Open Access Journal of Behavioural Science & Psychology
ISSN: 2642-0856

Academic Strive
Researcher's perception is our reality

Editorial *Volume 2; Issue 2*

Statistical Methodological Challenges in Social Sciences

Clarissa Ferrari*

Unit of Statistics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Italy

***Corresponding author:** Dr. Clarissa Ferrari, Unit of Statistics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Via Pilastroni 4, Brescia, Italy, Tel: +390303501722; Email: cferrari@fatebenefratelli.eu

Received Date: October 03, 2019; **Published Date:** October 15, 2019

Editorial

The term “Big Data” has recently become mainstream [1], especially thanks to the ease in sharing and producing data: think, for example, of open-source repository infrastructures of scientific or genetic data or of data collected through social media platforms. This gives rise to a substantial increase in the available data in all scientific fields, including social and behavioral sciences. The possibility of combining data from various sources promotes new chances of knowledge by allowing, as never before, to deeply analyze multiple interrelationships, across measures and variables. This is of particular interest in the social sciences where multiple domains and factors need to be evaluated and interpreted simultaneously. However, if on the one hand, the availability of large amount of data opens interesting scenarios for new applications and analyses, on the other hand, it gives rise to new challenges regarding the management and the analysis of these complex data. In such a context, expertise in research design and in statistical analysis becomes crucial to produce high quality scientific research [2].

Although the need to manage massive amount of complex data it is nowadays an acknowledged fact in almost all the scientific communities - including the social sciences - the involvement of a professional statistician in all the phases of a scientific research (study design, measurements quality check, analyses and interpretation of results) is still an exception. The inappropriate use of statistical methods is a serious problem leading to biased results, incorrect conclusions and to severe clinical, and thus unethical, consequences [3]. The risk is that the increasing availability of complex will exacerbate this problem. Moreover, it is not uncommon to find severe methodological errors and statistical pitfalls even in non-complex data contexts in which the basic knowledge of data analysis should be well-established.

In the following, the most common statistical methodological errors and misinterpretation of the results will be presented, by providing some simple rules to avoid and overcome them. The hope is that the following suggestions will serve as a guide for both clinical and social science studies. A good research starts from a proper definition of the experimental design that should include:

- A clear definition of the investigation hypotheses that drive the choice of the experimental design (e.g. descriptive studies vs analytical studies vs controlled studies)
- The setting of an appropriate sampling procedure (coherent with the design) with an adequate sample size.

The correct choice of the experimental design is crucial for all the subsequent steps of the data analysis and in particular for the sampling procedure. An adequate sample size, that has to be computed by an expert in statistical tests, will avoid underpowered studies and improve the robustness of results. Moreover, researchers should pay particular attention to the representativeness of the study group: if the analyzed group is not coherent with the studied population in terms of main features related with the experimental hypotheses, the research conclusions cannot be generalized to the population.

Another important assessment to check before running any analytical procedure regards the evaluation of the data distribution characteristics and scale of the target measurements, as well as the number of groups or experimental conditions to analyze. These issues are strictly related and should drive the choice of the statistical test and/or model to perform. Many tests and inferential models require to meet the Gaussian assumption of the data distribution: this holds e.g. for t-test, Analysis of Variance (ANOVA) test or for applying the standard linear models (regression, ANOVA/ANCOVA models) [4]. If the Normal assumption is not met, other tests or models have to be chosen, e.g. non-parametric tests as the Mann-Whitney or Kruskal-Wallis; or generalized linear models [5]. It is worth to note that the use of an incorrect test/model could produce biased and often untruthful results with repercussions on the research reproducibility and on ethical issues.

A very topical issue is the correct interpretation of the p-value and of the corresponding inferential deduction based on it [6]. In this regard, it is necessary to stress that the p-value strongly depends on the sample size: for this reason big-datasets can give rise to extremely low p-values regardless of the clinical effect. In addition, the presence of biases, as confounding, can indirectly affect the p-value. It is essential, thus, to discern between statistical and clinical significance by providing a range of potential explanations for the results. Factors such as background evidence and underlying mechanism are often as important as statistical measures like p-values [7].

Not less important, in terms of frequency with which the error occurs, it is the misuse and misinterpretation of correlation analysis results.

It is mandatory to clear the misunderstanding about the correlation-causality misuse: no conclusion on causality can be derived from correlation analysis. The correlation coefficient is a purely descriptive measure of the association between two quantitative variables and no further inference can be derived by its analysis. Moreover, the statistical test on correlation establishes whether the coefficient is statistically different from zero, but the evaluation of the strength of the association should be based only on the absolute value of the coefficient (see e.g. [8]). It is likewise worth to note that the correlation test should never be adjusted for multiple testing, except for the case of family wise error rate, i.e. when multiple tests have to be summarized in just one conclusion about the whole experiment. A clear explanation about this important issue can be found in [9].

Although many other methodological advises would deserve to be mentioned, an efficient guide line against methodological biases and statistical errors is based on reproducible research practice that includes:

i) an accurate planning of the study design,

ii) a careful choice of the statistical tests,

iii) A thoughtful selection of advanced statistical methods and analyses. To guarantee the reliability of these practices, the social sciences scientists should consider the full involvement in all the research phases of an expert with professional skills in statistical methodology.

References

1. Galeano P, Peña D (2019) Data science, big data and statistics. Test 28(2): 289-329.

2. Lam SW, Bauer SR, Yang W, Miano TA (2017) Statistics Myth Busters: Dispelling Common Misperceptions Held by Readers of the Biomedical Literature. Annals of Pharmacotherapy 51(5): 429-438.

3. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H (2007) Statistical errors in medical research - A review of common pitfalls. Swiss Med Wkly 137(3-4): 44-49.

4. Rosner B (2011) Fundamentals of Biostatistics, (5th edn), Pacific Grove, CA Duxbury 2000.

5. Agresti A (2015) Foundations of Linear and Generalized Linear Models. ISBN: 978-1-118-73003-4.

6. Ioannidis JPA (2019) What Have We (Not) Learnt from Millions of Scientific Papers with P Values? The American Statistician 73(1): 20-25.

7. Amrhein V, Greenland S, Mcshane B (2019) Statistical significance. Nature.

8. Dennis E. Hinkle, William Wiersma SGJ (2003) Applied Statistics for the Behavioral Sciences (5th edn), Boston, Mass: Houghton Mifflin.

9. Bender R, Lange S (2001) Adjusting for multiple testing - When and how? J Clin Epidemiol 54(4): 343-349.