Mus Musculus Statistics Challenge

Darren Liu

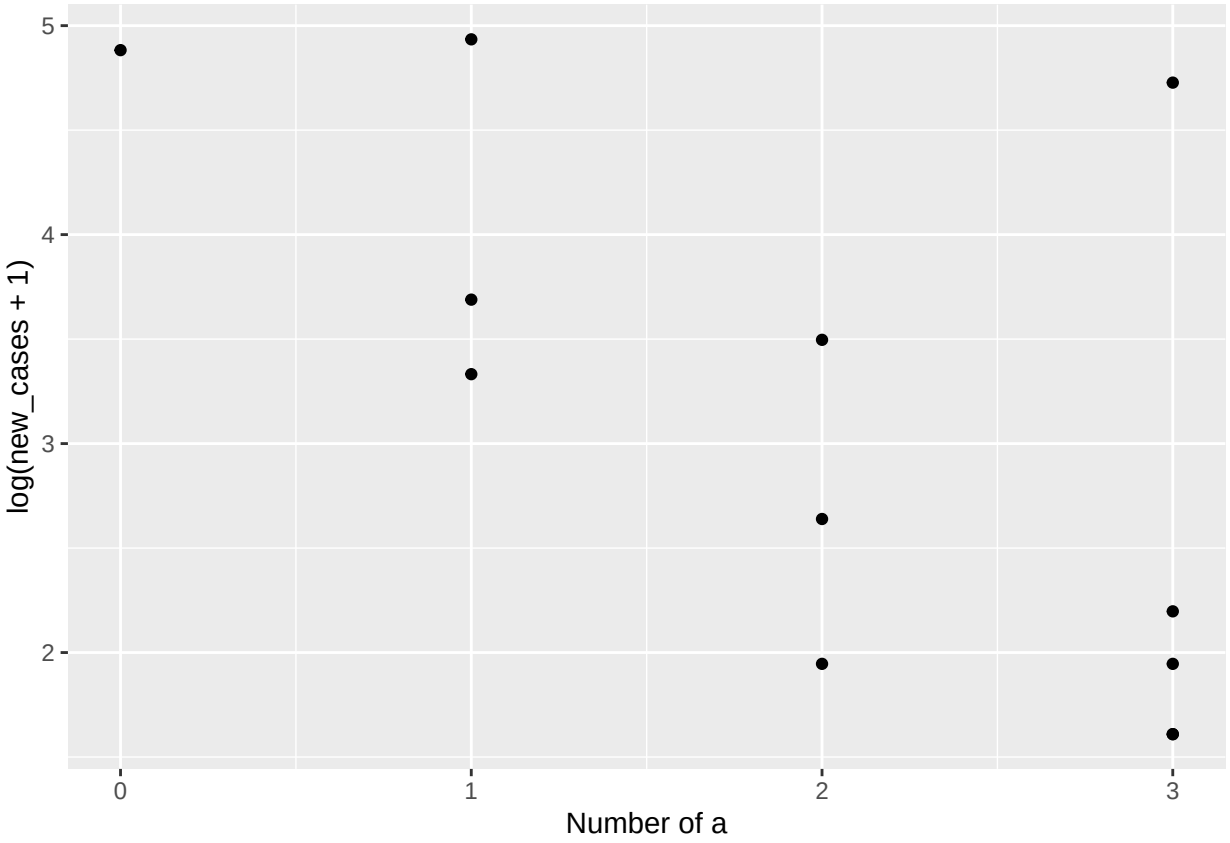
2022-09-12

Our approach

To begin, we first selected only for locations in North America on 03/28/2020 since North America is the most interesting location and March 28th is the most interesting day.We then found the amount of times the letter 'a' appeared in each location. To better understand our data, we then took the natural logarithm of the amount of new cases (1 was added in order to prevent log(0)). Finally, all outliers were tactfully removed (rows with $1.5 < \log(\text{new\_cases} + 1) < 5$ were omitted). We then took the Pearson coefficient between the count of the letter 'a' in the location as our indenpendent variable and the natural log of the new cases for that given location as our dependent variable. We found there to be a strong correlation between the number of 'a's and the amount of new COVID cases with a p-value = 0.02014

*Editor's note: This is entire paragraph is **very** sarcastic

```
##
##  Pearson's product-moment correlation
##
## data:  df$num_a and df$lnew_cases
## t = -2.7597, df = 10, p-value = 0.02014
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8940504 -0.1342956
## sample estimates:
##       cor
## -0.6575205
```



Statistical Criticism

To start with, it should be fairly obvious that there is no true correlation between the amount of 'a's in a location and the amount of COVID infections on any particular day. In order to get the 'significant' result, the data was manipulated in several different ways. For instance, the selection of only North America and the date (03/28/2020) is completely arbitrary and was largely done to reduce the sample size. From this selection, the data was graphed and there was a very weak negative correlation. In order to gain a significant result, points that did not follow this negative correlation was classified as 'outliers' and removed. The cherry-picking of the data allowed us to reduce the sample size (N = 11) to a point where we could find a spurious correlation.