SEEIUC
Sociedad Española de Enfermería
Intensiva y Unidades Coronarias

Enfermería *Intensiva*

www.elsevier.es/ei

COMMENTS ON RESEARCH ARTICLES

Ten common statistical mistakes to watch out for when writing or reviewing a manuscript☆

Diez errores estadísticos frecuentes que tener en cuenta al escribir o revisar un artículo

Makin TR, Orban de Xivry JJ. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. Elife. 2019;8:e48175. doi:10.7554/eLife.48175

Abstract

Inspired by broader efforts to make the conclusions of scientific research more robust, we have compiled a list of some of the most common statistical mistakes that appear in the scientific literature. The mistakes have their origins in ineffective experimental designs, inappropriate analyses and/or flawed reasoning. We provide advice on how authors, reviewers, readers can identify, resolve these mistakes and, we hope, avoid them in the future.

Absence of a suitable control group or condition

The problem. On occasions clinical research wishes to assess the effect of an intervention without the use of a control group. If follow-up of a group of patients is made after an intervention and the outcome variable is evaluated before and after the intervention, the change in this variable could be assumed to be due mainly to the effect of the intervention without bearing in mind the effect of time, and this is not habitually assumable. Appropriate inclusion of a control group is therefore important.

How to detect it. When we observe in an article that the data refer to a single group, or several groups, but there is no suitable control group.

Solutions for researchers. Experimental design is essential for preventing these biases, selecting both groups at the same time, assigning the participants, developing identical handling and promoting blinded studies both for the participants and for the researchers. If the experimental design does not allow for the separation of the effect of time from the effect of intervention, then the conclusions regarding impact of intervention must be presented as preliminary.

Interpret comparisons between 2 effects without them being directly compared

The problem. Conclusions on the impact of an intervention are often based on nothing more than the statistically significant effect of treatment in the experimental group compared with a non significant effect in the control group, based on 2 statistical tests made independently. In actual fact, the correlation between 2 variables in one group may be statistically significant and not be so in another group with a similar correlation coefficient. This may even occur if the relationship between the 2 variables is virtually identical in the 2 groups, and it should therefore not be inferred that one correlation is better than the other without using a single statistical proof to compare the 2 effects.

How to detect it. This is observed when the conclusion extracted with respect to the difference between 2 effects is given without having been statistically compared between them.

Solutions for the researchers. The correlation between 2 groups may be compared with Monte Carlo simulation, with an ANOVA test and even with non parametric statistical tests. The meta-analysis network procedures can compare multiple treatments simultaneously in a single analysis, combining direct and indirect tests within a systematic review of randomised clinical trials.

Exaggerate analysis units

The problem. The experimental unit is defined as the smallest observation which may be assigned randomly and independently. Without a clear identification of the appro-

DOI of original article: https://doi.org/10.1016/j.enfi.2020.09.002

☆ Please cite this article as: del Campo-Albendea L, Muriel-García A. Diez errores estadísticos frecuentes que tener en cuenta al escribir o revisar un artículo. Enferm Intensiva. 2021;31:42–4.

2529-9840/

priate unit for evaluating an effect, there may be a high and adulterated number of experimental units that lead to erroneous statistical inference.

How to detect it. In the section on methodology the appropriate unit of analysis has to be described. In other words, if the purpose of the study is to understand the group effects, the unit of analysis has to reflect the variance between the subjects. Often several measures are carried out on the same patient, for example when paired organs are assessed (eyes, kidneys, or lungs), when the same subject is evaluated in several measures over time or when the effect of a cluster level intervention is evaluated, for example, when nursing controls are randomized, but patient data are collected.

Solutions for the researchers. The best available solution is the use of lineal models of mixed effects, so that they can include all the data in the model without violating the supposed independence. However, advanced statistical knowledge is required and the outcomes must be cautiously interpreted.

Misleading correlations

The problem. Correlation is an extremely important tool in terms of evaluating the magnitude of an association between 2 variables, but parametric correlations (e.g. Pearson r) has a series of suppositions which, when not fulfilled, may lead to misleading correlations. Misleading correlations occur when one of many *outliers* (outside range values) are present in one of the 2 variables, since a single value may inflate the correlation coefficient. The *outliers* values may provide extreme observations that obey the phenomenon being studied, and therefore eliminating extreme data should be treated with caution.

How to detect it. This may be detected in the results, paying particular attention to the correlations which are not accompanied by a dispersion graph and considering whether enough justification has been provided when any extreme data has been eliminated.

Solutions for the researchers. Robust correlation methods (e.g. *bootstrapping*) are less sensitive to extreme values, since they take into consideration data structure.

Using small sample sizes

The problem. A small sample size may only detect major effects and is also more susceptible to the real effect present in the data not being found (type II error). Furthermore, distribution of a small sample tends to deviate from normal distribution and the limited size makes it often impossible to rigorously prove the assumption of normality.

How to detect it. The reviewers must critically examine the sample size used in the article and judge whether it is sufficiently statistically powerful enough to conclude the different results put forward.

Solutions for the researchers. The best way of solving these problems is to a priori develop statistical power analysis. The Bayesian statistics offers possibilities to determine sufficient statistical power to identify post hoc effects. In cases where sample size cannot be extended, it is necessary to provide replications or include sufficient controls (e.g. by establishing confidence intervals).

Circular analysis

The problem. This is based on selecting the data that characterise the dependent variables and using these data for an initial characterisation of study variables and then later carrying out statistical inferences with them. Regular circular forms of analysis are shown in the search for the effect of a treatment in subgroups created with explicit criteria prior to conducting the study, but based on the results of the study itself.

How to detect it. Initially it is always the case that statistical tests are weighted by the selection of a criterion in favour of the hypothesis that is being evaluated. The reviewers have to be alert to high impossible effects which are not plausible in theory and/or are based on relatively unreliable measurements. In these cases the authors should justify the Independence between the selection criterion and the effect of interest.

Solutions for the researchers. Defining the analysis criterion previously and independently from the data, with the most direct solution being to use different databases to specify analysis parameters and test out predictions. This division may be made at participant level (using a different group) or at a test level (using different tests for all participants). This may be achieved without losing statistical power by using approximations through bootstrapping.

Analysis flexibility: significant p value hunters

The problem. P-hacking refers to the practice of manipulating data (replacing response parameters, adding co variables, excluding subjects, etc.) until the result passes the threshold of statistical error. Estimating a p value in a database is not necessarily complicated and any researcher may provide a plausible explanation for any effect. Due to this, it is important to previously define statistical analyses to be used, experimentation design, or to later carry out study replication.

How to detect it. p-hacking is difficult to detect since all the necessary information is rarely broken down. It may be considered if all analysis choices are not well justified, if the same analytical proposal was not used in previous publications, if the researchers presented with a new variable that is not normal or if a large number of variables was collected only presenting with some significances in outcomes.

Solutions for the researchers. Exploratory analysis based on a preliminary study of flexible data are correct if they are reported and interpreted in a transparent way and especially if they serve as a replication base. These analyses may be a valid justification for additional research but may never be the basis of major conclusions. Possibly the best way of anticipating p-hacking is to show a certain tolerance for non significant results: if the experiment is well designed, executed and analysed, the reviewer cannot ''punish'' the researchers for their data.

Not correcting multiple comparisons

The problem. An effect is often explored in multiple variables, normally with an indeterminate hypothesis *a priori*, which is known as exploratory analysis. In frequentrist analysis, carrying out multiple comparisons during exploratory analysis may increase the probability of detecting a significant effect, even if this effect does not exist (false positive, type I error), due to the repetitive use of statistical evidence.

How to detect it. This may be detected in methodology and results. When exploratory analysis with multiple variables is performed, it is inacceptable to interpret the results which have not overcome the multiple comparison corrections without justification. Even if a robust prediction is proposed, if this prediction cannot be proven in multiple independent comparisons, a correction is required for multiple comparisons.

Solutions for the researchers. The researchers must reveal all variables measured and appropriately implement the use of corrections of multiple comparisons, justifying the reason why for a specific test.

Over interpreting insignificant outcomes

The problem. An insignificant p value may mean that a result is truly null and void, which is an effect without sufficient statistical power for its evaluation or an ambiguous effect. To interpret an insignificant result as evidence against the hypothesis, it would be necessary to demonstrate that this evidence is significant. This means that results found close to .05 should not be assumed to be unsatisfactory when really they provide preliminary evidence which requires further attention.

How to detect it. In the section on results or conclusions. An insignificant p value may be interpreted or described as indicative of the effect not being present at all. This error is very common and should be highlighted as problematic.

Solutions for the researchers. A major first step is to report the effect size together with the p value in order to provide information on the magnitude of the effect, which is equally important for any future meta-analysis. For example, if an insignificant effect in a study with a large sample size is also very small in magnitude, it is improbable that it be theoretically significant whilst one with a moderate effect size may potentially justify further research. Moreover, the researchers may already have determined a priori whether they have sufficient statistical power to identify the desired effect, or to determine whether the confidence intervals of this prior effect contain the zero. If not, the researchers should not over interpret insignificant results and should only describe them as non significant.

Correlation and causality

The problem. Correlation is commonly used to explore the relationship between 2 variables, usually assuming that one is the cause of the other. However, just because 2 variables appear to occur lineally does not necessarily mean there is a causal relationship between them, even if this association is plausible. Correlations cannot be used separately as evidence of a cause and effect relationship.

How to detect it. Researchers should used only causal language if the relationship between 2 or more variables is due to an appropriate analysis from a methodological and statistical viewpoint, and even then they should be cautious regarding the role of a third variable or factors of confusion.

Solutions for the researchers. An attempt should be made to explore the relationship with a third variable to provide further support in interpretations. For example, using mediation analysis or index of tendency. From the point of view of research design, the only study for the majority of authors which may provide answers to questions of causality is a randomized, controlled clinical trial whenever it is possible to perform it. If not, causal language should be avoided when evidence is from correlation.

L. del Campo-Albendea (MsC)[a],
A. Muriel-García (MsC, PhD)[b,*]
[a] *Graduada en Biología, Universidad Complutense de Madrid, Estudiante Máster de Bioestadística, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, Madrid, Spain*
[b] *Doctor por la Universidad Autónoma de Madrid, Unidad de Bioestadística Clínica, Hospital Ramón y Cajal, IRYCIS, CIBERESP, Madrid, Spain*

[*] Corresponding author.
E-mail address: alfonso.muriel@hrc.es (A. Muriel-García).