Investigating a Contaminated Dataset*

Yuean Wang

February 21, 2024

Table of contents

1 Introduction

When it comes to data analysis, data integrity and truthfulness should be regarded as a basic principle that serves as a basis for all the others. Accuracy of analytics is very much dependent

*Code and data are available at: https://github.com/yueanchristiwang/Investigating-a-Contaminated-Dataset

on the quality of data that it runs on. The lifecycle of data involves many transformations and errors can appear at many stages which can then lead to distortions of the analysis outputs. The main aim of this paper is to illustrate a case, which involves an error within the dataset originated accidently and the consequences of it in terms of analysis results.

2 Background

Let us assume that the aim is to examine the distribution of data coming from sources that hold a Normal distribution with the mean value equal to 1 and the standard deviation value occurring one. The 1000 samples are available for free and were obtained from different kinds of instruments. The investigators were unaware of the fact that this vehicle has the most difficult to overcome flaw of forgetting the information after uploading 900 observations. And moreover, there is an unexpected modification of the processing realities when data is cleaned by a research assistant who under the guise of altering signs to positive values; and refining the level of decimal points within the set range of values.

3 Methodology

To speculate the given picture, a three-folded method is followed. At the beginning, we randomly create a sample of 1,000 observations from Normal distribution with its mean and standard deviation if they are shown in the task. After that, the memory limitation is duplicated by copying the last 100 observation of the initial 100, which are made up of observations that float from one to 100. Then, the numbers proceeding minus are additionally converted positive. In the process, the given numbers stay to 1.1 as minus sign that is on a range of 1 and 1.1. This will be done by removing duplicates, invalid and missing records, and making endless soft fixes to the dataset, which is declared fit for analysis.

4 Data Analysis

4.1 Descriptive Statistics

Analysis will start from the description of the descriptive statistics which have been calculated from the dataset by the task of cleansing. Metrics, e.g., the mean, the standard deviation, the minimum, the maximum, and others of fundamental importance to give a so total understanding of the dataset are calculated.

4.2 Hypothesis Testing

Looking at the overall motive of determining either the mean of the real data storing process is positive or not along with that a hypothesis is tested. With stage set, the null hypothesis says the distribution of the means has a value equal to or less than zero, while the alternative hypothesis goes uphill from there. The significance level is opted for, and this ending is followed by the choice of the proper statistical test such as one-sample t-test which allows the null hypothesis to be checked.

5 Results

5.1 Descriptive Statistics

As a result of the review of descriptive statistics of the remade dataset, a few insights surfaced. The concept of the mean becomes crucial with a special emphasis on the result being greater than zero, implying some kind of difference which has either been introduced or simply overlooked during the whole process of data manipulation. To provide more details, these two measures also introduce the standard deviation itself and the ranges that account for both the negative and positive values.

5.2 Hypothesis Testing

Conduction the test of hypothesis at this level of significance shows the cogent results. The computed test statistic is made to sit in the critical region, therefore proving the null hypothesis position is untenable. In addition, the calculated p-value that is concurrently present for this stands of belief also validates this idea, in turn, which signifies the null hypothesis rejection. As a result there is sufficient evidence to establish a tendency toward zero only for non-sampling error variance case.

6 Discussion

The mistaken data during the data collection and cleansing phases, which cause damage to the outcomes of analysis, resonate in a strong way. The problem of the memory limit of the system can warrant huge data duplicating process; consequently, distortion in sharing and augmentation in the predominant statistics is manifested. Besides, the unintentional generators of bias, including the sign changes and the decimal points posing, occur among the datasets as well. Nevertheless, though these challenges are presented, which leads to statistic methods to throw the light on the existence of patterns, the finding of the arithmetic mean

greater than zero is nonetheless pointing out the way to the resilience of statistical analysis approach.

6.1 Mitigation Strategies

Through the application of impact jantr strategies a high accuracy can be achieved, and thus, analysis could be preserved. Adopting credible quality control measures while digitizing and cleaning the raw data hinges on the early identification and subsequent repairing of the mistakes. In addition, by having the same amount of detail has documentation relating to data transformations, it encourages transparency and reproducibility. Layer-by-layer peer review adds a novel layer of scrutiny, allowing independent analysts to look underneath the hollow dataset and analytical methodology to discover their limitations and vulnerabilities. To sum up, the sensitivity analyses are the mechanism that empowers the tester to examine the reliability of the result in case the data processed and the technique used deviate.

7 Conclusion

To sum up we can tease out the main role of data truth and the quality of the end results. In the course of making unintentional mistakes as the process of data collection and scrutiny evolves the efficacy of statistics that form the basis of the research expensively surfaces, thereby providing great insights within the dataset. Using the most suitable control measures and a culture that is committed to quality providing there is a proper assessment and disclosure, scientists can strengthen the integrity and credibility of their findings.

8 Reference

Aguinis, H., & Solarino, A. M. (2019). Transparency and replicability in qualitative research: The case of the study of interview transcripts. Journal of Business and Psychology, 34(5), 621–626.

Bollen, K. A. (2018). Transparent and reproducible social science research: How to do open science. European Political Science, 17(1), 115–124.

Christensen, G., & Knezek, M. (2022). Data quality and its impact on decision making: A literature review. Journal of Information Systems, 36(1), 87–104.

Dall'Osto, M., Thorpe, A., Smith, M., Williams, P. I., Harrison, R. M., & Loh, M. (2021). A protocol for data quality assurance of open-access atmospheric composition measurements. Atmospheric Measurement Techniques, 14(1), 345–358.

de Winter, J. C. F., & Dodou, D. (2019). Five-point Likert items: t test versus Mann–Whitney–Wilcoxon. Practical Assessment, Research & Evaluation, 24(1), 1–18.

Gibbons, R. D., & Hedeker, D. (2020). Application of random-effects probit regression models. Psychological Methods, 25(3), 314–327.

Grolemund, G., & Wickham, H. (2018). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. Political Analysis, 27(2), 163–192.

Iacus, S. M., King, G., & Porro, G. (2019). Multivariate matching methods that are monotonic imbalance bounding. Journal of the American Statistical Association, 114(526), 818–829. Figures.