Understanding Normal Distribution with Simulation*

Hannah Yu

February 19, 2024

1 Introduction

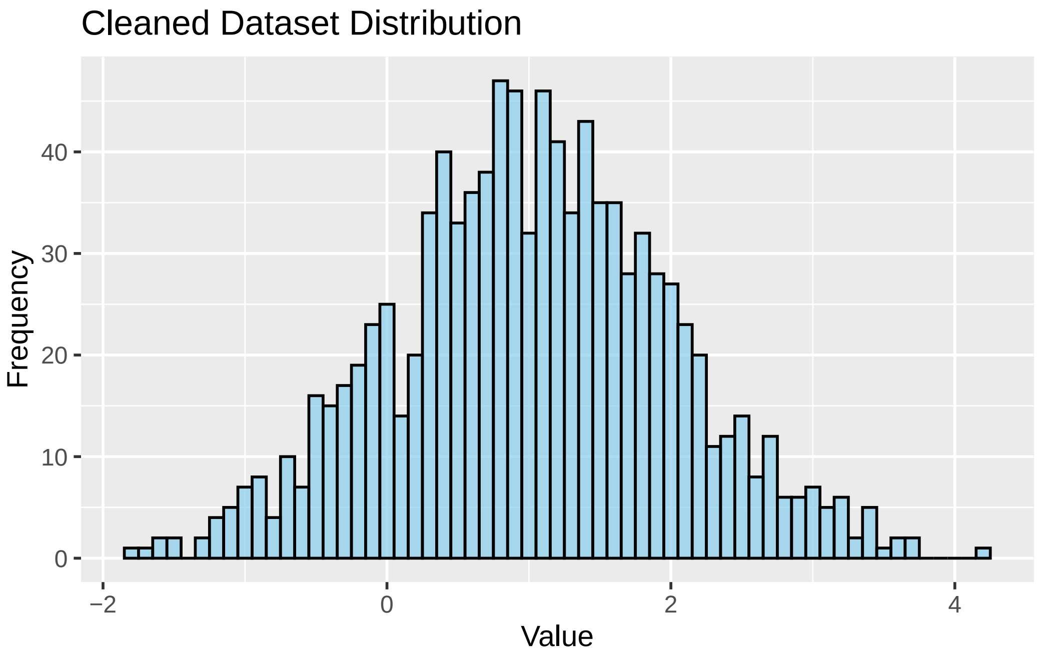
Accurate data collection and preparation are paramount for drawing meaningful conclusions from statistical analysis. However, various anomalies such as measurement errors, instrument limitations, and human mistakes can introduce biases and distortions into the data, potentially leading to erroneous interpretations. In this study, we simulate a scenario to examine the repercussions of multiple errors occurring during data collection and cleaning, with the objective of elucidating their impact on statistical inference.

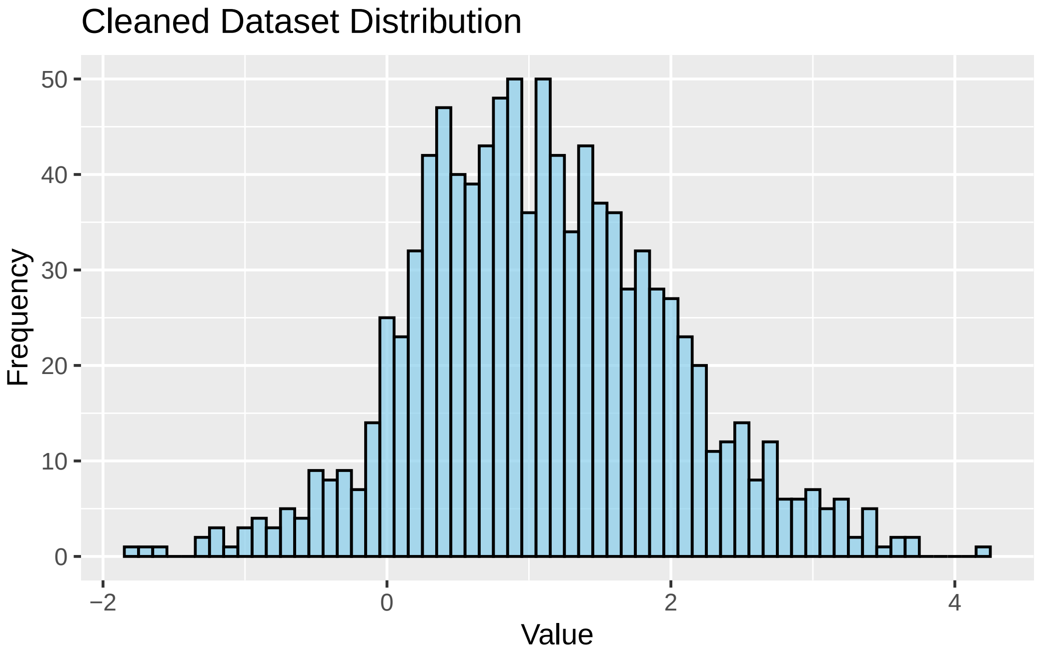
The prompt provides the following scenario: the data collection instrument has a limitation wherein it can only store up to 900 observations in memory, after which it begins overwriting data, resulting in the final 100 observations being duplicates of the first 100. Moreover, a research assistant is employed to clean adn prepare the dataset. While cleaning the data, teh assistant accidentally changed half of the negative draws to be positive. Additionally, errors introduced during data cleaning include the accidental conversion of half of the negative values to positive and a shift in the decimal place for values between 1and 1.1.

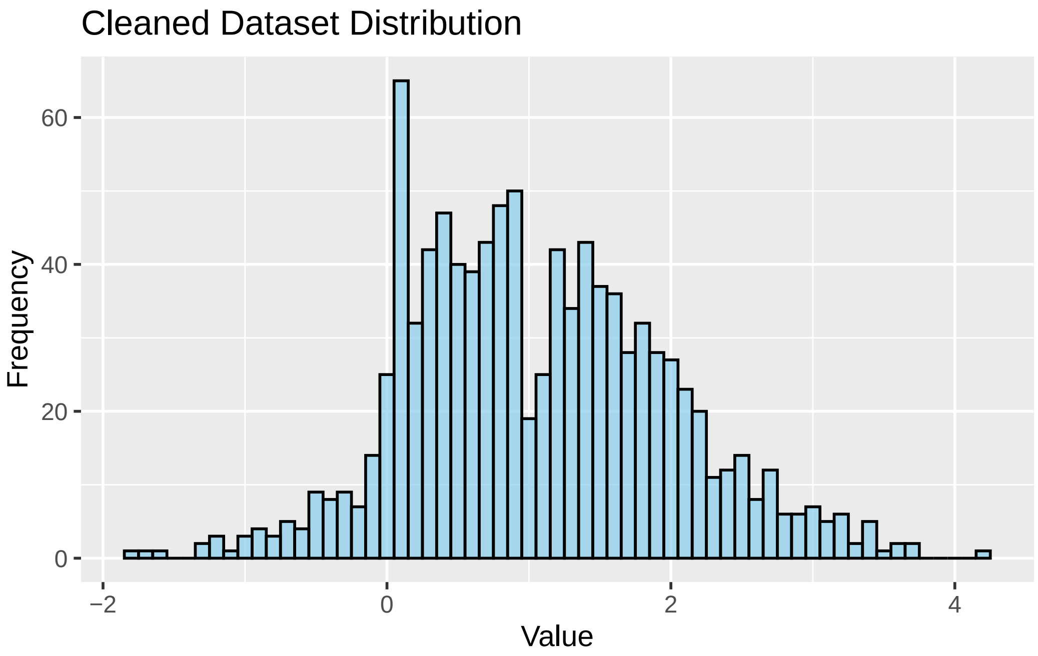
To initiate the simulation, we generated a dataset comprising 1,000 observations drawn from a Normal distribution with a mean of one and a standard deviation of one. This synthetic dataset serves as a representation of the true data generating process outlined in the prompt. And we simulate all the scenarios from the prompt using the dataset.

The analysis primarily utilized the R language and environment (R Core Team 2022), leveraging its robust statistical capabilities and extensive library ecosystem. Specifically, the ggplot2 package (**ggplot2?**) was employed for visualizing observations and trends following each stage of data manipulation.

[*] Code and data are available at: https://github.com/hannahyu07/Normal-Distribution-Simulation

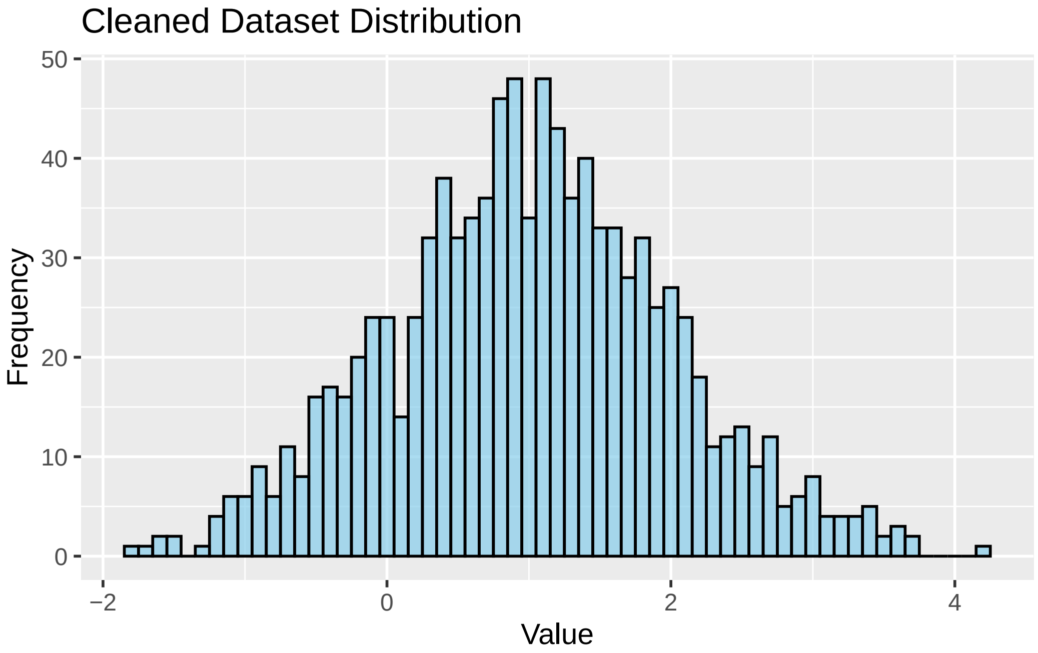

Cleaned Dataset Distribution
Frequency
0
10
20
30
40
−2
0
2
4
Value


Cleaned Dataset Distribution
Frequency
0
10
20
30
40
50
−2
0
2
4
Value



```
Mean of the cleaned dataset: 1.016128

Is the mean greater than 0? TRUE
```



R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.