NEW ERA AGRICULTURE MAGAZINE

Insights into the Use of Statistical Techniques in Food Science and Technology

Sandeep Kumar[1], Ashly Antony[2], Surnida Akash[3], Fathima Rimsa K[4] and Vimal Jahjhda[5]

Abstract: -

Statistical methods are crucial tools for identifying trends, analyzing relationships, and drawing conclusions from experimental data. However, it is common for researchers to use statistical tests without verifying their suitability for specific applications. This paper aims to introduce some of the key univariate and bivariate parametric and non-parametric statistical techniques, focusing on their practical applications in Food Science and Technology. It also discusses the prerequisites for using particular statistical tests, including their strengths and weaknesses in practical scenarios. For example, it emphasizes the importance of checking for normality and homogeneity of variances before conducting inference tests, as well as correlation and regression analyses, when comparing two or more sample sets.

Introduction:

Statistics, as a branch of mathematics, is applied extensively in the analysis of data. In Food Science, statistical methods are vital in planning, analyzing, and interpreting experimental data. This includes evaluating various aspects of food and beverages, such as their chemical, physical (e.g., rheological), sensory, and microbiological properties during product development and manufacturing. For example, statistical analysis might be used to assess the impact of increasing concentrations of a fruit extract on the acidity and sensory appeal of a product or to evaluate changes in biochemical markers like inflammation and oxidative stress in experimental animals treated with different doses of a food ingredient. The application of statistical tools is critical for research and development in both

Sandeep Kumar[1], Ashly Antony[2], Surnida Akash[3], Fathima Rimsa K[4] and Vimal Jahjhda[5]

[1]Ph.D. Research Scholar, Centre of Food Science and Technology, CCS, HAU, Hisar

[2]B.Sc. Scholar, College of Agriculture, Padannakad, Kerala Agricultural University

[3]Master's Scholar, Department of Plantation spices medicinal and aromatic crops, Sri konda laxman telangana state horticultural university

[4]B.Sc. Scholar, College of Agriculture, Padannakad, Kerala Agricultural University

[5]Master's Scholar, Department of Agricultural Economics, Institute of agriculture & natural sciences, Deen Dayal Upadhyay Gorakhpur University, Gorakhpur (U.P.)

academic and industrial settings, particularly within the food, chemical, and biotechnological industries. However, many researchers often select incorrect statistical tests or apply the correct tests in inappropriate situations. This issue can stem from several factors, such as a lack of interest in manual calculations, misinterpretation of statistical outcomes, or misuse of statistical software. With the advancements in computing power, statistical analysis has become more accessible through various software packages, which can quickly generate linear and non-linear models, create graphs, and solve complex algorithms—tasks that previously required significant time when done manually. Proper use of statistics in food research is essential for understanding data, accounting for variability in measurements and process controls, and summarizing experimental results. The objectives of this paper are threefold: **(1)** to explain key concepts related to data analysis in Food Science and Technology, **(2)** to provide relevant statistical information, and **(3)** to present and discuss published examples of mathematical modeling in food research. This includes exploring how statistical methods can be employed to optimize processes and study relationships between variables, helping to interpret and effectively present research findings.

Concepts of statistics applied in Food Science

Using appropriate statistical tools is crucial because it allows researchers to extract maximum information from experimental data. When research is published in a journal, it is important to provide sufficient detail so that readers can fully grasp the study's objectives and outcomes, and if necessary, replicate the work. However, many published studies often lack adequate information about the statistical methods used for data interpretation. Typically, the analysis is limited to basic descriptive statistics such as mean, median, minimum and maximum values, standard deviation, or coefficient of variation. In many cases, more advanced statistical techniques like correlation, regression, and comparison of mean values are applied using automated statistical software. However, this reliance on statistical packages may lead to inappropriate use if the researcher does not understand the underlying principles of these tests. Therefore, it is essential for researchers to have a solid understanding of inferential statistical tests before applying them. This knowledge is necessary for planning experiments effectively, analyzing data accurately, and interpreting results within the context of a comprehensive data framework, ultimately enabling them to draw valid conclusions from their work.

Regardless of the type of experimental design employed, it is vital to assess the statistical quality of the data before proceeding with further analysis. Poor-quality data can lead to incorrect conclusions. Data may be compromised if, for example, there is an inadequate sample size, non-random sampling, high measurement uncertainty, a lack of proper training for the analyst, or the presence of "censored" values. These issues should be addressed when planning an experiment, as they are usually within the researcher's control. Sometimes, experimental results may fall outside the detection limits of the analytical methods used. For example, the concentration of an analyte might be below the method's lowest detectable limit (left-censored) or, in rare cases, exceed the highest limit of quantification (right-censored). Handling such censored data is a widely debated topic across fields like food chemistry, microbiology, and toxicology. There are still ongoing discussions about the best methods to address these issues and ensure the reliability of the data analysis.

Parametric statistics in Food Science

The choice between parametric and non-parametric tests depends on the data's statistical distribution, sample size, and homoscedasticity (equal variance across samples). Parametric tests are appropriate when the data follow a normal distribution and the variances are homogeneous, as verified by the Shapiro–Wilk test (for normality) and Levene's test (or F-test, for homogeneity of variances). In such cases, the Student's t-test is applied to compare the means of two groups, while ANOVA (Analysis of Variance) is used to compare the means of three or more groups.

Non-parametric statistics in Food Science

Non-parametric tests analyze ranked data instead of raw data values. The data are ordered from smallest to largest, assigning ranks from 1 to n (with n being the total number of samples). These methods are particularly useful when there is no universally accepted scale for the original data or when parametric methods might be criticized for relying on an arbitrary metric. The key advantage of non-parametric tests is that they do not assume normality or homogeneity of variance, making them robust in cases where these assumptions are violated. They focus on comparing medians, which reduces the impact of outliers. However, a significant drawback is that non-parametric tests lack defined parameters, making it harder to quantify differences between populations. By relying on ranks, they preserve the data order but ignore actual values, leading to a potential loss of information. Consequently, non-parametric methods are generally less powerful (i.e., have a lower ability to detect differences) compared to parametric tests. Non-parametric tests are commonly used for nominal, categorical, or

ordinal data, or when dealing with arbitrary scales without numerical interpretations, such as in sensory evaluations of preferences. Each non-parametric test has specific sensitivities and limitations. It is often beneficial to use multiple non-parametric tests, and if results differ, it is important to investigate the reasons for the discrepancies.

Bivariate correlation analysis

Correlation analysis examines the potential relationship between two continuous variables. The correlation coefficient (r) quantifies this association, ranging from -1 to +1. A value closer to ±1 indicates a strong linear relationship, while a value near 0 suggests a weak or no linear association. The correlation is positive when both variables increase together, and negative if one variable decreases as the other increases. However, a low r-value does not always imply no relationship, as non-linear correlations or outliers can influence this measure. The Pearson correlation coefficient is commonly used to assess the strength of linear relationships between two normally distributed data sets. However, when dealing with more than five variables, interpreting the results becomes challenging. For instance, if analyzing five variables (A, B, C, D, E), calculating pairwise correlations (e.g., AB, AC, AD) may reveal associations, but understanding the broader data structure requires advanced techniques like principal component analysis (PCA), clustering, or linear discriminant analysis (LDA). If analyzing large datasets (≥30), it is important to check for normality. For non-normally distributed data or ordinal variables, the Spearman's rank correlation coefficient (ρ) is more suitable. Spearman's method assesses correlation without assuming linearity or normal distribution, making it ideal for small sample sizes or ranked data. Thus, Spearman's coefficient is often preferred when normality assumptions are violated or when dealing with ordinal measurements. This careful selection of statistical tests based on data characteristics ensures robust and accurate analysis, preventing misleading interpretations.

Regression analysis

Regression analysis is a statistical method used to evaluate the relationship between dependent and independent variables. It encompasses various techniques for modeling and analyzing multiple datasets. Key types of regression include:

- Linear regression
- Multiple regression (using several predictors),
- Probit regression
- Logistic regression, which is particularly useful for examining biological responses to different doses of stimulatory or inhibitory treatments.

At its simplest, regression can be applied to assess how variables such as increased treatment time or temperature affect the heat resistance of microorganisms, helping to determine thermal effects (D-values) under specific conditions. Additionally, it is often used to develop calibration curves for chemical, physical, and biological assays, involving continuous datasets. For accurate calibration, at least five concentration levels, including a blank, are required, with triplicate testing for each level.

Assumptions of Regression Analysis

For valid regression analysis, the following assumptions must be met:

1. **Representativeness:** Samples must accurately reflect the population for making inferences.
2. **Measurement Accuracy:** Independent variables should have minimal measurement error. If measurement errors are present, orthogonal linear regression should be used for correction.
3. **Random Error:** The error term should be a random variable with an average of zero.
4. **Linear Independence:** Predictors must be linearly independent, meaning no predictor can be expressed as a linear combination of others.
5. **Homoscedasticity:** The variance of errors should be constant across all levels of the independent variables.

Steps for Conducting Regression Analysis

Before carrying out regression analysis:

1. **Check for Outliers:** Use Grubbs' test at a 95% or 99% confidence level to detect any outliers in each concentration level.
2. **Verify Variance Homogeneity:** Ensure that variances are consistent across different levels of the calibration curve using appropriate statistical tests like Levene's or Bartlett's tests.

Other statistical techniques

As discussed earlier, the most commonly used univariate and bivariate statistical techniques have been outlined. However, there are other statistical and mathematical methods, particularly in chemometrics, that include techniques such as Principal Component Analysis, Cluster Analysis, Discriminant Analysis, K-nearest Neighbors, and more advanced methods like neural networks. These techniques can be seen as extensions of basic methods, designed for specific analytical needs. Examples of how these approaches are applied to complex data include sensory analysis physicochemical analysis, microbiological studies, metabolomics, and chemical data analysis.

References

1. 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.
2. Barnet, & Lewis (1978). Outliers in statistical data. Chichester, UK: Wiley.
3. 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.
4. Bergstrand,M., & Karlsson, M.O. (2009). Handling data belowthe limit of quantification in mixed effect models. The AAPS Journal, 11(2), 371–380.
5. 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.
6. 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.
7. 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.
8. 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.