STA 302 Tutorial 7*

Diana Liu

February 22, 2024

Introduction

In this exercise, we simulate errors in data collection and cleaning using R Core Team (2021) and discuss how they can effect the estimated mean of a dataset in the data section and potential ways to catch errors. The data set with errors has mean of 0.99, standard deviation of 1, and error of 0.03 (Table 5). Its probability density function and histogram seems to have binomial distribution with a peak between 0 and 1 and another between 1 and 2 (Figure 1) (Figure 2). Mean and standard deviation from the data set with errors do still suggest that true mean is greater than zero.

In general, errors can be detected by plotting the data set. How the graph differs from expectations can alert us to the fact that there could have been errors in the collection and cleaning of data. If we know that certain errors can occur, we can build functions that detects them, for example checking the frequency of negative observations.

Data

First we simulate a data set of 900 observations with normal distribution, mean of 1 and standard deviation of 1. In situation one, the last 100 observations of the data set has been replaced with the first 100. This is simulated by storing the first 100 observations in a variable then adding them to the end of the data set. Now the data set has 1000 observations and the first 100 are identical to the last 100 (Table 1).

In situation two half of the negative numbers in the data set are changed to positives. First we count the number of negative numbers and divide it by two to get the amount of negatives that has to be changed. Then we iterate through the data set and change negatives to positives by

[*] Code and data are available at: https://github.com/Diana-Guanzhi-Liu/STA-302-Tutorial-7. Peer Review by Hannah Yu: https://github.com/Diana-Guanzhi-Liu/STA-302-Tutorial-7/issues/1

Table 1: Comparing the first the first 10 observations to the 901 to 910 observations, we find that they are identical due to over-writing by the instrument

<table>
<tr><th>x</th><th>x</th></tr>
<tr><td>| head|</td><td>| tail|</td></tr>
<tr><td>|———-:|</td><td>|———-:|</td></tr>
<tr><td>| 0.4395244|</td><td>| 0.4395244|</td></tr>
<tr><td>| 0.7698225|</td><td>| 0.7698225|</td></tr>
<tr><td>| 2.5587083|</td><td>| 2.5587083|</td></tr>
<tr><td>| 1.0705084|</td><td>| 1.0705084|</td></tr>
<tr><td>| 1.1292877|</td><td>| 1.1292877|</td></tr>
<tr><td>| 2.7150650|</td><td>| 2.7150650|</td></tr>
<tr><td>| 1.4609162|</td><td>| 1.4609162|</td></tr>
<tr><td>| -0.2650612|</td><td>| -0.2650612|</td></tr>
<tr><td>| 0.3131471|</td><td>| 0.3131471|</td></tr>
<tr><td>| 0.5543380|</td><td>| 0.5543380|</td></tr>
</table>

multiplying by negative one until half of negatives have been changed (Table 2). If we scroll through the data set, negative numbers only being to appear after 500 observations, which is consistent with turning the first half of them positive.

In situation three, the decimal place of any value between 1 and 1.1 is shifter to the left by one digit. To do this we iterate through the data set and check if the observation is between 1 and 1.1, if true, multiply it by 0.1 (Table 3).

Discussion

Now the mean, standard deviation, and standard error from the simulated data set with errors can be compared to a simulated data set without errors. We know that the true mean and standard deviation would be one, and the error free data set has mean of 1.03 and standard deviation of 1 which are quite close (Table 4). Its density function looks to be perfectly normally distributed with mean of one (Figure 1). The data set with errors has mean of 0.99, standard deviation of 1, and error of 0.03 (Table 5). It seems to have binomial distribution with a peak between 0 and 1 and another between 1 and 2 (Figure 1). Mean and standard deviation from the data set with errors do still suggest that true mean is greater than zero.

In general, these errors can be detected by plotting the data set in a probability density graph. We expect the plot to be normally distributed centered around one, and we have enough draws according to the Law of Large Numbers to get the correct shape (Wasserman 2005). Because the graph is not the correct shape, this alerts us to the fact that there could have been errors in the collection and cleaning of data.

Table 2: Comparing the first 10 observations after the change to before, notice that the negative observation has been turned positive

<table>
<tr><th>x</th><th>x</th></tr>
<tr><td>| head|</td><td>| head_no_negatives|</td></tr>
<tr><td>|———-:|</td><td>|—————:|</td></tr>
<tr><td>| 0.4395244|</td><td>| 0.4395244|</td></tr>
<tr><td>| 0.7698225|</td><td>| 0.7698225|</td></tr>
<tr><td>| 2.5587083|</td><td>| 2.5587083|</td></tr>
<tr><td>| 1.0705084|</td><td>| 1.0705084|</td></tr>
<tr><td>| 1.1292877|</td><td>| 1.1292877|</td></tr>
<tr><td>| 2.7150650|</td><td>| 2.7150650|</td></tr>
<tr><td>| 1.4609162|</td><td>| 1.4609162|</td></tr>
<tr><td>| -0.2650612|</td><td>| 0.2650612|</td></tr>
<tr><td>| 0.3131471|</td><td>| 0.3131471|</td></tr>
<tr><td>| 0.5543380|</td><td>| 0.5543380|</td></tr>
</table>

Table 3: Comparing the first 10 observations from after this change to when the data set was first simulated, decimals for observations between 1 and 1.1 have been shifted to the left

<table>
<tr><th>x</th><th>x</th></tr>
<tr><td>| head|</td><td>| head_smaller_decimals|</td></tr>
<tr><td>|———-:|</td><td>|———————:|</td></tr>
<tr><td>| 0.4395244|</td><td>| 0.4395244|</td></tr>
<tr><td>| 0.7698225|</td><td>| 0.7698225|</td></tr>
<tr><td>| 2.5587083|</td><td>| 2.5587083|</td></tr>
<tr><td>| 1.0705084|</td><td>| 0.1070508|</td></tr>
<tr><td>| 1.1292877|</td><td>| 1.1292877|</td></tr>
<tr><td>| 2.7150650|</td><td>| 2.7150650|</td></tr>
<tr><td>| 1.4609162|</td><td>| 1.4609162|</td></tr>
<tr><td>| -0.2650612|</td><td>| 0.2650612|</td></tr>
<tr><td>| 0.3131471|</td><td>| 0.3131471|</td></tr>
<tr><td>| 0.5543380|</td><td>| 0.5543380|</td></tr>
</table>

Situation 1

The change in situation 1 caused the first 100 observations to be the same as the last 100 so there are only 900 unique observations compared to 1000 for the error free data set. This is unlikely to significantly effect the mean & standard deviation and distribution because the observations are randomly generated and the number of observations were not significantly effected.

This mistake is difficult to flag during analysis as it is difficult to tell if a pattern that exists in a data set is due to a mistake or not. If we are aware of the mistake in the instrument, we can manually review the data or build a function that compares the first 100 observations to the last 100 to detect if over writing occurred.

Situation 2

Situation 2 caused half of negative observations to become positive. This will increase the mean of the data set with errors. This is likely what creates the peak around 1.25 in the density plot and histogram (Figure 1) (Figure 2) as negative values on the left half of the graph are turned positive and moved to the right half. We expect observations to be normally distributed around 1 with standard deviation of 1 so most observations are between 0 and 2. This means that situation 2 is unlikely to effect standard deviation as any negative numbers that are converted to positive numbers influence standard deviation by the same amount.

This error can be detected by checking the frequency of negative values. We would expect the simulation to generate negative values that are evenly spaced throughout the observations. With a function that detects the frequency of negative observations of just by manually scrolling through the data set, we will be able to flag this error and similar errors.

Situation 3

Situation 3 decreased the mean by decreasing observations from 1 to 1.1 by a factor of ten. this is likely why there is a peak around 0.75 as those values are shifted left. This likely also decreases standard deviation as effects of observations that are closer to the mean are smaller than that of observations that are larger.

This error can be detected in a similar way as negative numbers by looking at the distribution of all numbers and noticing that there are no observations between 1 and 1.1 or seeing that the histogram bin that should contain observations between 1 and 1.1 has far less observations than what we expect (Figure 2).

Table 4: Error free data set has mean of 1.03, standard deviation of 1, and standard error of 0.03

True mean	True standard deviation	True standard error
1.03	1	0.03

Table 5: Data set with errors has mean of 0.11, standard deviation of 0.09, and standard error of 0

Estimated mean	Estimated standard deviation	Estimated standard error
1.06	0.92	0.03

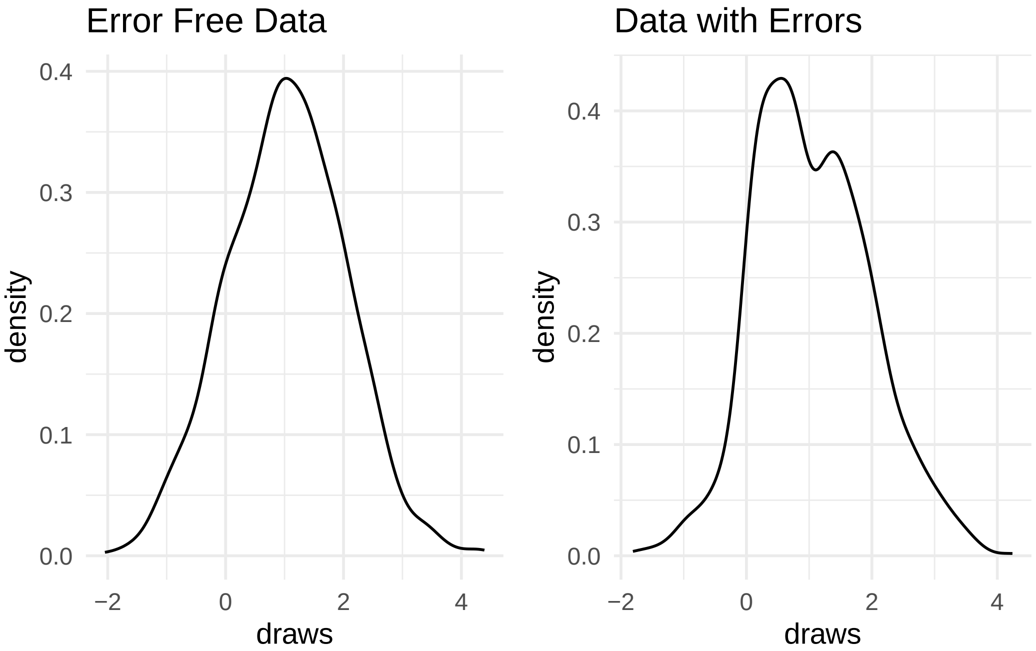

Figure 1: Error free data is normally distributed and has mean of one. Data with errors seems to be binomially distributed with one peak at 0.75 and another at 1.25.

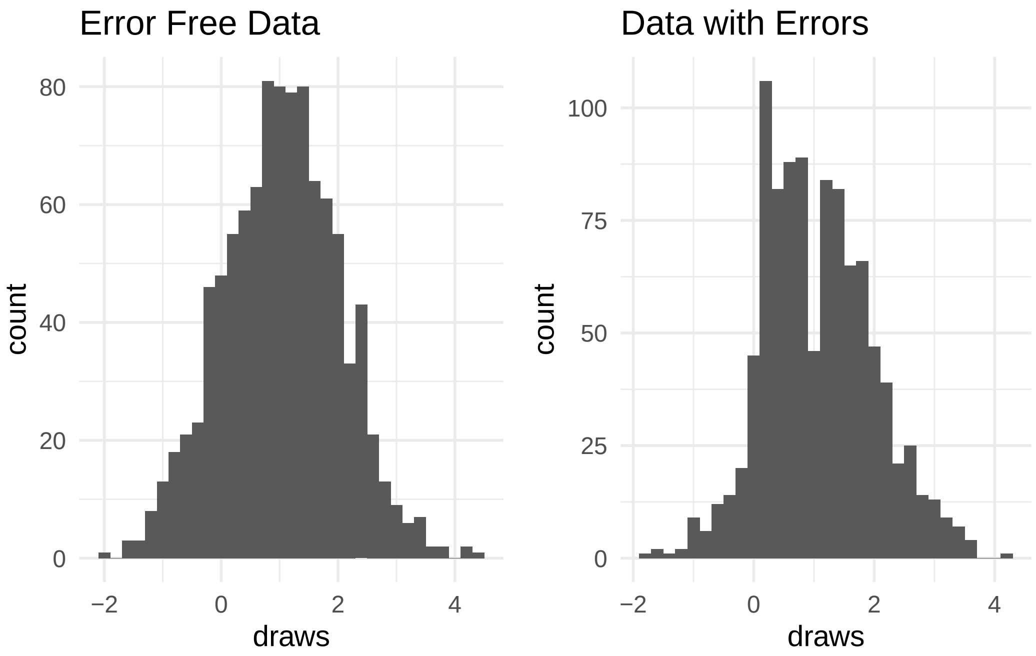

Figure 2: Error free data is normally distributed with mean of one. Data with errors is binomially distributed with a peak between 0 and 1 and another between 1 and 2

References

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wasserman, Larry. 2005. "All of Statistics." *Springer*.