ISSN: 2278-5213

Review

Methods and Application of Statistical Analysis in Food Technology

Bereket Abraha Gherezgihier[1,2*], Abdu Mahmud[1,2], Melake samuel[2] and Negasi Tsighe[2]
[1]School of Food Science and Technology, Jiangnan University, Wuxi 214122, PR China; [2]Dept. of Marine Food and Biotechnology, Massawa College of Marine Science and Technology, Eritrea
bereketmft@yahoo.com*; +86-186-1667226

Abstract

Statistics is a very broad subject with applications in a vast number of fields such as food technology, food engineering, pharmaceutical, biological and social sciences. Statistical methods are important not only in food technology, but also in other aspects to detect trends, evaluate food safety, food quality, consumer preferences, 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 review article is to present the important methods and applications of statistical analyses in food technology namely; basic statistics, descriptive statistics, collection of data, correlation and regression analyses, food sampling plan, testing of hypothesis and non-parametric statistical techniques as well as highlighting their uses based on some illustrations of tables and figures. The underlying requirements for use of particular statistical tests, together with their advantages or significance of methods and applications are also discussed.

Keywords: Basic statistical analysis, data collection, food technology, correlation, regression.

Introduction

Food issues are becoming important to consumers, most of who depend on the food industry and other food workers to provide safe, nutritious and palatable products (Keenan *et al.*, 2012). Data obtained not only from laboratory experiments, but also via surveys on consumers, as they are the users and receivers of the end products. Understanding such diverse information demands an ability to be at least aware of the process of analyzing data and interpreting results. In this way, communicating information is valid. This knowledge and ability gives undeniable advantages in the increasingly numerate world of food technology, but it requires that the practitioners have some experience with statistical methods (Granato *et al.*, 2012). Basic statistical concepts such as population size, sample size, sample space, variance, distribution, standard deviation, T-tests, hypothesis and so on are much needed in food technology to provide safe and quality food for consumers and people. Application of statistical methods in food technology is continually progressing and developing. Statistical analysis was identified, two decades ago, as one subject in a set of 'minimum standards' for training of food technologists at undergraduate level (Iwaoka *et al.*, 1996). Kravchuk *et al.* (2005) emphasized on the importance of application of statistical knowledge in the teaching of food technology disciplines, so as to ensure an ongoing familiarity by continual use. Some advantages of being conversant with statistics are obvious. An appreciation of the basis of statistical methods can aid making of conclusions and decisions on future work. Other benefits include the increased efficiency achieved by taking a statistical approach to experimentation (Granato *et al.*, 2013).

Significance of statistics in food technology

It is possible to evaluate scientific data without involving statistical analysis. Once data accumulate and time is limited, such judgment can suffer from errors. In these cases, simple statistical summaries can reduce large data blocks to a single value. Now, both the enlightened novice and the experienced analyst can judge what the statistics reveal. Consequent decisions and actions will now proceed with improved confidence and commitment (Keenan *et al.*, 2012). Additionally, considerable savings in terms of time and finance are possible. In some instances, decision-making based on the results of a statistical analysis may have serious consequences. Quantification of toxins in food and nutrient content determination rely on dependable methods of chemical analysis. Statistical techniques play a part in monitoring and reporting of such results. This gives confidence that results are valid and consumers benefit in the knowledge that certain foods are safe and that diet regimes can be planned with surety (Granato *et al.*, 2013).

*Corresponding author

Table 1. Application of statistics in the food technology (Ellendersen *et al*., 2012).

Method	Application
Summaries of results	Tables, graphs and descriptive statistics of instrumental, sensory and consumer measures of food characteristics
Analysis of differences and relationships	Research applications on differences in food properties due to processing and storage; correlation studies of instrumental and sensory properties
Monitoring of results	Statistical control of food quality and parameters such as net filled weight
Measurement system integrity	Uncertainty of estimates for pesticides and additives levels in food
Experimental design	Development and applications of balanced order designs in sensory research

Thus, manufactures and consumers both benefit from the application of these statistical methods. Generally, statistics provides higher levels of confidence and uncertainty is reduced. Food practitioners apply statistical methods, but ultimately, the consumer benefits (Tressou *et al*., 2004).

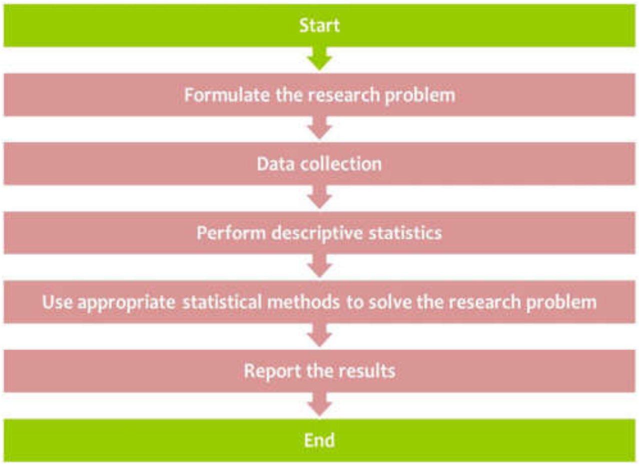
Applications of statistical analysis in food technology

There are many applications of statistics in the field of food technology. One of the earliest was in agriculture (Granato *et al*., 2012). Fisher (1966) used experimental design to partition variation and to enable more precise estimation of effects in crop plot experiments. There was even an early sensory experiment on tea tasting and since then statistical applications have increased as food technology emerged as a distinct applied science subject. Some forms of statistical applications in food are mentioned in Table 1. Preparation of data summaries is one general application of statistics that can be applied across the board. It is one of the simplest applications and can be done manually if necessary, depending on the requirements. A variety of simple graphs and table methods are possible, which allow rapid illustration of results. These summaries are taken further in statistical quality control where measures such as the mean value are plotted 'live', as a process is ongoing (Ellendersen *et al*., 2012). The graphs (control charts) used include limit lines which are set by using other statistical methods, which allow detection of out-of-limit material, for instance. Food product packs which are below statutory minimum net weight. Statistical methods can also be applied to evaluate the trustworthiness of data obtained by any method of measurement (Blumberg *et al.*, 2005). Food research application brings in analysis of differences and relationships, where hypotheses can be put forward on the basis of previous work or new ideas and then magnitudes of effects in sample statistics can be assessed for significance, for instance, examination of the change in colour pigment content during frozen storage of vegetables (Collis and Hussey, 2003). Examination of relationships requires that different measurement systems are applied and then compared. There are many examples of this in studies of food where data from instrumental, packaging, sensory and consumer sources are analyzed for interrelationships (Kirk and Sawyer, 1999).

Statistical data analysis in food technology

In any type of field, the goal of statistics is to gain understanding from data. Any data analysis should contain certain steps (Fig. 1) to be followed to ultimately achieve the goal of the research proposed (Freund, 2001; Shahbaba, 2012). Weiss (1999) stated in his work findings that to conclude an analysis in processed food of its quality, safety what so ever section, it should be noted that the major objective of statistics is to make inferences about processed food or population from an analysis of information contained in sample data. This includes assessments of the extent of uncertainty involved in these inferences.

Fig. 1. Steps in data analysis (Adapted from Freund, 2001).


Start
Formulate the research problem
Data collection
Perform descriptive statistics
Use appropriate statistical methods to solve the research problem
Report the results
End

Basic statistics

Statistics is a very broad subject, with applications in a vast number of different fields such as in food technology. In general, one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information. Putting it in other words, statistics is the methodology which scientists and mathematicians have developed for interpreting and drawing conclusions from collected data of different fields. Everything that deals even remotely with the collection, processing, interpretation and presentation of data belongs to the domain of statistics, and so does the detailed planning of that precedes all these activities (Shahbaba, 2012).

*Corresponding author

Basic concept and principles

Know-how in the basic concepts and principles is very crucial in the field of food technology. Randomness, population, data, variables, sample space, distribution are some of the basic principles and concepts of statistics (Blumberg *et al.,* 2005).

Randomness: This is a philosophically quite complex concept at any case, but in most cases it is satisfactory to think of it as lack of predictability in an 'experiment of food processed with many possible outcomes. An experiment with randomness in it is sometimes referred to as a stochastic experiment (Ellison *et al*., 2009).

Population versus data: In order to apply statistics properly in food aspects, it is necessary to distinguish between the amount of food processed (population) which is thought of as the underlying 'reality' that should to be understood and the data (in statistics often called a sample) that are collected in order to tell us something about this reality (Mann and Whitney, 1947). It can seldom or never obtain exact information about the population, but the data can, if collected in a reasonable way, be used to estimate or approximate important characteristics of it. The larger the sample is, the more precise information can be found about the population. The sample can be taken at random or in a more systematic way depending on situation.

Stochastic variable: The term stochastic variable (or just variable) refers to a quantity of food harvested, processed etc. that is measured in a stochastic experiment. The value of a stochastic variable cannot be predicted exactly; it will always have a component of randomness in it.

Sample space (or domain): This is the set of possible values/outcomes of a stochastic variable. Different types of food can be manufactured and needs to be assessed quality and safety before going out where by inspectors should know the samples space before doing the evaluation. The distribution of a stochastic variable describes how likely the different outcomes of the stochastic variable are (Ellison *et al*., 2009). It should be mentioned that the statistical term sample is a bit problematic in the food technology since it is used very differently in statistics and in applied science. In statistics, it is the term used for a set of observations taken from a certain population while in other disciplines, like for instance in chemistry, it is a term used for a physical object for which the stochastic variable is measured (Mann and Whitney, 1947).

Nature of data in food technology

Data are presented in many forms. In their original 'as-measured' state, data are 'raw' and they may convey readable details, but usually food processing' is required before they constitute information in the form of results (Greenfield and Southgate, 2003). Commonly, data exist as numbers, but other forms are possible. These can be letters, words, images, symbols, sounds, sensations, etc. Food technology data analyses have their own units for measurement of parameters such as absorbance, colour units, texture units, retention times (chromatography), etc. (Table 2). Other measurements can occur without units and use numbers to express counts of the occurrence of objects, for example the number of bacterial colonies on an agar plate, or the number of people preferring one food sample to another (Hanaf and Akiers, 2006).

Table 2. Examples of measurement types in food science (Adapted from Tressou *et al*., 2004).

Measurement/quantity	Example (typical unit)
Weight	Balance reading (gram)
Force	Mechanical firmness (newton)
Absorbance	Spectrophotometer (absorbance)
Colour	Spectrophotometer
Extensibility	Extensometer units
Sensory intensity	Intensity scale for bitterness
Rotation	Polarimeter (degree)
Volume	Colorimeter (temperature ◦C)

Population

Appropriate application of certain statistical analysis demands that there is some knowledge of the population from which the samples originate. The term 'population' refers generally to all the objects or subjects which make up a defined group (Blumberg *et al.,* 2005). Ultimately, it is the actual measurements or readings on the sample units themselves that make up the population under consideration. For example, in examination of the fat content of bacon samples, the fat content (%) values provide the data and a population of these values is assumed. A population of data would be real (rather than assumed) if all the food material within the defined population were to be analyzed (Ellison *et al*., 2009). Sample data can be displayed and analyzed in various ways to give a view of certain characteristics, which relate to the population. Such knowledge is fundamental to the understanding of sampling and statistical analysis, from organization of data to estimation of population parameters and significance testing. The first of these aspects relates to the distributional form of the population (MacFarland, 2012).

Collection of data and sampling of food

During experimentation of food, the readings or recordings, whether from instrument readout or a consumer survey sheet, provide the data. To obtain these, a sample is gathered from a defined population of produced food.

*Corresponding author

Table 3. Relationship of population with sample types (Adapted from Keenan *et al*., 2012).

Population	Sample nature	Sample unit	End determination subsample
Raw material batch	Weighed/measured amounts	Each amount	One or more parts of the amount
Final products of 1 day's production	Sample of the unit packs	Each pack	One or more measured amounts from each pack
Consumer population of China	Sample of consumers	Each consumer	
Food prototype batch	Weighed/measured amounts	Each amount	One or more parts of the amount

Sampling is a fundamental procedure in most investigations of food and food products, but it needs to be justified and placed in context (Ellison *et al*., 2009). Greenfield and Southgate (2003) explained that, sampling is viewed as a process of taking a fraction from a 'parent lot of food', source or population. Ideally, all of the population would be sampled (100% sample) because in one sense there would be 100% confidence that a 'trustworthy' result had been obtained. Instead of an estimate in the form of a statistic, the actual population parameter itself would be determined. However, with a large population of food and limited resources this is not possible–it would be too expensive, too time consuming and there are other disadvantages (Tressou *et al*., 2004). Thus, the practitioner has to rely on an estimate via a sample and this exposes the determination to sampling error. The nature of the sample itself involves definition of some terms and more detailed explanations. A sample of food or employees may have different units such as batch and lot and is made up of individual items, objects or people referred to as sampling units (Keenan *et al*., 2012). As indicated above, in food studies, the units can be weighed amounts of food, or individual consumer respondents in a survey, etc. In either case, the total quantity of the food material or the whole consumer population will not be subjected to measurement and a sample is taken as a small fraction. Relation of population and sample is necessary in evaluation of food and some illustration of the form of such procedure is shown in Table 3.

Descriptive statistics

At any conditions this type of statistics is necessary. Descriptive statistics is useful food technology to describe the basic features of a sample, providing a summary or profile. It consists of measures of central tendency, distribution, and dispersion. Descriptive statistics provide a "profile" of the subject in study conditions of foods, and may include means and standard deviations of food processed, packaged and often percent within categories of characteristics such as quality, safety, color, heat treated, and ranking preference (Keenan *et al*., 2012). The first stage of analyses is often those that summarize the data. Under this heading comes descriptive statistics, which are methods used to summarize the characteristics of a sample, for example, the average value, but which also includes displays with graphs and tables. Excel charts are produced by selecting the columns of data and then choosing the chart wizard (Johnson and Bhattacharyya, 1992).

Tabular and graphical displays

Graphical displays are basic form which provides indications of how the sample data are distributed (Shahbaba, 2012). This is also possible using table forms such as a frequency table. Graphs and tables can also be used to display descriptive statistics, which include a number of summary values such as the measures of central tendency and variation. Graphs and charts have the advantage of giving a more rapid overview, and they can indicate possible trends, effects and relationships. Graph icons on their own will mean that values will have to be estimated from the axes, unless the software allows a numerical display superimposed on the icons (Razali *et al*., 2011).

Summarizing nominal data

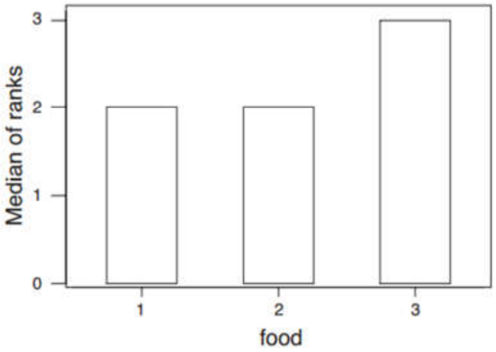
Many food researchers use nominal data which can be summarized using simple tables or graphs. These would indicate frequency of occurrence of the categories in the data. A graph or list of frequencies shows the distribution of the data in the sample. The method chosen for display must ensure that each item of data is treated exclusively and displayed once, i.e. there must be no overlap in the frequency counts (Keenan *et al*., 2012).

Summarizing ordinal data

Food industries can produce their food in a condition of the consumer acceptability (Shahbaba, 2012). With these data, order is present in the scale used as ranks or as points on an ordinal scale. Bar charts are more appropriate than pie charts as the order of the original scale is displayed along a horizontal or vertical axis. A ranking test produces ordinal data (Razali *et al*., 2011). These data can also be summarized by frequency tables and simple bar charts, but the results may be unsatisfactory. Ranking more than three items generates a compound bar or column graph with several bars per item.

*Corresponding author

Fig. 2. Median bar chart for preference ranking (Adapted from Razali *et al*., 2011).



Summaries such as median values for a ranking test can be displayed as a table and as a median bar chart (Fig. 2). Using excel would require calculation of the medians separately, but Minitab does this directly from the raw data using the chart option, which can plot summary measures (Fig. 2). In the example there are three food items, two have equal medians and the third is ranked lower in preference. Such like illustrations are quite important to figure out food that can be mostly liked (Johnson and Bhattacharyya, 1992).

Hypothesis testing and confidence intervals for the mean

A common aim in many studies is to check whether the data agree with certain predictions. These predictions are hypotheses about variables measured in the study. Hypotheses arise from the theory that drives the research. When a hypothesis relates to characteristics of a population, such as population parameters, one can use statistical methods with sample data to test its validity. A significance test is a way of statistically testing a hypothesis by comparing the data to values predicted by the hypothesis. Data that fall far from the predicted values provide evidence against the hypothesis. All significance tests have five elements: assumptions, hypotheses, test statistic, p-value, and conclusion. All significance tests require certain assumptions for the tests to be valid. These assumptions refer, e.g., to the type of data, the form of the population distribution, method of sampling, and sample size (Johnson and Bhattacharyya, 1992).

The above section handled estimation of the mean and variance. Sometimes one is also interested in testing hypotheses about the mean. A typical situation is that a certain product standard has a mean equal to a fixed value and one is interested in testing whether the actual distribution has a mean which is comparable to this standard value. The estimated value is in practice always different from the standard because it is a random variable, but the question is whether it is different enough in order to be determined as so-called significantly different. This means different enough so that we can safely say that there is a very small chance that a difference of this size can happen if the hypothesis is true (Keenan *et al*., 2012). A typical way of investigating this is to first set up an hypothesis, called the null hypothesis (H_o), in this equal to H_o: $\mu = \mu_o$, where the μ_o is the fixed reference value. Then the data are used to check if this hypothesis seems reasonable. A natural way to do this is to compute the empirical mean and compare this with the fixed value μ_o. If the difference is large, there is reason to believe that the hypothesis is not true. The problem is to decide how far away an estimated mean has to be from the fixed hypothesis value before one can conclude that the hypothesis must be rejected. The way to solve this problem in practice is to compare the difference between the measured and the hypothesis values with an estimate of the natural variability of the estimator. It can be shown that if the underlying distribution for *y* is the normal one and the hypothesis H_o is true (MacFarland, 2012).

Statistical process control

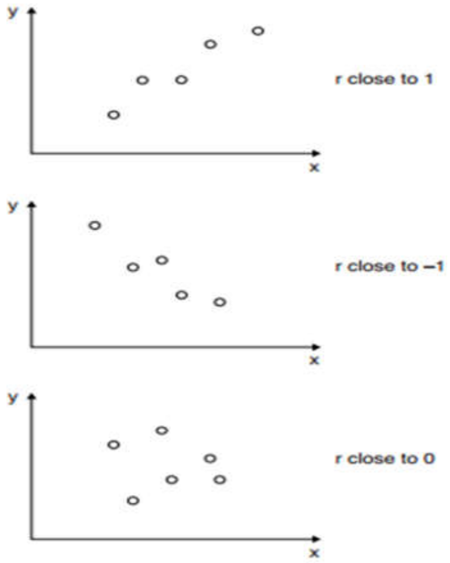
Montgomery (1997) explained briefly the statistical process control as a methodology applied for checking whether an actual measurement is within the normal range of variability. The SPC methodology is based on so-called control charts. These are plots of the data over time with so-called control limits superimposed. The plotting aspect is the most important, but the limits can also play a certain role for detecting drift and outliers. The assumption behind the most basic SPC control chart, the so-called Shewart control chart, is that when the process in under control (within the normal variation range) all observations are independent identically distributed. Usually, they are assumed to have a normal distribution.

Relationships between two or more variables

In the same way as for one variable, one can define a joint distribution of two or more variables. This distribution can as for the univariate case, be estimated using a generalization of the univariate histogram (Kleinbaum *et al*., 2007). The simplest way of studying the empirical relationship between two variables is to use a scatter plots between two and two variables. Figure 3 gives some examples of this. An important measure of the degree of linear relationship between *x* and *y* is the correlation. As for the mean and standard deviation, there will be one correlation for the underlying population and one for the sample. As usual the estimate will approach the true value when the *N* increases. The estimated correlation coefficient is, however, in this case biased in the sense that its expectation is larger than the true value of the correlation and can therefore, lead to overoptimistic assessments of the true correlation coefficient (MacFarland, 2012).

***Corresponding author**

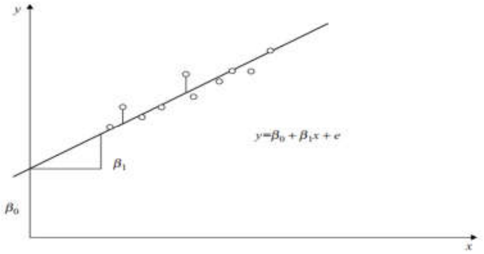
Fig. 3. Correlation. The three plots indicate three very different situations. The 1st and 2nd, correlation is strong while in the last it is very weak. In the 1st figure the relation is positive and results in a positive correlation coefficient. In the 2nd the relation is negative, which gives a correlation to -1. The correlation is always -1 and 1 (Adapted from MacFarland, 2012).



Simple linear regression

Assume again that one measures two variables *x* and *y* and that there is a relationship between them. Assume further that *x* is simple to measure while *y* is more difficult or time consuming to measure, but also more interesting. One will then be interested in building a model between the two such that *y* can be predicted from measurement of *x* (Martens and Naes, 1989). The simplest model that can be used for the relationship between *x* and *y* is the linear model (Fig. 4). This is a model assumption which may be reasonable or not, but often it is a useful approximation in real applications (Kleinbaum *et al*., 2007).

Fig. 4. Linear regression illustration. The y is a linear function of x plus noise. The points indicate the observations. The problem is to fit the best possible line. This is obtained by computing the distances from each point to a potential line (adapted from Kleinbaum *et al*., 2007).



Conclusion

Food is one of the basic needs of human beings. Quality and safety conditions are the first and the most that has to be confirmed before dispatching food and food products to the market. The important tool that can certify situations of food from the farm to the fork is statistics via gathering, evaluating, analyzing data, interpreting results and comparing with standards to decide its acceptability. Therefore statistical analysis has great role in food technology and at large to the health of people throughout the world. Health aspect of the world population would not be secured in the absence of statistical application in various fields, particularly in food technology.

Acknowledgements

Authors are thankful to the Jiangnan University, China and Massawa College of Marine Science and Technology-Eritrea for their assistance and guidance. Authors also express their thanks to anonymous reviewers of the review article for their valuable comments.

References

1. Alezandro, M., Granato, D., Lajolo, F. and Genovese, M. 2011. Nutritional aspects of second generation soy foods. *J. Agricult. Food Chem.* 59: 5490–5497.
2. Blumberg, B., Cooper, D. and Schindler, P. 2005. Business Research Methods. McGraw-Hill Education, Maidenhead, UK, pp.18-25.
3. Collis, J. and Hussey, R. 2003. Business Research. Palgrave MacMillan, Basingstoke, UK, pp.46–79.
4. Ellendersen, L., Granato, D., Guergoletto, K. and Wosiacki, G. 2012. Development and sensory profile of a probiotic beverage from apple fermented with Lactobacillus case. *Engg. Life Sci.* 12(4): 475-485.
5. Ellison, S., Barwick, V. and Farrant, T. 2009. Practical statistics for the analytical scientist, A bench guide (2nd Edition). Royal society of chemistry (RSC) Publishing. United Kingdom, pp. 48-60.
6. Fisher, R. 1966. The Design of Experiments (8th ed). Hafner Publishing Company, INC. New York,N.Y., pp. 11-40.
7. Freund, J. 2001. Modern elementary statistics (4th Edition). Prentice-Hall. London, pp. 621-730
8. Gacula, M. and Singh, J. 1984. Statistical Methods in Food and Consumer Research (2nd Edition). Food Science and Technology International series. Academic Press, Orlando, Elsevier Inc, USA, pp.128-194.
9. Granato, D., Calado, V., Oliveira, C. and 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 review in food science and nutrition. http://ddx.doi.org/10.1080/10408369.2017.750233
10. Granato, D., Ribeiro, J. and Masson, M. 2012. Sensory acceptability and physical stability assessment of a prebiotic soy-based dessert developed with passion fruit juice. *Sci. Technol. Food.* 32(1): 119–125.

*Corresponding author

11. Greenfield, H. and Southgate, D. 2003. Food Composition Data (2nd Edition). Food and Agriculture Organization of the United Nations, Rome, pp.63–82.
12. Hanafi, M. and Akiers, H. 2006. Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Comput. Stat. Data Anal.* 51: 1491–1508.
13. Iwaoka, W., Britten, P. and Dong, F. 1996. The changing face of food science education. *Trend. Food Sci. Technol.* 7: 105–112.
14. Johnson, R. and Bhattacharyya, G. 1992. Statistics: Principles and Methods, 2nd Edition. Wiley and Sons. New York, pp. 67-160.
15. Keenan, D., Brunton, N., Mitchell, M., Gormley, R. and Butler, F. 2012. Flavour profiling of fresh and processed fruit smoothies by instrumental and sensory analysis. *Food Res. Int.* 45: 17–25.
16. Kirk, R. and Sawyer, R. 1999. Pearson's Composition and Chemical Analysis of Foods (9th Edition). Longman Scientific and Technical, Harlow, UK, pp. 85-120.
17. Kleinbaum, D., Kupper, L. and Azhar, N. 2007. Applied regression analysis and multivariate methods (4th Edition.) An Imprint of Brooks/Cole Publishing Company/An International Thomson; Duxbury, Belmont, USA, pp. 14-212.
18. Kravchuk, O., Elliott, A. and Bhandari, B. 2005. A laboratory experiment, based on the maillard reaction, conducted as a project in introductory statistics. *J. Food Sci. Educat.* 4: 70–75.
19. MacFarland, T. 2012. Two-way analysis of variance: Statistical tests and graphics using R. Springer verlag. New York, pp. 50-150.
20. Mann, H.B. and Whitney, D. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Mathemat. Stat.* 18: 50–60.
21. Martens, H. and Naes, T. 1989. Multivariate Calibration. Chichester: John Wiley & Sons, Ltd. UK, pp.341-349
22. Montgomery, D. 1997. Introduction to Statistical Quality Control (6rd Edition). John Wiley & Sons, Inc. New York, 3-42
23. Razali, N. and Wah, Y. 2011. Power comparisons of Shapiro–Wilk, Kolmogorov– Smirnov, Lilliefors and Anderson–Darling tests. *J. Stat. Modeling Analyt.* 2(1): 21–33.
24. Shahbaba, R. 2012. Biostatistics with R: an introduction to statistics through biological data. Springer, New York, pp. 1-12.
25. Tressou, J., Leblanc, J., Feinberg, M. and Bertail, P. 2004. Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: Application to ochratoxin A. *Reg. Toxicol. Pharmacol.* 40: 252–263.
26. Weiss, N. 1999. Introductory Statistics (5th Edition). Addison Wesley Longman. New York, pp. 270-287.

*Corresponding author