Opinion

Missing values: inputting and alternatives

Karpenko Dmitriy Vladimirovich*

* d_@list.ru

https://orcid.org/0000-0002-0691-4079

National Medical Research Center for Hematology, Moscow, Russian Federation (Russia) 125167, Moscow, Novyi Zykovskii proezd, 4

Abstract

In the contemporary era, data processing has emerged as a pivotal activity, exerting a profound influence on the advancement of numerous domains, including those pertaining to economic, social, and technological development. In recent decades, there have been notable developments in the methods and tools used for data processing. Notwithstanding the considerable advances that have been made, the adaptation of tools and personnel to changes in specific fields of activity is occurring with a notable lag. The following discussion will focus on the processing of data with missing values in the context of the tasks of biology and bioinformatics.

Keywords:

Missing values, outliers, inputting

It is inevitable that data will be lost during the processes of data acquisition, whether the data is obtained from measuring devices or from surveys of people. In addition, medical data may be identified in instances where the configuration of an experiment or the acquisition of new data is challenging. Adherence to the prescribed procedures can reduce the amount of inaccessible information, although it cannot be entirely eliminated [1]. This issue is particularly pertinent in the context of orphan diseases, where each patient represents a crucial case for study. In such cases, data are accumulated by a variety of organizations over an extended period of time. The data obtained over an extended period may reflect alterations associated with the natural progression of treatment and the utilization of specific tools. The existence of discrepancies in the data may result in the emergence of entirely new categories of unmeasured values for a portion of the sample. The handling of individual characteristics in the presence of missing data does not typically give rise to concerns. The situation becomes more complex when it is necessary to simultaneously analyses a large number of studied features. It is important to note that the issue affects the problem of outliers. The deletion of values considered as outliers is a standard technique, which in turn generates the problem of lost data. Many standard tools require complete data matrices without missing values. In the literature, a wide range of techniques for working with lost data can be found [2].

The primary question is whether the data is lost in a random or biased manner. The analysis of data in which loss is a non-accidental occurrence represents a considerably more intricate undertaking, necessitating an examination of this pattern [3]. The most straightforward methods involve the removal of features or objects that contain lost values. Such approaches are acceptable when the proportion of missing data is small; however, they are problematic for studies with small groups or large proportions of missing values.

A variety of approaches for filling in lost values are presented in the literature. This process is referred to as inputting. The values to be inserted can be selected from other samples included in the analysis or from external sources. This is known as cold or hot inputting, respectively. The data may be entered according to the estimated mean or in consideration of the relationship with the values of other features within the sample. The data may be completed in accordance with the information regarding the k-nearest neighbors [4]. It is possible to note approaches that suggest the performance of multiple imputations with subsequent analysis of the results [5]. It is also noteworthy that some studies have proposed the replacement of not only missing data but also outliers [6].

It is crucial to acknowledge that during the input phase, we incorporate data that, at best, will not be crucial to the outcome of the analysis. The additional information may be biased and thus distort the result. The injected information is not directly connected to the object of study; at best, it can be linked to the object through other values. This emphasizes the significance of analytical tools that are capable of processing data in a direct manner, obviating the necessity for the imputation of missing values.

Approaches that do not involve data inputting are described in the literature [7,8]. Additionally, there are ongoing research projects whose findings are presented in the form of preprints [9,10]. It is important to note that these approaches do not require inputting, either in advance or implicitly. Rather, they compensate for anomalies through the use of alternative solutions. Notwithstanding the existence of solutions at the algorithmic level, they cannot be directly applied in the context of biology and medicine. It is essential to integrate these approaches into user-friendly tools that can be readily employed by specific user groups. In a series of recent papers, we demonstrated the efficacy of an algorithm that is insensitive to lost

data in addressing the challenges of prediction and classification in small groups [11]. Further work is currently underway to develop the algorithm as a standalone tool. Another algorithm for analyzing differential gene expression is capable of robustly processing both lost data and outliers. It is assembled as a ready-to-use tool and is presented in detail in a preprint [12].

A review of recent literature reveals a significant number of studies employing data inputting for a variety of purposes. Notwithstanding the aforementioned criticism, data inputting is performed correctly in many cases without distortion of the result. Furthermore, alternative methods may be prohibitively expensive in terms of computational resources and implementation complexity, which can result in their de facto rejection. As experts in our respective fields, we often lack a comprehensive understanding of related disciplines, which impedes our ability to rapidly address multidisciplinary challenges. It is crucial to not only develop the algorithms themselves, but also to implement them in the form of final tools. The issue of missing values illustrates the significance of this observation. It is not feasible to leave the task of developing tools to biologists and medical professionals, who are the primary users of such tools. It is essential that mathematicians, programmers, and end-users collaborate to ensure the accurate formulation of the problem and the development of an effective solution. The fields of biology and medicine present complex and intriguing challenges that may initially appear straightforward but require a more nuanced examination.

Competing interests
The author declare the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1 Li, T., Hutfless, S., Scharfstein, D.O., Daniels, M.J., Hogan, J.W., Little, R.J.A., Roy, J.A., Law, A.H. and Dickersin, K. (2014) Standards Should Be Applied in the Prevention and Handling of Missing Data for Patient-Centered Outcomes Research: A Systematic Review and Expert Consensus, *Journal of Clinical Epidemiology*, **67**, 15–32, doi: 10.1016/J.JCLINEPI.2013.08.013.

2 Josse, J. and Reiter, J.P. (2018) Introduction to the Special Section on Missing Data, *https://doi.org/10.1214/18-STS332IN*, **33**, 139–141, doi: 10.1214/18-STS332IN.

3 Samuelson, D.A. and Spirer, H.F. (2016) Chapter 3. Use of Incomplete and Distorted Data in Inference About Human Rights Violations, *Human Rights and Statistics*, 62–78, doi: 10.9783/9781512802863-006/HTML.

4 Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001) Missing Value Estimation Methods for DNA Microarrays, *Bioinformatics*, **17**, 520–525, doi: 10.1093/BIOINFORMATICS/17.6.520.

5 Liu, Y., Wang, Y., Feng, Y. and Wall, M.M. (2016) Variable Selection and Prediction with Incomplete High-Dimensional Data, *https://doi.org/10.1214/15-AOAS899*, **10**, 418–450, doi: 10.1214/15-AOAS899.

6 Little, R.J.A. and Smith, P.J. (1987) Editing and Imputation for Quantitative Survey Data, *Journal of the American Statistical Association*, **82**, 58, doi: 10.2307/2289125.

7 Jiang, W., Josse, J. and Lavielle, M. (2020) Logistic Regression with Missing Covariates—Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework, *Computational Statistics & Data Analysis*, **145**, 106907, doi: 10.1016/J.CSDA.2019.106907.

8 Chechik, G., Heitz, G., Elidan, G., Abbeel, P. and Koller, D. (2007) Max-Margin Classification of Incomplete Data, *NIPS 2006: Proceedings of the 19th International Conference on Neural Information Processing Systems*, 233–240, doi: 10.7551/MITPRESS/7503.003.0034.

9 Josse, J., Chen, J.M., Prost, N., Varoquaux, G., Scornet, E., Josse, J. and Scornet, E. (2019) On the Consistency of Supervised Learning with Missing Values, *Preprint*, https://arxiv.org/abs/1902.06931v5.

10 Lounici, K. and Pacreau, G. (2023) Robust Covariance Estimation with Missing Values and Cell-Wise Contamination, *Preprint*, https://arxiv.org/abs/2306.00752v3.

11 Karpenko, D.V. and Bigildeev, A.E. (2023) Small Groups in Multidimensional Feature Space: Two Examples of Supervised Two-Group Classification from Biomedicine, *Journal of Bioinformatics and Computational Biology*, doi: 10.1142/S0219720023500257.

12 Karpenko, D. (2023) DEAr – Differential Expression Analyzer, *Preprint*, doi: 10.21203/RS.3.RS-2957165/V3.