How to perform properly statistical analysis on food data? An e-learning tool: Advanced Statistics in Natural Sciences and Technologies

T. Eftimov[1,2], **P. Korošec**[1,3], **D. Potočnik**[2,4], **N. Ogrinc**[2,4], **D. Heath**[4] and **B. Koroušić Seljak**[1,2]

[1] Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000, Ljubljana, Slovenia
[3] Faculty of Mathematics, Natural Science and Information Technologies, Glagoljaška ulica 8, 6000 Koper, Slovenia
[4] Department of Environmental Sciences, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

Many scientists have problems and difficulties in making a statistical analysis of their data, which they need to interpret their results. One problem is that applying a statistical method requires knowledge of the conditions (assumptions) about the data that must be met in order to apply it. These initial conditions are not usually checked and researchers simply apply a statistical method, in most cases taken from a similar published work, which is unsuited to their data set. As a result, their conclusions can be incorrect. This kind of misunderstanding is all too common in the research community. To become familiar with statistical methods, we provide a short tutorial on how to perform statistical analysis of food data. The study uses authentic food data on fatty acid profiles and the isotope composition of milk samples. In addition, we present an in-house developed and freely available e-learning tool for advanced statistics in natural sciences and technologies that has the benefit of checking the required conditions of each statistical method and offering only those methods that are appropriate for analysing the experimental data.

Keywords: statistical analysis; e-learning tool; food data analysis

1. Introduction

Many researchers have problems and difficulties performing a statistical analysis of their data, which is often crucial in interpreting their results. The problems appear because each statistical method has some conditions (assumptions) about the data that must be satisfied in order to apply the test [1]. Similarly, in food research, as in other research domains [2], researchers do not check for these conditions and simply apply a statistical method based on one used in a similar study. As a result, their conclusions can be incorrect. In addition, authors often do not provide the necessary information on the required conditions for selecting an appropriate statistical method. For example, to compare data from different data sets, two commonly used statistical tests are the paired t-test (if a comparison is made between two data sets) [3] and ANOVA (if a comparison is made between three or more data sets) [4]. The problem is that these tests are used even when the conditions for their safe use are not satisfied i.e., checking for normality, homoscedasticity (equality of variances) and independence. If the required conditions are not satisfied, an analyst will need to apply an alternative version of these tests. In many cases, calculating the $(1-\alpha)100\%$ confidence intervals [5] usually assumes that the data is normally distributed, when, in reality, the experimental data does not follow a normal distribution.

2. Statistical tutorial

To become familiar with statistical methods, we provide a short tutorial on how to perform a statistical analysis of food data. The tutorial is focused on descriptive statistics, hypothesis testing, and confidence intervals.

2.1 Population, representative data samples, and data types

A *population* is a set of similar items or events, which are of interest for some question or experiment, while a *representative data sample* is a set of data collected and/or selected from a statistical population by a defined procedure [1]. To see the difference between them, let us consider that a researcher is interested in how many people drink milk with breakfast in Slovenia. In this example, the population contains each person who lives in Slovenia. It would be unrealistic for us to ask each individual about his or her milk drinking habits, so instead we need a representative sample of people. The information obtained from the representative data sample allows a researcher to develop hypotheses about the larger population. In this example, let us consider that the researcher is a vegan and randomly selects a sample in which many of the participants are his or her friends many of whom are vegans. In this case, the number of people who drink milk with their breakfast will be lower when in reality the number is larger. This is an example of sample selection bias. In order to have a representative data sample, the researcher needs to be unbiased in their selection i.e., there are no outside factors influencing sample selection.
Once a representative data sample is selected, it becomes important to know what type of data has been collected [1]. There are two types of data: *qualitative* (categorical) and *quantitative* (numeric). Each type is split into two sub-types:

ordinal and *nominal* for qualitative data, and *discrete* and *continuous* for quantitative data. Ordinal data is categorical, where the values have natural, ordered categories and the distances between the categories are not known. For example, a question from a food questionnaire might ask: "How often do you eat fish during the week?". Possible answers include "Never", "one to three times per week", "four to six times per week", or "Every day". In this case, the answers are examples of ordinal data. Nominal data is also a categorical data, but in this case there is no natural order between the categories. Examples include eye colour, gender, and the region where one lives. In contrast to qualitative data, quantitative data is numerical. Discrete data can only take on a finite number of numeric data, while continuous data can take on an infinite number of possible values. Let us assume that a researcher is interested in the weight of participants in the data sample. The weight of each participant is stored as a continuous variable. However, he/she can split the participants into four groups according to their weight: (1) less than 50 kg, (2) 50 kg - 65 kg, (3) 66 kg - 80 kg, and (4) greater than 80 kg. In this case, weight is a discrete variable because it can take only one value from the four weight groups. This is an example of discretization, where a continuous variable is transformed into a discrete variable.

2.2 Descriptive statistics and inferential statistics

The next step is to apply statistics to the representative data sample. There are two different statistical analyses that we can apply: *descriptive statistics* and *inferential statistics* [6]. Descriptive statistics summarizes data from a sample using measures, while inferential statistics draws conclusions from the data set subject to random variation.

Descriptive statistics includes a distribution of single variable, measures of *central tendency*, which include the *mean*, *median*, and *mode*, and measures of *variability*, which include *standard deviation*, *variance*, and the *minimum* and *maximum* value. The distribution is a summary of the frequency of individual values or range of values for a variable. It represents every value of the variable and the number of how many times the value appears in the data sample. The central tendency of a distribution is an estimate of the "centre" of a distribution. The mean (average) is the most commonly used measure for central tendency and in order to compute it we sum all the values in our data sample and then divide the sum by the sample size (number of values in the data sample). The median is the value that separates the higher half of a data sample from the lower half, or it can be assumed as the "middle" value of an ordered data sample. The mode is the value from a data sample that appears with the highest frequency. The variability of a distribution refers to the spread of the data values around the central tendency. The minimum and the maximum value are the minimum and the maximum value that appear in the data sample. The standard deviation is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data values are close to the mean, while a high standard deviation indicates that the data values are spread out over a wider range of values. The variance is the square of the standard deviation.

Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected perfectly to represent the population. The methods of inferential statistics include estimation of distribution parameters and the testing of statistical hypotheses.

When we talk about statistical analysis, it is also important to distinguish between *univariate*, *bivariate* and *multivariate* statistical analysis. Univariate analysis works when the data sample has only one variable. It does not deal with relationships and the major purpose is to describe the data. Bivariate analysis involves analysing two variables in order to find a relationship between them. Bivariate analysis is a special case of multivariate analysis when multiple relations between multiple variables are analysed. To explain better the difference between them, let us assume that the researcher is interested in testing whether or not consulting with a nutritionist can decrease a person's calorie intake. For this purpose, a researcher works with participants who consult with a nutritionist each day after work. The participants form two groups: one group that has daily consultations and a control group who do not receive any nutritional advice. After a week, let us assume that the researcher discovered that those who consult with a nutritionist have decreased their calorie intake for 40% over the control group. This is an example of univariate inferential statistics because it analysed the relationship between one independent variable (consulting with a nutritionist) with a single dependent variable (calorie intake). If a researcher is interested in the relationship between the times a person consulted a nutritionist and calorie intake over the week, then he/she needs to perform a bivariate analysis. If the researcher is interested to adding another variable i.e., how many times a person visited a gym, to determine the effectiveness of visiting a nutritionist and exercise, on two groups of people (20-40 yrs., and 41-60 yrs.), then he/she needs to perform multivariate analysis to find the relationships between all of the variables.

2.3 Hypothesis testing

One of the most commonly used approaches for testing relationships between two or more data samples is to use *hypothesis testing* [7]. Hypothesis testing is a procedure in which sample(s) data is employed to evaluate a hypothesis. The procedure involves two hypotheses: a *null hypothesis* (H_0) and an *alternative hypothesis* (H_A). The null hypothesis is a statement of "no effect" or "no difference", while the alternative hypothesis is a statement indicating the presence of an effect or a difference. We can define hypotheses about a single population or about the relationship between two or more populations e.g., between either the means or variances of a single, two or multiple distributions or even about the whole distribution.

After the data is collected, the next step is to apply hypothesis testing in order to evaluate the data using an appropriate *statistical test*. Each statistical test has a test statistic that has some distribution. A *test statistic* is a mathematical formula used to obtain a value using the collected data sample(s). Then the obtained value is compared with a value from a special table that contains information about the distribution of the test statistic. Such tables contain extreme values of the test statistic that are highly unlikely to occur if the null hypothesis is true. In order to obtain a value from such a table, a researcher first needs to specify a *level of significance* (α). The significance level is a probability threshold below which the null hypothesis will be rejected. Instead of specifying the significance level, a *p-value* can be used, which is the smallest level of significance that results in the rejection of the null hypothesis. The smaller p-value indicates that the null hypothesis is rejected, while a larger p-value indicates that the null hypothesis is not rejected.

In reality, the null hypothesis can be either true or false, and the result of a statistical test can be that the null hypothesis is either rejected or is not rejected. When performing hypothesis testing, two types of errors can occur: *type* I and *type* II. A *type I* error is the probability of rejecting the null hypothesis when it is actually true, while a *type II* error is the probability of not rejecting the null hypothesis when it is actually false. The probability of a type I error is the level of significance (α), so before the study we usually assign it a small value (e.g., 0.05, 0.01) because researchers do not want to have a type I error. The probability of type II error is marked as β and it is related to the *power of a statistical test*. The power is the probability that it will reject a false null hypothesis, or power = 1- β.

Power analysis is an important aspect in experimental design [8]. It allows researchers to determine the sample size required to detect an effect of a given size with a given degree of confidence. It allows researchers to find the probability of detecting an effect of a given size with a given significance level, under sample size constraints. If the probability is very low then researchers can change the sample size of the experiment.

The next step is to select an appropriate statistical test. There exist two types of tests: *parametric* and *nonparametric*. In order to distinguish between, we must check the conditions for the safe use of a parametric. These conditions include checking for *independence*, *normality of the data*, and *homoscedasticity of the variances*. Statistically, two events are independent if the occurrence of one does not influence the probability of the other occurring. Normality indicates that the data is normally distributed, which we can check by using statistical tests such as the Kolmogorov-Smirnov [9], Anderson-Darling [10], Shapiro-Wilk [11], and D'Agostino-Pearson test [12]. The result from a statistical test can be graphically proven using histograms or Q-Q plots (quantile-quantile). In probability, quantiles are cut points dividing the range of a probability distribution into intervals with equal probabilities. The homoscedasticity of variances indicates the hypothesis of equality of variances (homogeneity of variances). The Levene's test [13] can be used to check the homoscedasticity of the variances. If the required conditions for the safe use of parametric tests are satisfied, we must use a parametric test because it has higher power then a nonparametric test, otherwise we must select a nonparametric test.

Additionally to conditions for the safe use of parametric tests, other parameters that are also related to the selection of an appropriate statistical test is the number of data samples that need to be compared (two or more then two) and if the data samples are *paired* or *unpaired*. Paired samples (also called dependent samples) [1] are samples in which natural or matched couplings occur. So in the data sample each data value in one sample is uniquely paired to a data value in the second sample. Examples of paired samples are found in food questionnaires. Let us suppose that researchers are interested if there is a difference between two populations (males and females) according to the food questionnaire. In this case the obtained result of the male population for the first question is paired with the obtained result of the female population for the same question, and so on. The choice between paired and unpaired samples depends on experimental design, and researchers need to be aware of this when designing their experiment.

The first step for performing hypothesis testing is to check all the conditions involving the data sample(s), and then to decide which statistical test is the most appropriate. Table 1 presents the different statistical test classified according to the conditions that must be met.

If a researcher wishes to compare more than two data samples, first he/she needs to check the conditions for independence, normality of data, and homoscedasticity of variances. If they are satisfied, then he/she needs to determine if the samples are either paired or unpaired. If he/she has unpaired samples, an appropriate statistical test to use is the one-way ANOVA. If at least one of the conditions for the safe use of a parametric test is not satisfied, then he/she needs to select a nonparametric alternative such as the Kruskal-Wallis test.

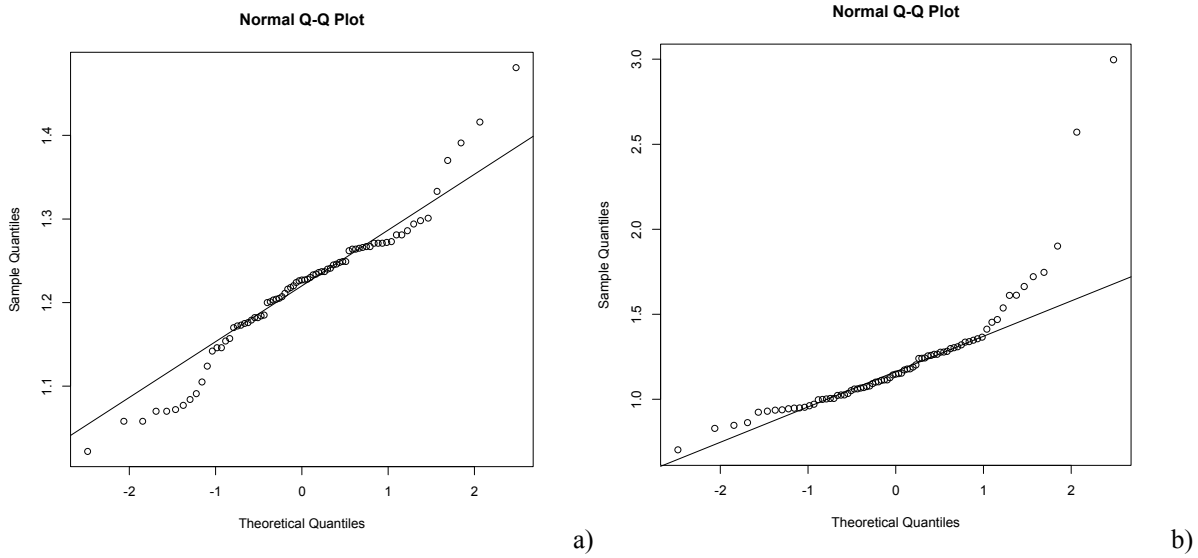
Table 1 Classification of different statistical test.

	Two data samples	More then two data samples
Parametric	t-test (unpaired) [14] Paired t-test (paired) [3]	One-way ANOVA (unpaired) [4] Repeated-measures ANOVA (paired) [15]
Nonparametric	Mann-Whitney U (unpaired) [16] Wilcoxon signed-rank (paired) [18]	Kruskal-Wallis (unpaired) [17] Friedman, Friedman-aligned, Iman-Davenport (paired) [19]

2.3.1 Example of comparing two data samples

This example involves comparing two data samples. The data is authentic food data and represents the fatty acid profile of Slovenian milk samples. Fatty acids composition in milk can be significantly altered by the nutrition of the dairy cows and their metabolisms and thus can provide information about the provenance of milk and dairy products. One batch of samples contains the fatty acid profile of 77 milk samples collected in 2013, while the other fatty acid profile of 77 samples collected in 2014. Comparing both data samples, we are interested to test if there is difference between the fatty acid profiles of milk sampled in 2013 and 2014.

First, we started by checking the required conditions for the safe use of the parametric tests. The independence is satisfied because the milk samples are independent. The normality condition is checked using the Shapiro-Wilk test, for which the significance level is set to 0.05. The p-value using this test with the data from 2013 is 0.02, while the p-value obtained with the data from 2014 is 0.00. In both cases the p-values are smaller then the significance level, which indicates that the data from both samples is not normally distributed. Figure 1 shows a graphical representation of the data from both data samples using normal quantile-quantile (Q-Q) plots. If the data is normally distributed, the values in the plot will lie on a straight diagonal line, which is not the case for our both samples.



a) b)

Fig. 1 Normal Q-Q (quantile-quantile) plots for the data samples from a) 2013 and b) 2014.

Using the result form the Shapiro-Wilk test and the normal Q-Q plots, we see that normality condition is not satisfied, so we do not need to check the homoscedasticity of variances, because we can not select a parametric test, and instead we need to select a nonparametric test. However, using the Levene's test, the p-value is 0.00, which is smaller than 0.05, and there is no homogeneity of variances. Because there are two data samples that are unpaired, an appropriate test is the Mann-Whitney U test. Using it, the p-value is 0.04, which is smaller than 0.05, the significance level set prior to analysis, and the null hypothesis is rejected, so there is a significant statistical difference between the fatty acid profiles measured between two different years.

2.3.2 Example of comparing three data samples

This example involves comparing three data samples. The data presented is authentic food data on the fatty acid profile (different from the fatty acid profile in the first example) in Slovenian milk samples. The first sample contains the fatty acid profile measured in 77 milk samples from 2012, the second contains the same fatty acid profile measured in 77 samples form 2013, while the third sample contains the same fatty acid profile measured in 77 samples from 2014. Comparing the three data samples, we are interested to test if there is a difference between the fatty acid profile of milk collected in 2012, 2013 and 2014.

As in the case of comparing two data samples, we started by checking the required conditions for the safe use of parametric tests. In this case independence is satisfied because the milk samples are independent. To check for

normality we use the Shapiro-Wilk test, for which the significance level is set to 0.05. The p-value obtained for the sample form 2012 is 0.00, the p-value for the sample from 2013 is 0.00, and the p-value for the sample from 2014 is 0.00. All p-values are smaller than 0.05, so the normality condition is not satisfied. Since, therefore, we cannot use a parametric test we must select a nonparametric one. The p-value for homoscedasticity of variances is 0.03 and is smaller than our selected significance level 0.05, which means that we reject the null hypothesis and there is no homogeneity of variances. Because there are three data samples that are unpaired, an appropriate statistical test is the Kruskal-Wallis test. Using it, the p-value is 0.87, which is greater than 0.05 and we do not reject the null hypothesis, so there is no significant statistical difference between the fatty acid profiles of milk collected from the three different years. If there is a significant statistical difference and we are interested to see from which pairs of data samples this difference appears, we need to continue with a post-hoc statistical test appropriate for the Kruskal-Wallis test.

2.4 Confidence interval

In some experiments, it can happen that the distribution of the data collected in a data sample is known, but the parameters that describe the distribution are unknown and must be estimated from the data sample. For example, let us assume that a researcher has data that is normally distributed, but he/she does not know the mean and the standard deviation of the normal distribution. One way to estimate them is to use the (1-α)100% *confidence interval* (CI) [5]. A CI is a type of interval estimate of an unknown parameter. It is calculated using the collected experimental data and it potentially includes the unobservable parameter of interest. The parameter that indicates how frequently the interval contains the true parameter is the confidence level (α). For example, a 95% CI indicates there is 95% probability that the calculated CI contains the true value of the population parameter, but it does not mean that for a given interval there is a 95% probability that the population parameter lies within the interval.

The most common approach for calculating the CI is to use the sigma, two-sigma, three-sigma rule of thumb (68-95-99.7 rule) [20]. However, we can only apply this rule when the data is normally distributed. So before using it, we must first check to see if the data is normally distributed. In reality, many papers include CIs calculated using this rule even when the normality condition is not satisfied resulting in incorrect results. So the question that arises here is how to calculate the CI when the normality condition is not satisfied. The answer is to use bootstrap confidence intervals. Bootstrapping involves random sampling with replacement, which means randomly selecting a subset of values from the data sample in which a value may appear multiple times in the selected subset [21]. We use it to estimate parameters by measuring the properties when sampling form an approximating distribution, which in practice is the empirical distribution function of the observed data. In the case where we can assume that a set of values are independent and identically distributed, bootstrapping can be performed using a number of resamples with replacement of the observed data sample, which have equal size to the observed data sample.

2.4.1 Example of calculating confidence interval

This example involves finding a 95% CI for the parameters: means and variance, using a data sample that consists of the fatty acid profiles measured in milk samples (77) from 2014. First, we check for normality using the Shapiro-Wilk test at a significance level 0.05. The p-value is 0.00, so it indicates that the data is not normally distributed, meaning we cannot use the two-sigma rule and instead use bootstrapping. Using bootstrapping the CI for the mean is [1.22; 1.41] and the CI for the variance is [0.10; 0.47]. The calculated CI for the mean parameter indicates that there is 95% probability that the calculated CI contains the true value of the mean, while the calculated CI for the variance indicates that there is 95% probability that the calculated CI contains the true variance.

3. An e-learning tool: Advanced Statistics in Natural Sciences and Technologies

3.1 An in-house developed open-source e-learning tool

We developed an in-house freely available e-learning tool for advanced statistics in natural sciences and technologies, which covers the statistical tutorial outlined in this chapter. At present, it contains methods for linear regression, which is an approach for modelling the relationship between a scalar dependent variable and one or more independent variables [22] and principal component analysis (PCA), which is an approach to convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables [23]. The benefit of the e-learning tool is that it checks for the required conditions of each statistical method and offers only those methods that are appropriate for analysing the experimental data. The e-learning tool is available free to use on registration at http://ws.ijs.si/statTool/. The tool includes “Basic statistics”, “Hypothesis testing”, “Confidence interval”, “Regression analysis”, and “Dimensionality reduction”. There is also a template for organizing the data, which must be a comma-separated values file (csv). After uploading the data, the e-learning tool checks the data and will give an error message if either the decimal values are represented by a decimal point, the data contains any character values or if any values missing. Figure 2 shows the registration tab of the e-learning tool.

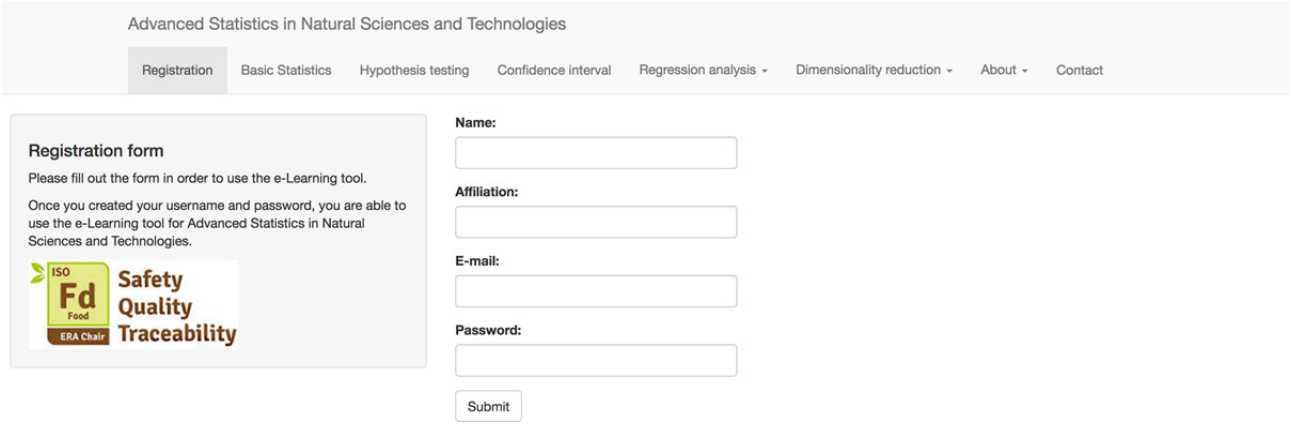


Fig. 2 The registration part of the e-learning tool.

The idea for such a tool came about after asking a sample of 28 researchers (master and doctoral level) who work in food analysis to select a method to do the following: i) compare two data samples, ii) compare three or more data samples, and iii) calculate a confidence interval. In the first case, 36.36% selected a paired-t test, while 45.46% did not know which statistical test to use. Only 18.18% chose the correct answer, which was to first check the conditions necessary for each method based on their data and then decide on the method to use. In the second case, 81.82% selected to use ANOVA, while again only 18.18% chose to check the conditions of each method before selecting which statistical method to use. Finally, when calculating the confidence interval just over half of the participants (54.54%) selected the standard formula that requires normally distributed data, even though their data is not normally distributed, i.e., they did not know which formula to use. In this case, none of the participants chose to check first the conditions of the data. It is clear that only a small percentage of the participants know how to perform correctly a statistical analysis.

A presentation on statistical analysis was presented to participants of the ISO-FOOD spring school on the use of isotopes in food organized from 4th to 7th April 2017 at Jožef Stefan Institute, Ljubljana, Slovenia [24]. After an introductory presentation, the participants (master and doctoral level researchers) where asked to complete the same three tasks as earlier. After dividing the participants into three groups they were given thirty minutes to perform one of the three examples using a statistical software package in which they were experienced. The data used for the tasks is available at (http://cs.ijs.si/opendata/DataSets.zip). Each group then presented their results. The group that worked on task one, began by using Excel, but switched to RStudio [25] since Excel was taking too much time. First, they checked the required conditions for the safe use of a parametric test and because the tests were not satisfied, they selected an appropriate nonparametric test, which in their case was the Mann-Whitney U test, whereas before the presentation, the majority of opinion was to use a paired-t test. The second group chose to compare the three data samples using three different pairwise comparisons by comparing the variances of the data samples using the Fisher test [26]. However, this is not correct because these pairwise comparisons are independent and combined independent p-values need to be additionally calculated to control the family wise error (FWER), which is the probability of making type I errors when performing multiple hypotheses tests [27]. The group also explained that they performed multiple pairwise comparisons because they could not find the ANOVA function in Excel. However, the appropriate statistical test in this example was Kruskal-Wallis test. The third group, which worked on calculating CI, also used Excel using the equations for normally distributed data despite this not being the case in the test data. Afterwards, they switched from Excel to SPSS [28], and performed the same experiment reporting the bootstrapping confidence intervals, which were the correct results for this experiment. Prior to the presentation, when asked, the suggestion was to use the sigma, two-sigma, three-sigma rule.

Each group was then set the task of reviewing three scientific papers published in a journal with high impact factor and then to judge if the statistical analysis used is correct and what if any information relating to statistics is missing. All groups agreed that in the three reviewed papers the authors only said that they used the t-test to compare two data samples and a one-way ANOVA to compare more then two data samples. They did not provide the results for checking the required conditions for the safe use of parametric tests, so it is not clear how they selected the statistical test or if they also followed a similar study. In addition, the CIs were reported using the sigma, two-sigma, three-sigma rule without first checking if the data was normally distributed or not.

Finally, after a short presentation of the e-learning tool each group repeated each of the three tasks. All three groups needed only a few minutes to perform the three examples. For the first example, they needed only to upload the test data, and to indicate if they have paired or unpaired samples. The e-learning tool checks all other conditions and offers them with the most suitable statistical test, in this case the Mann-Whitney U test. For the second example, the tool offered only the Kruskal-Wallis test (Fig. 3), which meant that they did not have to check for either normality or homoscedasticity of the variances. Finally for the third task, the participants needed to specify only which parameter they want to calculate the CI and the tool provides the CI together with the method used to obtain it.

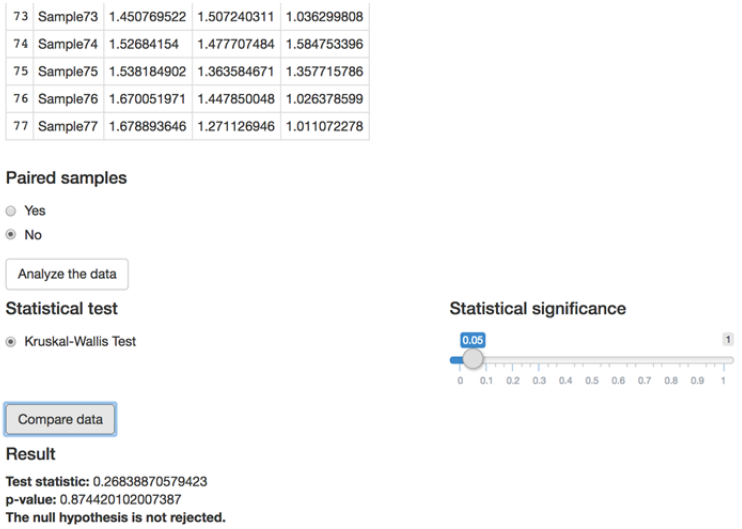


Fig. 3 The result from the e-learning tool when used to compare three data samples. The tool checks the required conditions for the safe use of parametric tests, which in this case are not satisfied, and offers the user the Kruskal-Wallis test, which is a nonparametric test.

4. Conclusion

In this chapter, we present a statistical tutorial of how to perform properly statistical analysis on food data. In addition, an e-learning tool is presented. The e-learning tool automatically checks the conditions for each method and offers only those methods that are appropriate for the data. Overall, the e-learning tool not only reduces the time needed to perform a statistical analysis but importantly, it can help in interpreting results and increase awareness of using an inappropriate statistical method. In the future, we plan to upgrade the e-learning tool with a power analysis to help researchers to select a relevant sample size for their experiments.

Acknowledgements This work is supported by the project ISO-FOOD, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 621329 (2014-2019). The authors acknowledge the financial support from the Slovenian Research Agency for the project "The content of trans fats in foods and population intakes - public health implications" (research core funding No. L3-7538).

References

[1] Ostle B, Mensing RW. Statistics in research: Basic concepts and techniques for research workers. Ames, IA: Iowa State University Press.1975.
[2] Eftimov T, Korošec P, Koroušić Seljak B. Disadvantages of Statistical Comparison of Stochastic Optimization Algorithms. Proceedings of the Bioinspired Optimizaiton Methods and their Applications, BIOMA. 2016.
[3] Zimmerman DW, Teacher's corner: A note on interpretation of the paired-samples t test. Journal of Educational and Behavioral Statistics. 1997; 22.3:349-360.
[4] Stoline MR. The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. The American Statistician. 1981; 35.3: 134-141.
[5] Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in medicine. 1998; 17.8: 857-872
[6] Lapin LL. Probability and statistics for modern engineering. Thomson Books/Cole. 1983.
[7] Winer BJ, Brown DR, Michels KM. Statistical principles in experimental design. New York: McGraw-Hill. 1971; 2.
[8] Erdfelder E, Faul F, Buchner A. GPOWER: A general power analysis program. Behavior research methods, instruments & computers. 1996; 28:1-11.
[9] Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American statistical Association. 1967; 62.318:399-402.
[10] Pettit AN. Testing the normality of several independent samples using the Anderson-Darling statistic. Applied Statistics. 1997. 156-161.
[11] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965; 52.3-4:591-611.
[12] Pearson ES, D'agostino RB, Bowman KO. Tests for departure from normality: Comparisons of powers. Biometrika. 1977;231-246.
[13] Schultz BB. Levene's test for relative variation. Systematic Biology. 1985; 34.4: 449-456.
[14] Cressie NAC, Whitford HJ. How to Use the Two Sample t-Test. Biomedical Journal. 1986; 28.2:131-148.
[15] Dien J, Santuzzi AM. 4 Application of Repeated Measures ANOVA to High-Density ERP. Event-related potentials: A methods handbook. 2004; 57-81.

[16] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger that the other. The annals of mathematical statistics. 1947; 50-60.
[17] Breslow N. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrica. 1970; 579-594.
[18] Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician. 1981; 35.3:124-129.
[19] Derrac J, Garcia S, Molina D, Herrera F. A practical tutorial on the use of nonparametric statistical tests as a methodology for computing evolutionary and swarm intelligence algorithms. Swarm and Evolutionary Computation. 2011; 1.1:3-11.
[20] Kriegel HP, Kroger P, Schubert E, Zimek A. LoOP: local outlier probabilities. Proceedings of the 18th ACM conference on Information and knowledge management. 2009.
[21] Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science. 1986; 54-75.
[22] Seber GAF, Lee AJ. Linear regression analysis. John Wiley & Sons. 2012; 936.
[23] Jolliffe I. Principal Component Analysis. John Wiley & Sons, Ltd. 2002.
[24] Era Chair for isotope techniques in food quality, safety and traceability – ISO-FOOD, http://isofood.eu/
[25] Team RStudio. RStudio: integrated development for R. R Studio, Inc., Boston, MA URL http://www.rstudio.com. 2015.
[26] Cochran WG. Approximate significance levels of the Behrens-Fisher test. Biometrics. 1964; 20.1:191-195.
[27] van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. Statistical applications in genetics and molecular biology. 2004; 3.1:1041.
[28] Green SB, Salkind NJ. Using SPSS for Windows and Macintosh: Analyzing and understanding data. Prentice Hall Press. 2010.