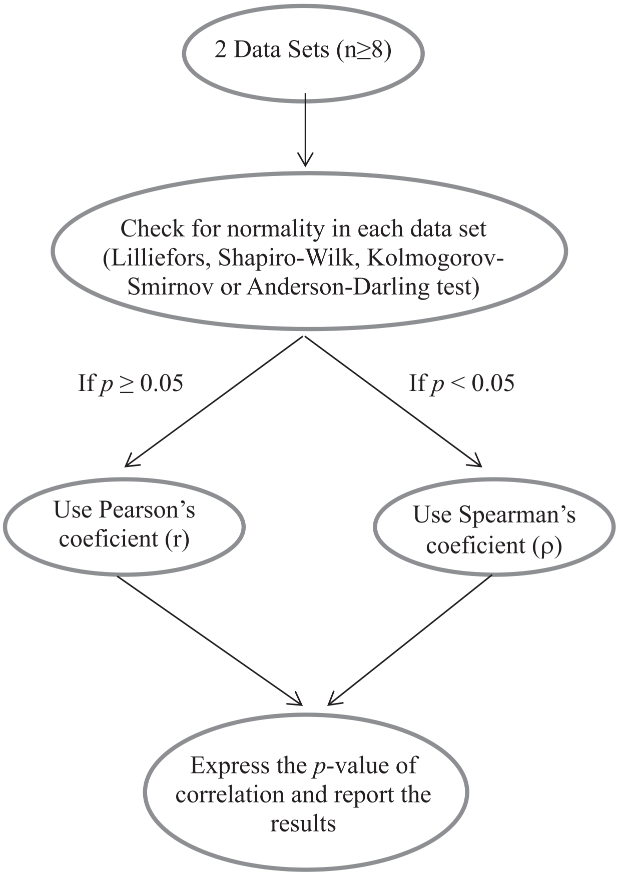

2 Data Sets (n≥8)
Check for normality in each data set (Lilliefors, Shapiro-Wilk, Kolmogorov-Smirnov or Anderson-Darling test)
If p ≥ 0.05
If p < 0.05
Use Pearson's coeficient (r)
Use Spearman's coeficient (ρ)
Express the p-value of correlation and report the results

Fig. 9. Steps to follow when two data sets (usually with $n \geq 8$) are to be analyzed for correlation.

co-variants may contribute to the response, i.e. correlated occurrences may be due to a common cause. For example, suppose one researcher, studying the effects of increasing sugar levels on the sensory acceptance of coffee using a consumer panel, obtains a correlation coefficient of $+0.72$ with $p = <0.05$ but another researcher in another place, using the same trial conditions repeats the test and obtains very different correlation coefficients and probability value. The explanation is possibly related to differences in consumer habits and experience but might also reflect the variety and composition of the coffee beans, their roasting and the preparation of coffee for the test. Hence, although correlation studies can be extremely useful, they do not automatically imply a cause and effect relationship between the variables. Another aspect of correlation that often needs to be considered is the intra-class correlation coefficient. This provides a measure of the correlation between two sets of variables when the data are paired or can be organized into groups. It is analogous to the use of a paired, rather than an independent, *t*-test. The procedure, described in Bland and Altman (1996), estimates the proportion of total variance that is due to the between-group component; it has many applications but is particularly useful for examination of taste panel results from examination of replicate samples of a product.

Although a correlation coefficient provides a measure of the strength of a relationship between two sets of data, it does not prove equivalence between the results. Lack of equivalence is due to bias and can be determined using the Bland and Altman (1986, 1995) procedure. For example, comparison of two analytical methods may give results that are highly correlated but when bias is assessed, one method may give consistently larger or smaller values for the same set of samples; thus the two methods do not provide equivalent results.

2.5. Regression analysis

Regression analysis is used to examine the relationship between sets of dependent and independent variables and includes many techniques for modeling and analyzing two or more sets of data. There are many forms of regression including linear, multi-linear, probit and logistic approaches, the latter being particularly important in studying biological responses to dose levels of inhibitory or stimulatory treatments. In its simplest form, regression is used to assess the relative effects of e.g. increasing treatment times (or temperatures) on the heat resistance of microorganisms in order to determine thermal effects (D-values) under defined process conditions. It may be used also to establish calibration curves for chemical, physical and biological assays for *continuous* data sets. Calibration curves require at least five concentration levels including a blank value with adequate (at least triplicate) replication of tests at each concentration level.

Assumptions for regression require that:

- The samples are representative of the population for the inference prediction;
- The concentrations of the independent variables are measured with zero error, i.e. they are ‘absolute’ values; if this is not the case then the more complex orthogonal linear regression must be used in order to correct for errors in the predictor variable (cf. Carroll, Ruppert, Stefanski, and Crainiceanu (2012)
- The error term of the estimates is a random variable with zero mean;
- The predictors are linearly independent, i.e. each value cannot be expressed as a linear combination of the other values;
- The variance of the errors is homoscedastic.

In order to perform a regression analysis it is essential to:

- Test data for the presence of outliers (at 95 or 99% of confidence) using the Grubb's test for each concentration level;
- Ensure the homogeneity of variances in the concentration levels of the calibration curve by using one the tests described above (Section 2.1.2).
- Build the model (i.e. the graph) to display the analyte concentration versus the response (absorbance, area, etc.);
- Test the significance of the regression and its lack of fit through the F-test and a one-factor ANOVA. Provided that the response is directly and linearly correlated to the concentrations then the regression coefficient should be significant. Evidence for lack of fit ($p < 0.05$), may be due to a non-linear response, to excessive variation in the replicates at one or more of the test values or the use of an over-extended independent variable range. In this case, removing the highest values and repeating the statistical analysis should reduce the range of concentrations. Evidence for lack of linearity may indicate that a nonlinear model (quadratic, for example) might be more appropriate for the method, and therefore, alternative models should be evaluated.
- Determine the following statistical parameters by means of the regression analysis:
  - The regression equation ($y = ax + b$), where y is the dependent estimate at independent concentration level (x), a is the slope of the line and b is the linear intercept when $x = 0$;
  - The standard deviation of the estimated parameters and model;
  - The statistical significance of the estimated parameters;
  - The coefficient of determination (R^2; regression coefficient) and the adjusted R^2.

The regression model is considered suitable to the experimental data when:

1. The standard deviation of the parameter is at least 10% lower than the corresponding parameter value;
2. The standard deviation of the proposed mathematical model is small;
3. The parameters of a model are statistically significant otherwise they will not contribute to the model.

It is a myth to consider that if $R^2 > 90\%$ the model is excellent (Montgomery & Runger, 2011). This is only one criterion to evaluate the goodness of fit of the model. If R^2 is low (<70%), the mathematical model is not good; on the other hand, if R^2 is high (>90%), it means that you should continue the analysis and check the other criteria. It is noteworthy that in some applications, depending on the type of analysis, e.g. evaluation of sensory data, the coefficient of determination may be considered good if $R^2 > 60\%$;

4. The statistical significance, obtained from the *F*-test of an ANOVA analysis of the proposed mathematical model is at least $p < 0.05$;
5. Analysis of the residuals (experimental value for a response variable minus value predicted by the mathematical model) must conform to ND and have a constant variance, as described above. This is a necessary condition for the application of some post-hoc tests such as *t* and *F*.

It is important to recognize that the regression and correlation coefficients describe different parameters. Regression describes the goodness of fit of a model; correlation estimates the linear relationship of two variables.

A common mistake is to use R^2 to compare models. R^2 is always higher if we increase the order of a model (linear in comparison to quadratic, for example). For example, a third order polynomial has a higher R^2 than a second order polynomial because there are more terms, but it does not necessarily mean that the first is the better model. An analysis of the degrees of freedom (number of experimental points minus number of parameters from the model) needs to be carried out. A model with more terms requires estimation of more coefficients so fewer degrees of freedom remain. Thus, another criterion needs to be used: the adjusted regression coefficient $-$ R^2_{adj}. This coefficient adjusts for the number of explanatory terms in a model relative to the number of data points and its value is usually less than or equal to that of R^2. When comparing models, the one with the highest adjusted coefficient is the best model.

We have noted above that in addition to simple 'Generalized Linear Models' of regression other, more complex, forms of regression are available for use in specific circumstances. The reader is directed to other works, such as Kleinbaum, Kupper, and Azhar (2007), for information and guidance on such procedures.

2.6. Other statistical techniques

We have discussed above the most frequently used univariate and bivariate statistical techniques. However, other statistical and mathematical procedures, especially chemometrics, and including Principal Component Analysis, Cluster Analysis, Discriminate analysis, K-nearest neighbors and other complex techniques (neural networks) can be considered to be extensions of these methods developed for specific purposes. Examples of several approaches to analysis of complex data using such procedures include the analysis of sensory (Keenan, Brunton, Mitchell, Gormley, & Butler, 2012), physicochemical (Gómez-Meire, Campos, Falqué, Díaz, & Fdez-Riverola, 2014), microbiological (Zhou et al., 2013), metabolomics (Oms-Oliu, Odriozola-Serrano, & Martín-Belloso, 2013) and chemical data (Granato et al., in press; Zielinski et al., in press).

3. Final remarks

In this paper, we have sought to explain the use of some the more common statistical tests in analysis of data generated in Food Science and Technology by means of theoretical and practical examples. We have tried, so far as is practical, to avoid the use of statistical jargon and to provide reference to more advanced works where appropriate. The reader is encouraged to ponder the advantages and disadvantages of these tests in practical applications and to apply the most suitable methods for analysis of their experimental data.

References

Alezandro, M. R., Granato, D., Lajolo, F. M., & Genovese, M. I. (2011). Nutritional aspects of second generation soy foods. *Journal of Agricultural and Food Chemistry*, *59*, 5490–5497.

Altman, D.G. (1999). *Practical statistics for medical research* (8th ed.)Boca Raton: Chapman & Hall/CRC, 611.

Anonymous (1994). Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic methods for the determination of repeatability and reproducibility of a standard method. *ISO 5725-2:1994*. Geneva, Sw: International Organization for Standardisation.

Anonymous (2001a). *Robust statistics: A method of coping with outliers.* AMC technical brief, number 6. London: Royal Society of Chemistry.

Anonymous (2001b). *MS Excel add-in for robust statistics:* Royal Society of Chemistry, Analytical Methods Committee ([Online] http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/RobustStatistics.asp (last accessed 24 June 2013))

Baert, K., Meulenaer, B., Verdonck, F., Huybrechts, I., Henauw, S., Vanrolleghem, P. A., et al. (2007). Variability and uncertainty assessment of patulin exposure for preschool children in Flanders. *Food and Chemical Toxicology*, *45*(9), 1745–1751.

Barnet, & Lewis (1978). *Outliers in statistical data.* Chichester, UK: Wiley.

Benincá, C., Granato, D., Castro, I. A., Masson, M. L., & Wiecheteck, F. V. B. (2011). Influence of passion fruit juice on colour stability and sensory acceptability of non-sugar yacon-based pastes. *Brazilian Archives of Biology and Technology*, *54*, 149–159.

Bergstrand, M., & Karlsson, M.O. (2009). Handling data below the limit of quantification in mixed effect models. *The AAPS Journal*, *11*(2), 371–380.

Bernardson, C. S. (1975). Type error rates when multiple comparison procedures follow a significant test ANOVA. *Biometrics*, *31*, 229–232.

Besten, M.A., Jasinski, V. C. G., Costa, A. G. L. C., Nunes, D. S., Sens, S. L., Wisniewski, A., Jr., et al. (2012). Chemical composition similarity between the essential oils isolated from male and female specimens of each five *Baccharis* species. *Journal of the Brazilian Chemical Society*, *23*, 1041–1047.

Besten, M.A., Nunes, D. S., Sens, S. L., Wisniewski, A., Jr., Granato, D., Simionatto, E. L., et al. (2013). Chemical composition of essential oils from male and female specimens of *Baccharis trimera* collected in two distant regions of Southern Brazil: A comparative study using chemometrics. *Quimica Nova*, *36*, 1096–1100.

Bland, J. M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 307–310.

Bland, J. M., & Altman, D.G. (1995). Comparing methods of measurement: Why plotting difference against standard method is misleading. *Lancet*, *346*, 1085–1087.

Bland, J. M., & Altman, D.G. (1996). Measurement error and correlation coefficients. *British Medical Journal*, *313*, 41–42.

Boardman, T. J., & Moffitt, D. R. (1971). Graphical Monte Carlo type Y error rates, for multiple comparison procedures. *Biometrics*, *27*, 738–744.

Brown, M. B., & Forsythe, A.B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, *69*, 364–367.

Calado, V. M.A., & Montgomery, D. C. (2003). Planejamento de Experimentos usando o Statistica. *Rio de Janeiro: e-papers* (1st ed.) (Available at: www.e-papers.com.br (last accessed 2 July 2013))

Carmer, S. G., & Swanson, M. R. (1973). Evaluation of ten multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association*, *68*, 66–74.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2012). *Measurement error in non-linear models* (2nd ed.): Chapman Hall/CRC Press.

Conagini, A., Barbin, D., & Demétrio, C. G. B. (2008). Modifications for the Tukey test procedure and evaluation of the power and efficiency of multiple comparison procedures. *Scientia Agricola*, *65*(4), 428–432.

Corry, J. E. L., Jarvis, B., Passmore, S., & Hedges, A. J. (2007). A critical review of measurement uncertainty in the enumeration of food micro-organisms. *Food Microbiology*, *24*, 230–253.

Elison, S. L. R., Rosslein, M., & Williams, A. (Eds.). (2000). *Quantifying uncertainty in analytical measurement* (2nd ed.): Eurachem/Citac.

Ellendersen, L. S. N., Granato, D., Guergoletto, K. B., & Wosiacki, G. (2012). Development and sensory profile of a probiotic beverage from apple fermented with *Lactobacillus casei*. *Engineering in Life Sciences*, *12*(4), 475–485.

Ellison, S. L. R., Barwick, V. J., & Farrant, T. J.D. (2009). *Practical statistics for the analytical scientist — A bench guide.* Cambridge: RSC Publishing.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675–701.

Friedman, M. (1939). A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *34*, 109.

Gabriel, K. R. (1964). A procedure for treating the homogeneity of all set of means in analysis of variance. *Biometrics*, *20*, 459–477.

Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, *24*, 343–360.

Gibbons, J.D. (1993). *Nonparametric statistics: An introduction.* Newbury Park: Sage Publications.

Gómez-Meire, S., Campos, C., Falqué, E., Díaz, F., & Fdez-Riverola, F. (2014). Assuring the authenticity of North West Spain white wine varieties using machine learning techniques. *Food Research International.* http://dx.doi.org/10.1016/j.foodres.2013.09.032.

Govaerts, B., Beck, B., Lecoutre, E., le Bailly, C., & Vanden Eeckaut, P. (2005). From monitoring data to regional distributions: A practical methodology applied to water risk assessment. *Environmetrics*, *16*, 109–127.

Granato, D., Calado, V., Oliveira, C. C., & Ares, G. (2013). Statistical approaches to assess the association between phenolic compounds and the in vitro antioxidant activity of *Camellia sinensis* and *Ilex paraguariensis* teas. *Critical Reviews in Food Science and Nutrition*. http://dx.doi.org/10.1080/10408398.2012.750233.

Granato, D., Caruso, M. S. F., Nagato, L. A. F., & Alaburda (in press). Feasibility of different chemometric techniques to differentiate commercial Brazilian sugarcane spirits based on chemical markers. *Food Research International.* http://dx.doi.org/10.1016/J.FOODRES.2013.09.044 (in press).

Granato, D., Castro, I. A., Ellendersen, L. S. N., & Masson, M. L. (2010). Physical stability assessment and sensory optimization of a dairy-free emulsion using response surface methodology. *Journal of Food Science*, *73*, 149–155.

Granato, D., Freitas, R. J. S., & Masson, M. L. (2010). Stability studies and shelf life estimation of a soy-based dessert. *Ciência e Tecnologia de Alimentos*, *30*, 797–807.

Granato, D., Katayama, F. C. U., & Castro, I. A. (2011). Phenolic composition of South American red wines classified according to their antioxidant activity, retail price and sensory quality. *Food Chemistry*, *129*, 366–373.

Granato, D., & Masson, M. L. (2010). Instrumental color and sensory acceptance of soy-based emulsions: A response surface approach. *Ciência e Tecnologia de Alimentos*, *30*, 1090–1096.

Granato, D., Ribeiro, J. C. B., Castro, I. A., & Masson, M. L. (2010). Sensory evaluation and physicochemical optimisation of soy-based desserts using response surface methodology. *Food Chemistry*, *121*(3), 899–906.

Granato, D., Ribeiro, J. C. B., & Masson, M. L. (2012). Sensory acceptability and physical stability assessment of a prebiotic soy-based dessert developed with passion fruit juice. *Ciência e Tecnologia de Alimentos*, *32*, 119–125.

Hedges, A. J. (2008). A method to apply the robust estimator of dispersion, Qn, to fully-nested designs in the analysis of variance of microbiological count data. *Journal of Microbiological Methods*, *72*, 206–207.

Hedges, A. J., & Jarvis, B. (2006). Application of 'robust' methods to the analysis of collaborative trial data using bacterial colony counts. *Journal of Microbiological Methods*, *66*, 504–511.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods.* New York: John Wiley & Sons, Inc.

Horwitz, W. (1995). Protocol for the design, conduct and interpretation of method performance studies. *Pure and Applied Chemistry*, *67*, 331–343.

Jarvis, B. (2008). *Statistical aspects of the microbiological examination of foods* (2nd ed.) London: Academic Press.

Jongenburger, I. (2012). *Distributions of microorganisms in foods and their impact on food safety.* (PhD Thesis). NL: University of Wageningen.

Keenan, D. F., Brunton, N.P., Mitchell, M., Gormley, R., & Butler, F. (2012). Flavour profiling of fresh and processed fruit smoothies by instrumental and sensory analysis. *Food Research International*, *45*, 17–25.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researchers handbook* (4rd ed.) Upper Saddle River, NJ: Pearson.

Keselman, H. J., & Wilcox, R. R. (1999). The 'improved' Brown and Forsyth test for mean equality: Some things can't be fixed. *Communications in Statistics — Simulation and Computation*, *28*, 687–698.

Kleinbaum, D., Kupper, L., & Azhar, N. (2007). *Applied regression analysis and multivariate methods* (4th ed.) Thomson Brooks/Cole; Duxbury, Belmont, Ca, USA.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Society*, *47*, 583–621.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292): Stanford University Press.

Lim, T. S., & Loh, W. Y. (1996). A comparison of tests of equality of variances. *Computational Statistics and Data Analysis*, *22*, 287–301.

Lorrimer, M. F., & Kiermeier, A. (2007). Analyzing microbiological data — Tobit or not Tobit? *International Journal of Food Microbiology*, *116*, 313–318.

Macedo, L. F. L., Rogero, M. M., Guimarães, J. P., Granato, D., Lobato, L. P., & Castro, I. A. (2013). Effect of red wines with different in vitro antioxidant activity on oxidative stress of high-fat diet rats. *Food Chemistry*, *137*, 122–129.

MacFarland, T. W. (2012). *Two-way analysis of variance — Statistical tests and graphics using R.* Springer, 150.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50–60.

Mon, S. Y., & Li-Chan, C. Y. (2007). Changes in aroma characteristics of simulated beef flavour by soy protein isolate assessed by descriptive sensory analysis and gas chromatography. *Food Research International*, *40*, 1239–1248.

Montgomery, D. C. (2009). *Design and analysis of experiments* (5th ed.)New York: Wiley.

Montgomery, D. C., & Runger, G. C. (2011). *Applied statistics and probability for engineers* (5th ed.)New York: Wiley.

O'Neill, R., & Wetheril, G. B. (1971). The present state of multiple comparison. *Journal of the Royal Statistical Society*, *33*, 218–250.

Oms-Oliu, G., Odriozola-Serrano, I., & Martín-Belloso, O. (2013). Metabolomics for assessing safety and quality of plant-derived food. *Food Research International*. http://dx.doi.org/10.1016/j.foodres.2013.04.005.

Oroian, M. (2012). Physicochemical and rheological properties of Romanian honeys. *Food Biophysics*, *7*, 296–307.

Oroian, M., Amariei, S., Escriche, I., & Gutt, G. (2013). Rheological aspects of Spanish honeys. *Food Bioprocess Technology*, *6*, 228–241.

Rasmussen, J. L., & Dunlap, W. P. (1991). Dealing with non-normal data: Parametric analysis of transformed data vs nonparametric analysis. *Educational and Psychological Measurement*, *51*, 809–820.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*, *2*, 21–33.

Rousseuux, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*, 1273–1283.

Shapiro, S. S., & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.): Chapman-Hall/CRC.

Siegel, S. (1956). *Nonparametric statistics.* New York: Mc Graw-Hill Book Company, Inc.

Snedecor, G. W., & Cochrane, W. G. (1989). *Statistical methods* (8th ed.)Iowa, USA: Iowa State University Press.

Tressou, J., Leblanc, J. Ch., Feinberg, M., & Bertail, P. (2004). Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: Application to ochratoxin A. *Regulatory Toxicology and Pharmacology*, *40*, 252–263.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83.

Youden, W. J., & Steiner, E. H. (1975). *Statistical manual of the AOAC.* Washington DC, USA: Association of Official Analytical Chemists.

Zhou, Y., Pan, Z., Li, Y., Kang, X., Wang, X., Geng, S., et al. (2013). Epidemiological analysis of *Salmonella* isolates recovered from food animals and humans in eastern China. *Food Research International*, *54*, 223–229.

Zielinski, A. A. F., Haminiuk, C. W. I., Alberti, A., Nogueira, A., Demiate, I. M., & Granato, D. (in press). A comparative study of the phenolic compounds and the in vitro antioxidant activity of different Brazilian teas using multivariate statistical techniques. *Food Research International*. http://dx.doi.org/10.1016/j.foodres.2013.09.010 (in press).