STA 302 Tutorial 7*

Diana Liu

February 22, 2024

Introduction

In this exercise, we simulate how errors in data collection and cleaning can effect the estimated mean of a dataset and discuss potential ways to catch errors.

Data

First we simulate a data set of 900 observations with normal distribution, mean of 1 and standard deviation of 1. In situation 1, 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).

Then 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 multiplying by -1 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.

Now we change the decimal place of any value between 1 and 1.1. 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, shifting the decimal to the left (Table 3).

*Code and data are available at:

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

x	x
head	tail
---:	---:
0.4395244	0.4395244
0.7698225	0.7698225
2.5587083	2.5587083
1.0705084	1.0705084
1.1292877	1.1292877
2.7150650	2.7150650
1.4609162	1.4609162
-0.2650612	-0.2650612
0.3131471	0.3131471
0.5543380	0.5543380

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

x	x
head	head_no_negatives
---:	---:
0.4395244	0.4395244
0.7698225	0.7698225
2.5587083	2.5587083
1.0705084	1.0705084
1.1292877	1.1292877
2.7150650	2.7150650
1.4609162	1.4609162
-0.2650612	0.2650612
0.3131471	0.3131471
0.5543380	0.5543380

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

x	x
head	head_smaller_decimals
------:	------:
0.4395244	0.4395244
0.7698225	0.7698225
2.5587083	2.5587083
1.0705084	0.1070508
1.1292877	1.1292877
2.7150650	2.7150650
1.4609162	1.4609162
-0.2650612	0.2650612
0.3131471	0.3131471
0.5543380	0.5543380

Discussion

Now the mean, standard deviation, and standard error from the simulated data set with errors can be comared 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).

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.

Situation 2 caused half of negative observations to become positive. This will increase the mean of the data set with errors. 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 sd by the same amount.

Situation 3 decreased the mean by decreasing observations from 1 to 1.1 by a factor of ten. this is likely why the estimated mean is smaller than the true mean by approximately a factor of ten. This also decreases standard deviation as observations between 1 and 1.1, on the right hand side of the distribution are decreased

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

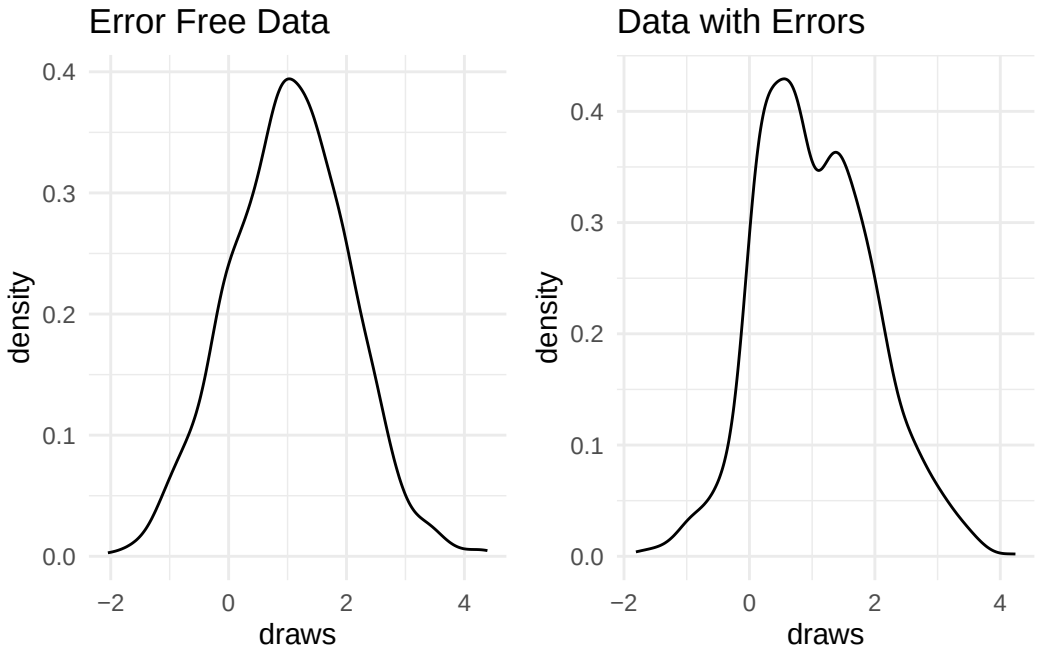

Figure 1