Errors in Cleaning Data*

Shreya Sakura Noskor

February 25, 2024

1 Introduction

Whne taking measurements of data in the real world, many statisticians face the difficulties of cleaning raw data. This may be due to incorrect or missing values in the data or due to the fact that some data needs to be combined in order to perform meaningful analysis. For example it is impossible to take in to consideration the opinions of the entire population of Toronto but we can gather a smaple and clean the data so that they best represent the population.

This paper covers multiple scenarios where errors could happen when taking measurements or cleaning the data. We stimulate this by generating a set using normal distribution and seeing what would happen if: the measurement was wrong, the numbers got changed during the process by being turned into positives or turned into decimals. We also look at the mean for each of these scenarios and graph the normal distribution better to compare them. We see that all the cleaning procedures do give us a better mean but not a better normal distribution compared to the raw/cleaned data.

The main data we will be looking at is stimulated by a normal distribution whose first couple of values are presented below in Table 1.

Table 1: The Raw Data

x
-0.0944949
2.0961629
-0.3558643
-0.0384344
0.9677574
-0.5694278

*Code and some data from this paper are available at: github repo

2 The Errors

This paper was created using R (R Core Team (2022)) as well as using the help from other packages like Wickham (2016), Wickham et al. (2019), Wickham et al. (2023), and Xie (2014).

2.1 Situation 1

```
[1] 0.946091
```

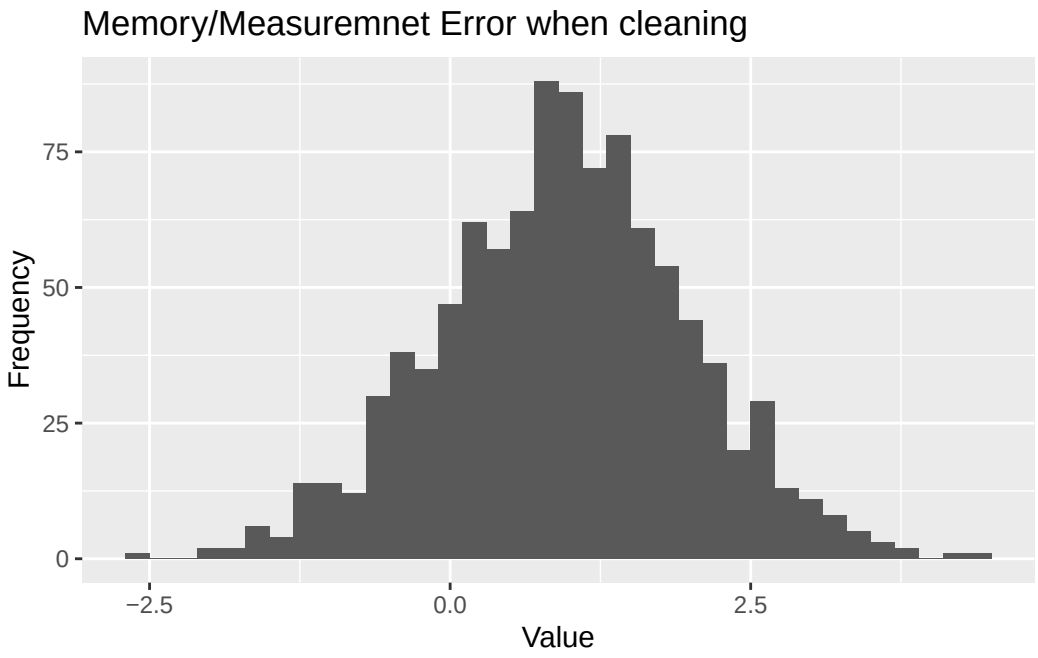

Figure 1: Memory/Measuremnet Error when cleaning

The first scenario that may lead to an error in cleaning is when we are taking the actual measurement. In this case what happens is that there is an error when try to gather 1000 data points where the amount of data an equipment can hold is 900. so the last 100 data points is actually the repetition of the first 100. In this graph Figure 1 we see that the graph does follow normal distribution, visually but the values that are repeated are stored in the same bin meaning that the data is now bias and if not it is inaccurate. Upon a closer investigation at the output of the cleaning we notice that the first 100 elements are positive and hense the historgram is now skewed to the right.

2.2 Situation 2

```
[1] 1.055946
```

Figure 2: Handleing Negative Error when cleaning

Figure 2 looks at the case when there is a human error when handling the data. In this case we stimulate the scenario where half of negative values in the original data is turned into a positive. We again see that the distribution is less normal now with more values in the 0 to 1 bins. This means there are more positive numbers then the original data hence it is more left skewed data and it is also inaccurate compared to the original raw data. In this stimulation we had to randomly choose the negative indices that we would turn positive as picking what values to turn negative would not have showcased the entire effect of this stimulation.

2.3 Situation 3

```
[1] 0.9042228
```

The third scenario that we stimulate in this graph (Figure 3), is the case when the values get changed again but this time it may seem small. Here we stimulate by filtering the original data for values between 1 to 1.1 and divide by 10 so that those values are now from 0.1 to 0.11 respectively. Initially it may seem like a very small change has been done to a very small

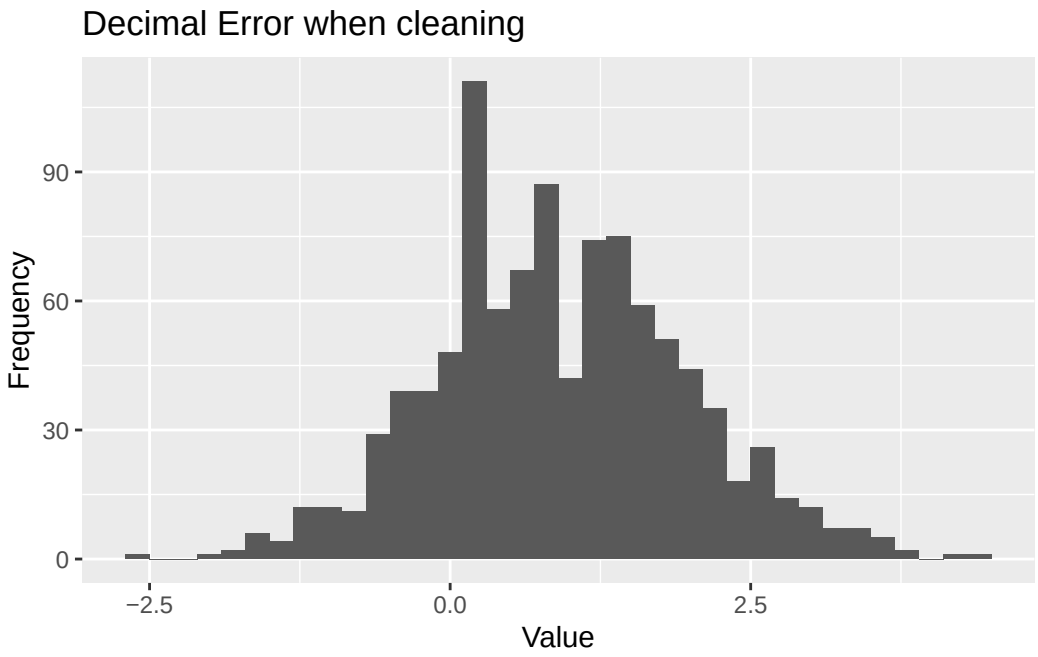

Figure 3: Decimal Error when cleaning

subset of the whole data but from Figure 3, we see that this makes the graph very uneven. This means that there is one bin where it contains the values from 1 to 1.1 and the 0.1 to 0.11 values making it the largest bin in the entire graph. This leads to a very left skewed graph as our stimulated random normal distribution had a mean of 1 and standard deviation of 1 as well. In this stimulation we had to make sure that we only took a small subset of the data to really show the amount of error it would cause.

3 The Correct Cleaned Set

```
[1] 0.951202
```

In this graph Figure 4, we just stimulate the data with the `rnorm()` function because there is nothing we need to clean in this stimulation. What we see is a mostly normal distribution with a mean of 0.951202 which is very accurate because when calling the rnorm function we have mean set to 1. There are no major outliars in this data set that throws the distribution off. Having a mean above 0 is necessary as if we had a mean less then that it mean that our data is the opposite of the raw data.

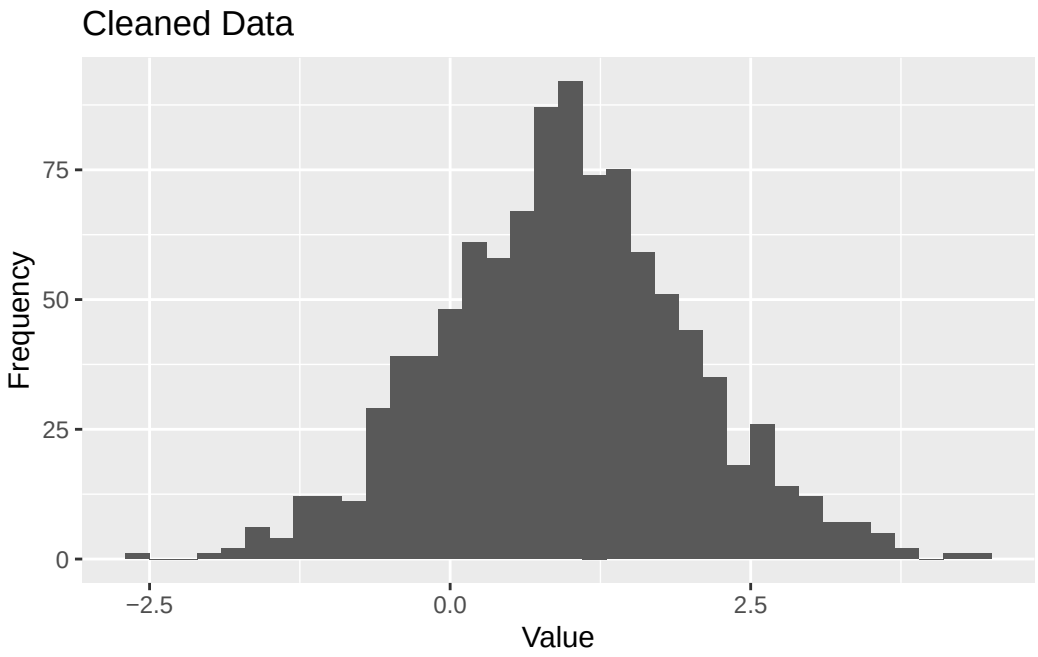

Figure 4: The Distirbution for Raw data

4 Discussion

4.1 Summary

In this paper we saw the effects of wrong methods in data cleaning by using stimulation's. We had 3 error cases: mistaking in equipment leading to memory error, mistaking in cleaning leading to change in signs of data, and mistaking in cleaning leading to change in data itself. In Figure 1, we see that the effect of having data repeating itself is dangerous as it can lead to inaccurate data. The way we can fix is this by making sure our equipment is up to date and can handle the sample size. If not we can go back and check the data to see if there are any obvious repeats, although in some cases that may not work. In Figure 2 we see that converting some negative values in to positive values can skew the data. Also logically thinking, this makes the data inaccurate as well since the data has been tampered with and is now fake. A way to fix this error is to set up systems that doesn't allow anyone to actually change the raw data and make sure certain rules are set in place to diminish such errors. In Figure 3 we see the similar thing as changing the negative values where some bins now carry more data then they originally did. This lead to a slight left skew in the normal distribution. Similar to the negative stimulation the solution to this is to just set protocols in place that do not allow anyone to change the actual raw file without consulting multiple people.

4.2 Compared to Clean Data

We see that the mean to each scenario is 0.946091 Figure 1, 1.0559458 Figure 2, 0.9042228 Figure 3 and lastly the actual mean from the original raw data, 0.951202. We see that in each scenario the mean from all the stimulation is very close to 1 but the closets is the repeated data scenario and the original data. The other 2 scenarios have a much higher difference with the mean due to the fact that they changed the actual values of the data so they got farther away. The reason why the repeated scenario sometimes has a much closer mean to 1, as opposed to the original raw data stimulated by using `rnorm()` with mean 1, is due to the fact that it has repeated data. What happened was that the first 100 values may have been close to 1 so when we had repeats of it we essentially increased the weight of those values. The problem with this case is that the weight goes to the first 100 values of the data so if the first 100 data points were outliars it would change the normal distribution and the mean by a lot. This is very unreliable.

4.3 Concluding thoughts

Data cleaning may seem like a very simple step but it is very crucial that we are careful with it as one change may lead to a very different result and outcome. There for having simple protocols in place or working with multiple people when cleaning data is a very good idea to pipeline the mistakes so they can get caught early on rather, then later.

References

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

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

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. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.