Linear models[*]

Jierui Miao

27 February 2024

Contents

0.1 Introduction

In the field of data analysis and statistical research, the integrity and reliability of data is of paramount importance. However, both human errors and technical failures can severely distort results and lead to misleading conclusions. In this analysis, we simulated a complex data collection and preparation scenario involving multiple mechanical and human errors to understand their impact on statistical inference. The real data generation process was a normal distribution with mean 1 and standard deviation 1, from which we drew a sample of 1,000 observations. However, due to instrumental limitations, the last 100 observations were overwritten by the first 100. In addition, half of the negative values were incorrectly converted to positive values during the data cleaning process, and the decimal positions of values between 1 and 1.1 were incorrectly shifted. Our goal is to assess whether these errors significantly affect our ability to correctly infer whether the mean value of the real data generation process is greater than zero.

0.2 Method

The simulation is performed in the statistical computing environment R. The first step involves the use of "set.seed" to ensure that the simulation values are the same every time it is opened. Then 1,000 observations are generated based on a specified normal distribution. However, the instrument used for data collection only has a maximum memory capacity of 900 observations, causing the first 100 entries to overwrite the last 100. This creates a significant bias on the initial data points and may skew the analysis. We randomly selected half of the negative observations and converted them to positive values. For observations between 1 and 1.1, we divided them by 10, misplacing the decimal point. After these manipulations, we performed a one-sample t-test to assess the hypothesis that the mean of the data was greater than zero.

By using R (R Core Team 2020), and R packages "tidyverse" (Wickham et al. 2019), "dplyr" (Wickham et al. 2021) and "knitr" (Xie 2021), The result of the simulation is shown below.

[*] Code and data are available at: https://github.com/jerry-maker-765/Linear-models.git

```
##
##  One Sample t-test
##
## data:  true_data
## t = 34.744, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.015941      Inf
## sample estimates:
## mean of x
##  1.066477

## [1] 1.066477
```

0.3 Result

As shown in the output provided, the results of the one-sample t-test indicate a t-value of 34.744, with degrees of freedom (df) equal to 999, yielding a p-value of less than 2.2e-16. This p-value is considerably lower than any of the traditional significance levels, such as 0.05 or 0.01, which strongly suggests that the mean of our manipulated dataset is significantly greater than 0. The 95% confidence interval ranges from approximately 1.015941 to infinity, and the value of the mean of the sample estimates is 1.066477, further supporting the hypothesis that the true mean is greater than 0. 1.015941 to infinity, and the value of 1.066477 for the mean of the sample estimates further supports the hypothesis that the true mean is greater than zero.

These results, while superficially illuminating, must be interpreted in the context of known errors affecting the data set. Due to the memory limitations of the instrument, a non-random sample was created by replicating the first 100 observations, which could have erroneously raised or lowered the estimated mean, depending on their values. Research assistants changed negative values to positive values, increasing the upward bias of the data and artificially inflating the mean. Finally, decimal place error introduces a random but systematic tightening of some values that may offset some upward bias but is somehow unrepresentative of the original data distribution.

0.4 Discussion

After correcting the dataset for the above errors, a one-sample t-test was conducted to assess whether the mean of the manipulated dataset significantly exceeded 0. The results of this test are critical because they are expected to highlight how the introduced errors may affect the statistical conclusions drawn from the flawed data.

The analysis revealed a distorted view of the initial DGP. The memory limitations of the instrument resulted in artificially inflated early data points, and modifications by the research assistants further exacerbated deviations from the true mean. In particular, negative-to-positive conversions artificially increased the overall mean, and misplaced decimal places for specific values introduced random variability unrelated to the true distribution.

The accuracy and precision of data analysis depends on the integrity of the data. Unless a comprehensive error detection and prevention strategy is employed, errors of all kinds (from instrument failure to human error) can affect data quality.(Svanks 1988)

Instruments require regular calibration checks and protocols to detect and log memory overwrites or failures. It uses algorithms to identify duplicate data, anomalous data points, or values that deviate from established standards so that corrective action can be taken as soon as possible. Human oversight plays a key role in advocating a rigorous data cleaning and validation process, including secondary reviews by independent analysts to reduce the risk of human error. Another indispensable tool available to researchers is sensitivity analysis to determine how data modifications affect the results of the analysis, thereby exposing potential problems with data integrity. (Batini et al. 2009)In addition, statistical methods should be robust and flexible enough to accommodate anomalies and non-standard distributions of data, which may indicate deeper problems. These strategies help to improve the credibility and reliability of statistical analyses and prevent a wide range of errors that could compromise data integrity.

0.5 Conclusion

This study shows that mistakes in collecting and organizing data can indeed mess up the data, leading us to make bad guesses. We tested the data that was messed up and found that the mean should be much higher than zero. But we can't trust this result because we know there were some glitches, such as some data being replaced and some confusion when we entered the numbers.

This tells us that we have to be extra careful when conducting research to catch and stop these kinds of errors in time. We need better tools for collecting data and checking for errors, and we also need to make sure that the people collecting the data know what they are doing and double check their work.

The tests we do also show why it's important to check to see how a slight change in the data here or there would change our findings. This helps us make sure that our findings are reliable and stand up even when things aren't perfect.

All in all, even though our data shows that the mean is much higher than zero, we have to be cautious due to the presence of errors. Making sure we catch and correct errors and use robust methods to analyze the data is especially important if we want to get to the truth and keep science on the right path.

##Peer review Dingning Li

You did a very detailed job of writing the steps clearly, but you need to do more discussion to show the importance of the experiment you did, and what it needs to be improved in being applied to real situations, and how we should avoid mistakes in our statistics. I believe adding these parts will make your essay better!

Reference

Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. "Methodologies for Data Quality Assessment and Improvement." *ACM Comput. Surv.* https://doi.org/10.1145/1541880.1541883.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Svanks, MI. 1988. "Integrity Analysis: Methods for Automating Data Quality Assurance." *Information and Software Technology* 30 (10): 595–605. https://doi.org/https://doi.org/10.1016/0950-5849(88)90116-4.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.