Understanding Normal Distribution with Simulation*

Hannah Yu

February 26, 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.

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 and prepare the dataset. While cleaning the data, the 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 (Alexander 2023).

The analysis primarily utilized the R language and environment (R Core Team 2022). Specifically, we employed the ggplot2 package (Wickham 2016) to visualize observations and trends.

2 Set-up

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. From Table 1, we observe the summary statistics of the true dataset, with a minimum of -1.81,

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

Table 1: Summary Statistics of Clean Data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.81	0.37	1.01	1.02	1.66	4.24

Table 2: Summary Statistics of Situation 1 Data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.81	0.40	1.02	1.03	1.68	4.24

maximum of 4.24, median of 1.01, and mean of 1.02. And we simulate all the scenarios from the prompt using the dataset.

2.1 Situation 1

In this section, we examine the consequences of a memory limitation in the data collection instrument, which results in the duplication of observations towards the end of the dataset. Specifically, the dataset has a maximum memory of 900 observations, and overwrites the final 100 observations as a repeat of the first 100.

We observe the overall distribution of the dataset from Figure 1. Despite the distortion introduced by the duplication of observations, the overall shape of the distribution remains consistent with a normal distribution. While certain values may be more prevalent than others due to the duplication, the fundamental characteristics of the distribution remain unchanged. This observation is confirmed by Table 2, its summary statistics with a minimum of -1.51, a mean of 1.07, and a maximum of 4.24 still represents the shape of normal distribution.

2.2 Situation 2

In this segment, we address another common data cleaning error: accidentally converting negative values to positive ones. First, we identify the indices of negative values in the dataset. Subsequently, we randomly select half of these negative values and convert them to their positive counterparts.

The summary statistics of this simulation does not vary much compared to the previous simulation with the exception of a slight increase in median. The resulting distribution, visualized through histogram Figure 2 reflects the effects of this error on the dataset. While the overall shape of the distribution still resemble normal distribution, certain values may be inflated due to the conversion of negative values to positive. Notably, there's a slight increase in frequency

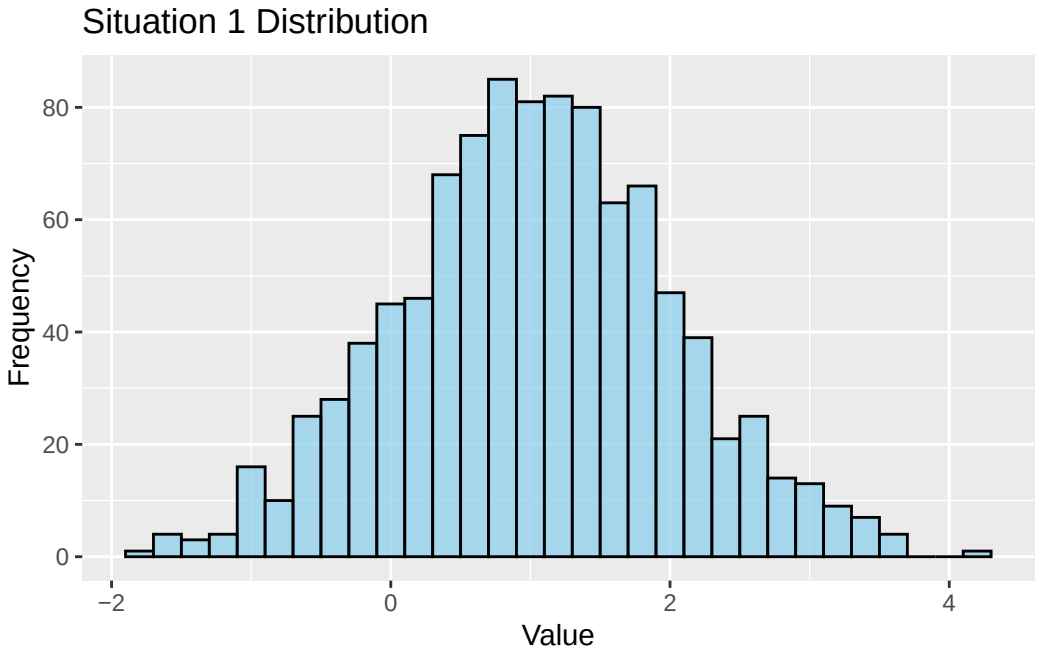

Figure 1: Situation 1 Distribution

Table 3: Summary Statistics of Situation 2 Data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.51	0.49	1.05	1.11	1.69	4.24

for all values above 0, while the frequency of negative numbers is halved. It's mean of 1.11 shows that there is a increase in the number of positive values.

2.3 Situation 3

In this section, we simulate yet another data cleaning error: the inadvertent alteration of decimal places in values falling between 1 and 1.1. We identify these values within the dataset and divide them by 10, effectively shifting their decimal place by one position to the left. This manipulation mimics a scenario where data cleaning procedures introduce unintended changes to the dataset.

The resulting distribution, visualized through histogram Figure 3, demonstrates the effects of this error on the dataset. While the overall shape of the distribution may still exhibit characteristics of a normal distribution, the alteration of decimal places can lead to discrepancies in

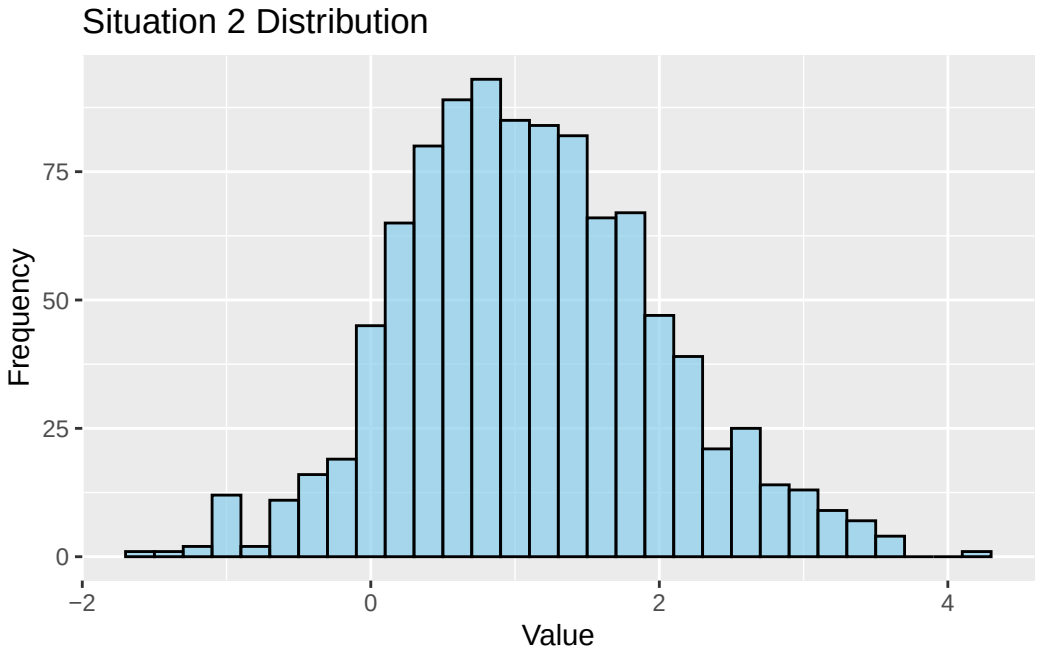

Figure 2: Situation 2 Distribution

Table 4: Summary Statistics of Situation 3 Data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.51	0.40	0.94	1.07	1.69	4.24

the frequency distribution. Notably, values between 1 and 1.1 exhibit a decrease in frequency, while values between 0 and 0.11 experience a corresponding increase in frequency.

2.4 Situation 4

In the previous section, we addressed various data cleaning errors and their effects on the dataset. To conclude, we restored the dataset by reverting to the original true data and visualized the corrected cleaned dataset in histogram Figure 4.

Upon examination, we observe a distribution consistent with a normal distribution. This distribution comprises 1000 observations, centered around a mean of 1 and with a standard deviation of 1, aligning precisely with the parameters specified in the original true data generating process.

The restoration of the dataset to its original form underscores the importance of thorough validation and scrutiny during the data cleaning process.

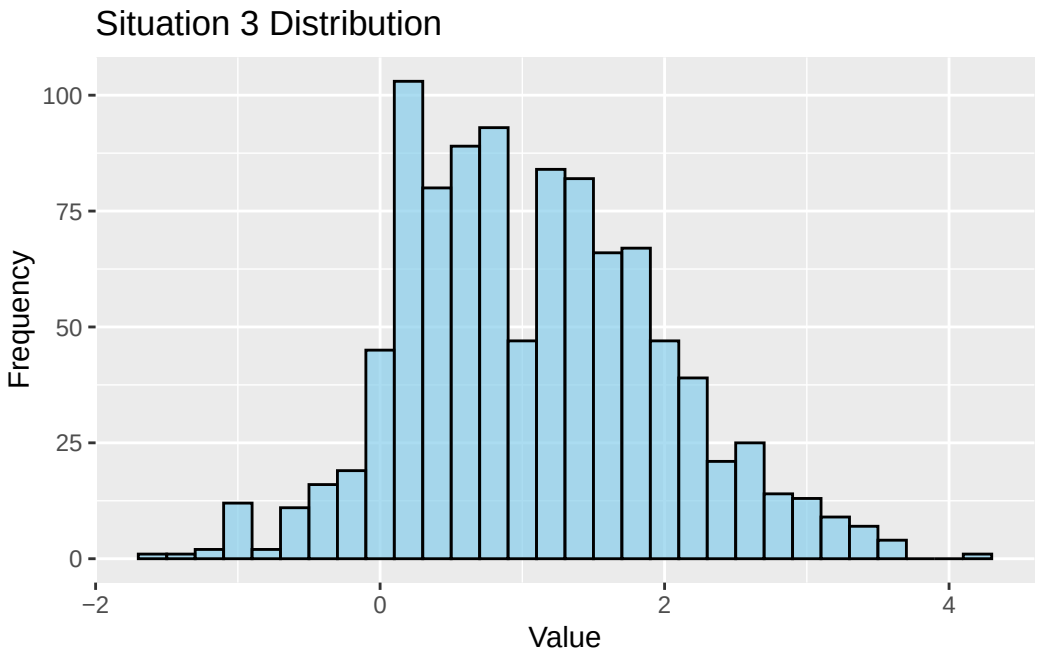

Figure 3: Situation 3 Distribution

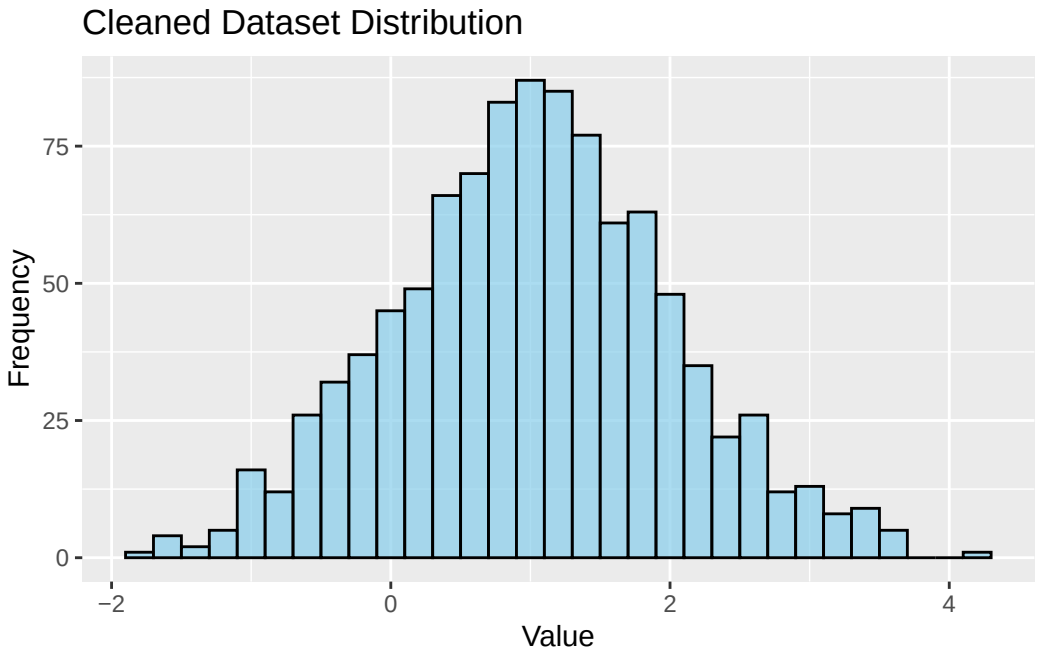

Figure 4: Situation 4-Cleaned Dataset Distribution

3 Discussion

The errors introduced during the data collection and cleaning processes have significant implications for the integrity of the dataset and subsequent analyses. Each error resulted in distortions to the distribution of the dataset, potentially biasing statistical inferences and leading to erroneous conclusions. Just like Van den Broeck et al. (2005) emphasized the importance to address errors in data cleaning to prevent false study conclusions.

In Situation 1, where the data collection instrument had a memory limitation causing the duplication of observations, we observed a distortion in the dataset's distribution. While the overall shape remained similar to a normal distribution, certain values were overrepresented due to the duplication. This highlights the importance of validating data collection instruments and implementing measures to prevent memory limitations from compromising data integrity.

Similarly, in Situation 2, the inadvertent conversion of negative values to positive ones introduced biases into the dataset. Although the overall shape still resembled a normal distribution, the frequency of values above 0 increased while the frequency of negative values decreased. To prevent such errors, robust data cleaning protocols should be established, including thorough checks for unintended alterations to the dataset.

In Situation 3, where the decimal places of values between 1 and 1.1 were altered, we observed further distortions in the dataset's distribution. While the overall shape remained similar to a normal distribution, the alteration of decimal places led to discrepancies in the frequency distribution. Values between 1 and 1.1 experienced a decrease in frequency, while values between 0 and 0.11 saw a corresponding increase. Implementing quality control measures during data cleaning can help mitigate such errors.

To ensure that similar issues are flagged during actual analyses, several steps can be implemented. Firstly, conducting thorough data validation checks before and after data cleaning processes can help identify inconsistencies or anomalies (Chu 2019). Additionally, implementing automated data quality checks and validation scripts can assist in detecting errors early in the analysis pipeline. Furthermore, establishing clear documentation and communication channels among team members involved in data collection and analysis can facilitate the identification and resolution of potential issues.

In conclusion, the errors introduced during data collection and cleaning processes can have profound effects on the integrity of the dataset and subsequent analyses. By implementing robust quality control measures and validation checks, researchers can mitigate the risk of errors and ensure the reliability of their findings.

References

Alexander, Rohan. 2023. "Telling Stories with Data." *Tellingstorieswithdata.com*. https://tellingstorieswithdata.com/.

Chu, Xu. 2019. "Data Cleaning." *Springer eBooks*, January, 535–41. https://doi.org/https://doi.org/10.1007/978-3-319-77525-8_3.

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

Van den Broeck, Jan, Solveig Argeseanu Cunningham, Roger Eeckels, and Kobus Herbst. 2005. "Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities." *PLoS Med* 2 (10): e267. https://doi.org/10.1371/journal.pmed.0020267.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.