Pharmacological Reports (2020) 72:481–485
https://doi.org/10.1007/s43440-020-00110-5

EDITORIAL

Reporting statistical methods and outcome of statistical analyses in research articles

Mariusz Cichoń[1]

Published online: 15 June 2020

Introduction

Statistical methods constitute a powerful tool in modern life sciences. This tool is primarily used to disentangle whether the observed differences, relationships or congruencies are meaningful or may just occur by chance. Thus, statistical inference is an unavoidable part of scientific work. The knowledge of statistics is usually quite limited among researchers representing the field of life sciences, particularly when it comes to constraints imposed on the use of statistical tools and possible interpretations. A common mistake is that researchers take for granted the ability to perform a valid statistical analysis. However, at the stage of data analysis, it may turn out that the gathered data cannot be analysed with any known statistical tools or that there are critical flaws in the interpretation of the results due to violations of basic assumptions of statistical methods. A common mistake made by authors is to thoughtlessly copy the choice of the statistical tests from other authors analysing similar data. This strategy, although sometimes correct, may lead to an incorrect choice of statistical tools and incorrect interpretations. Here, I aim to give some advice on how to choose suitable statistical methods and how to present the results of statistical analyses.

Important limits in the use of statistics

Statistical tools face a number of constraints. Constraints should already be considered at the stage of planning the research, as mistakes made at this stage may make statistical analyses impossible. Therefore, careful planning of sampling is critical for future success in data analyses. The most important is ensuring that the general population is sampled randomly and independently, and that the experimental design corresponds to the aims of the research. Planning a control group/groups is of particular importance. Without a suitable control group, any further inference may not be possible. Parametric tests are stronger (it is easier to reject a null hypothesis), so they should always be preferred, but such methods can be used only when the data are drawn from a general population with normal distribution. For methods based on analysis of variance (ANOVA), residuals should come from a general population with normal distribution, and in this case there is an additional important assumption of homogeneity of variance. Inferences made from analyses violating these assumptions may be incorrect.

Statistical inference

Statistical inference is asymmetrical. Scientific discovery is based on rejecting null hypotheses, so interpreting non-significant results should be taken with special care. We never know for sure why we fail to reject the null hypothesis. It may indeed be true, but it is also possible that our sample size was too small or variance too large to capture the differences or relationships. We also may fail just by chance. Assuming a significance level of $p = 0.05$ means that we run the risk of rejecting a null hypothesis in 5% of such analyses. Thus, interpretation of non-significant results should always be accompanied by the so-called power analysis, which shows the strength of our inference.

Experimental design and data analyses

The experimental design is a critical part of study planning. The design must correspond to the aims of the study presented in the Introduction section. In turn, the statistical methods must be suited to the experimental design so

✉ Mariusz Cichoń
mariusz.cichon@uj.edu.pl

[1] Institute of Environmental Sciences, Jagiellonian University, Gronostajowa 7, 30-376 Kraków, Poland

that the data analyses will enable the questions stated in the Introduction to be answered. In general, simple experimental designs allow the use of simple methods like t-tests, simple correlations, etc., while more complicated designs (multifactor designs) require more advanced methods (see, Fig. 1). Data coming from more advanced designs usually cannot be analysed with simple methods. Therefore, multifactor designs cannot be followed by a simple t-test or even with one-way ANOVA, as factors may not act independently, and in such a case the interpretation of the results of one-way ANOVA may be incorrect. Here, it is particularly important that one may be interested in a concerted action of factors (interaction) or an action of a given factor while controlling for other factors (independent action of a factor). But even with one factor design with more than two levels, one cannot use just a simple t-test with multiple comparisons between groups. In such a case, one-way ANOVA should be performed followed by a post hoc test. The post hoc test can be done only if ANOVA rejects the null hypothesis. There is no point in using the post hoc test if the factors have only two levels (groups). In this case, the differences are already clear after ANOVA.

Description of statistical methods in the Materials and methods section

It is in the author's interest to provide the reader with all necessary information to judge whether the statistical tools used in the paper are the most suitable to answer the scientific question and are suited to the data structure. In the Materials and methods section, the experimental design must be described in detail, so that the reader may easily understand how the study was performed and later why such specific statistical methods were chosen. It must be clear whether the study is planned to test the relationships or differences between groups. Here, the reader should already understand the data structure, what the dependent variable is, what the factors are, and should be able to determine, even without being directly informed, whether the factors are categorical or continuous, and whether they are fixed or random. The sample size used in the analysis should be clearly stated. Sometimes sample sizes used in analyses are smaller than the original. This can happen for various reasons, for example if one fails to perform some measurements, and in such a case, the authors must clearly explain why the original sample size differs from the one used in the analyses. There must be a very good reason to omit existing data points from the analyses. Removing the so-called outliers should be an exception rather than the rule.

A description of the statistical methods should come at the end of the Materials and methods section. Here, we start by introducing the statistical techniques used to test predictions formulated in the Introduction. We describe in detail the structure of the statistical model (defining the dependent variable, the independent variables—factors, interactions if present, character of the factors—fixed or random). The variables should be defined as categorical or continuous. In the case of more advanced models, information on the methods of effects estimation or degrees of freedom should be provided. Unless there are good reasons, interactions should always be tested, even if the study is not aimed at testing an interaction. If the interaction is not the main aim of the study, non-significant interactions should be dropped from the model and new analyses without interactions should be carried out and such results reported. If the interaction appears to be significant, one cannot remove it from the model even if the interaction is not the main aim of the study. In such a case, only the interaction can be interpreted, while the interpretation of the main effects is not allowed. The author should clearly describe how the interactions will be dealt with. One may also consider using a model selection procedure which should also be clearly described.

The authors should reassure the reader that the assumptions of the selected statistical technique are fully met. It must be described how the normality of data distribution and homogeneity of variance was checked and whether these assumptions have been met. When performing data transformation, one needs to explain how it was done and whether the transformation helped to fulfil the assumptions of the parametric tests. If these assumptions are not fulfilled, one may apply non-parametric tests. It must be clearly stated why non-parametric tests are performed. Post hoc tests can be performed only when the ANOVA/Kruskal–Wallis test shows significant effects. These tests are valid for the main effects only when the interaction is not included in the model. These tests are also applicable for significant interactions. There are a number of different post hoc tests, so the selected test must be introduced in the materials and methods section.

The significance level is often mentioned in the materials and methods section. There is common consensus among researchers in life sciences for a significance level set at $p = 0.05$, so it is not strictly necessary to report this conventional level unless the authors always give the I type error (p-value) throughout the paper. If the author sets the significance level at a lower value, which could be the case, for example, in medical sciences, the reader must be informed about the use of a more conservative level. If the significance level is not reported, the reader will assume $p = 0.05$. In general, it does not matter which statistical software was used for the analyses. However, the outcome may differ slightly between different software, even if exactly the same model is set. Thus, it may be a good practice to report the name of the software at the end of the subsection describing the

statistical methods. If the original code of the model analysed is provided, it would be sensible to inform the reader of the specific software and version that was used.

Presentation of the outcome in the Results section

Only the data and the analyses needed to test the hypotheses and predictions stated in the Introduction and those important for discussion should be placed in the Results section. All other outcome might be provided as supplementary materials. Some descriptive statistics are often reported in the Results section, such as means, standard errors (SE), standard deviation (SD), confidence interval (CI). It is of critical importance that these estimates can only be provided if the described data are drawn from a general population with normal distribution; otherwise median values with quartiles should be provided. A common mistake is to provide the results of non-parametric tests with parametric estimates. If one cannot assume normal distribution, providing arithmetic mean with standard deviation is misleading, as they are estimates of normal distribution. I recommend using confidence intervals instead of SE or SD, as confidence intervals are more informative (non-overlapping intervals suggest the existence of potential differences).

Descriptive statistics can be calculated from raw data (measured values) or presented as estimates from the calculated models (values corrected for independent effects of other factors in the model). The issue whether estimates from models or statistics calculated from the raw data provided throughout the paper should be clearly stated in the Materials and methods section. It is not necessary to report the descriptive statistics in the text if it is already reported in the tables or can be easily determined from the graphs.

The Results section is a narrative text which tells the reader about all the findings and guides them to refer to tables and figures if present. Each table and figure should be referenced in the text at least once. It is in the author's interest to provide the reader the outcome of the statistical tests in such a way that the correctness of the reported values can be assessed. The value of the appropriate statistics (e.g. F, t, H, U, z, r) must always be provided, along with the sample size (N; non-parametric tests) or degrees of freedom (df; parametric tests) and I type error (p-value). The p-value is an important information, as it tells the reader about confidence related to rejecting the null hypothesis. Thus one needs to provide an exact value of I type error. A common mistake is to provide information as an inequality ($p < 0.05$). There is an important difference for interpretation if $p = 0.049$ or $p = 0.001$.

The outcome of simple tests (comparing two groups, testing relationship between two variables) can easily be reported in the text, but in case of multivariate models, one may rather report the outcome in the form of a table in which all factors with their possible interactions are listed with their estimates, statistics and p-values. The results of post hoc tests, if performed, may be reported in the main text, but if one reports differences between many groups or an interaction, then presenting such results in the form of a table or graph could be more informative.

The main results are often presented graphically, particularly when the effects appear to be significant. The graphs should be constructed so that they correspond to the analyses. If the main interest of the study is in an interaction, then it should be depicted in the graph. One should not present interaction in the graph if it appeared to be non-significant. When presenting differences, the mean or median value should be visualised as a dot, circle or some other symbol with some measure of variability (quartiles if a non-parametric test was performed, and SD, SE or preferably confidence intervals in the case of parametric tests) as whiskers below and above the midpoint. The midpoints should not be linked with a line unless an interaction is presented or, more generally, if the line has some biological/logical meaning in the experimental design. Some authors present differences as bar graphs. When using bar graphs, the *Y*-axis must start from a zero value. If a bar graph is used to show differences between groups, some measure of variability (SD, SE, CI) must also be provided, as whiskers, for example. Graphs may present the outcome of post hoc tests in the form of letters placed above the midpoint or whiskers, with the same letter indicating lack of differences and different letters signalling pairwise differences. The significant differences can also be denoted as asterisks or, preferably, p-values placed above the horizontal line linking the groups. All this must be explained in the figure caption. Relationships should be presented in the form of a scatterplot. This could be accompanied by a regression line, but only if the relationship is statistically significant. The regression line is necessary if one is interested in describing a functional relationship between two variables. If one is interested in correlation between variables, the regression line is not necessary, but could be placed in order to visualise the relationship. In this case, it must be explained in the figure caption. If regression is of interest, then providing an equation of this regression is necessary in the figure caption. Remember that graphs serve to represent the analyses performed, so if the analyses were carried out on the transformed data, the graphs should also present transformed data. In general, the tables and figure captions must be self-explanatory, so that the reader is able to understand the table/figure content without reading the main text. The table caption should be written in such a way that it is possible to understand the statistical analysis from which the results are presented.

Guidelines for the Materials and methods section:

- Provide detailed description of the experimental design so that the statistical techniques will be understandable for the reader.
- Make sure that factors and groups within factors are clearly introduced.
- Describe all statistical techniques applied in the study and provide justification for each test (both parametric and non-parametric methods).
- If parametric tests are used, describe how the normality of data distribution and homogeneity of variance (in the case of analysis of variance) was checked and state clearly that these important assumptions for parametric tests are met.
- Give a rationale for using non-parametric tests.
- If data transformation was applied, provide details of how this transformation was performed and state clearly that this helped to achieve normal distribution/homogeneity of variance.
- In the case of multivariate analyses, describe the statistical model in detail and explain what you did with interactions.
- If post hoc tests are used, clearly state which tests you use.
- Specify the type of software and its version if you think it is important.

Guidelines for presentation of the outcome of statistical analyses in the Results section:

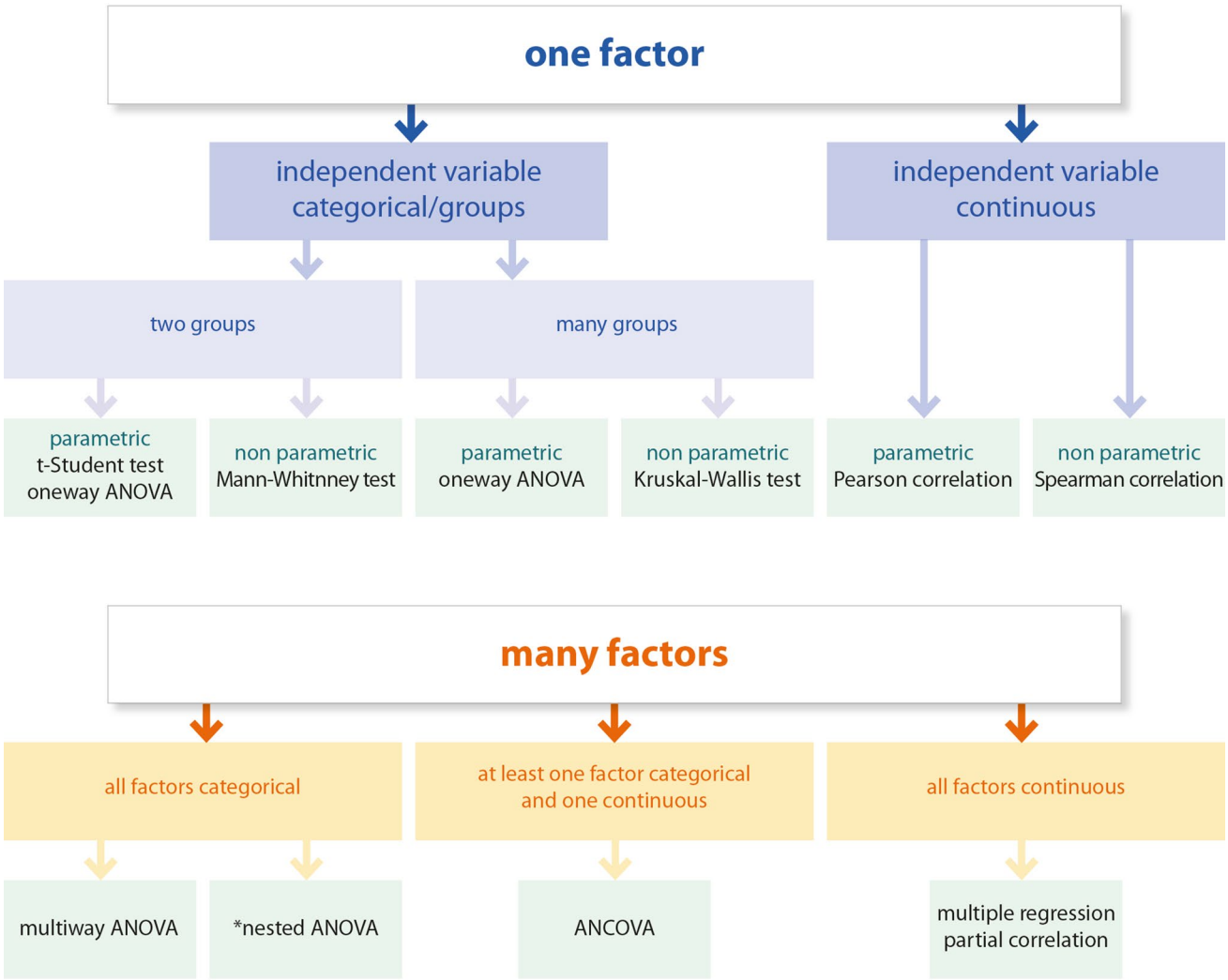
- Make sure you report appropriate descriptive statistics—means, standard errors (SE), standard deviation (SD), confidence intervals (CI), etc. in case of parametric tests or median values with quartiles in case of non-parametric tests.
- Provide appropriate statistics for your test (t value for t-test, F for ANOVA, H for Kruskal–Wallis test, U for Mann–Whitney test, χ^2 for chi square test, or r for correlation) along with the sample size (non-parametric tests) or degrees of freedom (df; parametric tests).

Examples:
$t_{23} = 3.45$ (the number in the subscript denotes degree of freedom, meaning the sample size of the first group minus 1 plus the sample size of the second group minus 1 for the test with independent groups, or number of pairs in paired t-test minus 1).
$F_{1,23} = 6.04$ (first number in the subscript denotes degrees of freedom for explained variance—number of groups within factor minus 1, second number denotes degree of freedom for unexplained variance—residual variance). F-statistics should be provided separately for all factors and interactions (only if interactions are present in the model).
$H = 13.8$, $N_1 = 15$, $N_2 = 18$, $N_3 = 12$ (N_1, N_2, N_3 are sample sizes for groups compared).
$U = 50$, $N_1 = 20$, $N_2 = 19$ for Mann–Whitney test ($N_1$and N_2 are sample sizes for groups).
$\chi^2 = 3.14$ df = 1 (here meaning e.g. 2×2 contingency table).
$r = 0.78$, $N = 32$ or df = 30 (df = N − 2).

- Provide exact p-values (e.g. $p = 0.03$), rather than standard inequality ($p \leq 0.05$)
- If the results of statistical analysis are presented in the form of a table, make sure the statistical model is accurately described so that the reader will understand the context of the table without referring to the text. Please ensure that the table is cited in the text.
- The figure caption should include all information necessary to understand what is seen in the figure. Describe what is denoted by a bar, symbols, whiskers (mean/median, SD, SE, CI/quartiles). If you present transformed data, inform the reader about the transformation you applied. If you present the results of a post hoc test on the graph, please note what test was used and how you denote the significant differences. If you present a regression line on the scatter plot, give information as to whether you provide the line to visualise the relationship or you are indeed interested in regression, and in the latter case, give the equation for this regression line.

Further reading in statistics:

1. Sokal and Rolf. 2011. Biometry. Freeman.
2. Zar. 2010. Biostatistical analyses. Prentice Hall.
3. McDonald, J.H. 2014. Handbook of biological statistics. Sparky House Publishing, Baltimore, Maryland.
4. Quinn and Keough. 2002. Experimental design and data analysis for biologists. Cambridge University Press.



* when one factor has no full representation in other factors (factors do not cross)

Fig. 1 Test selection chart