Fraudulent Data Detection Presentation

Cooper Strahan

November 2020

1 Introduction

My talk on Detecting Fraudulent Data was given on Friday the 13th, after Kemal and Andrew's Discussion on Intrusion Detection Systems. The talk was composed of three main sections being Cases of Data Fraud in Research, describing Benford's Law, and the application of Benford's Law. The medium that the presentation was built on was a Google Slide document.

2 Purpose

The talk began with a description of two cases of research fraud. Bengü Sezen's falsifying Nuclear Magnetic Resonance data, and David Anderson replacing the values of data points that were outliers with the mean value. Bengü Sezen's fraud caused the termination of three graduate students two of which who were fired over their inability to recreate her results, and one who dropped out of the program for the same reason. David Anderson's fraud caused 4 papers to be redacted from publication. Clearly the first instance of fraud directly affected more people, but both are ethically problematic.
After describing these instances of fraud I spoke about Benford's law. Benford's law is that in many sets of naturally occurring numbers, that the leading digit is likely to be small (Wikipedia Benford's Law). Next the history of Benford's Law was discussed, stating that it was originally found by Simon Newcomb in 1881, but that the extensive amount of data tested by Benford did a better job to display the law's use and gained more popularity for its discovery. The law was then named after Frank Benford rather than Simon Newcomb.
I then went on to describe the math surrounding Benford's Law. The idea that the distribution of the probability of the first digit where the digit $d \in 1, 2, ...9$ follows the equation $log_{10}(1 + \frac{1}{d})$. This is followed by the probability of the second digit which is described by the equation $\sum_{k=10^{n-2}}^{10^{n-1}-1} log_{10}(1 + \frac{1}{10k+d})$. The formula's results were then shown in a table that displayed the probability of the first, second and third digits. The first digits were shown to decrease exponentially from 1 to 9 where the probability of 1 was 30% and then probability of 9 was just under 5%. The probabilities of the second and third numbers were shown to be around 10%.

We then went over data sets that tend to follow the distribution and those that do not. Examples of the former were data sets with a mean larger than the median, that had a positive skew and sets that were composed of numbers that were generated with mathematical combinations. Examples of the latter were sequentially assigned numbers, numbers influenced by human thought, numbers with a built in minimum/maximum, and distributions that didn't span magnitudes. Following that we spoke about types of real-world data sets that follow the law, like COVID-19 data, and other fields of research where the data would conform to the law like accounting data, economic data, genome data, and scientific data. The last part of this section discussed data that does not follow the law, for instance, square roots and reciprocals, telephone directories, height, weight, and IQ data.

The last section of the talk spoke about how Andreas Diekmann performed a study surrounding the phenomenon of this law in data. He compared the distribution of the first and second digit of real and fraudulent data to the distributions described by Benford's Law. I then showed his results, that the real data followed the distribution almost perfectly and the constructed fraudulent data for the first digit mostly followed the distribution but fell apart when compared to the second digits distribution.

3 Conclusion

I wrapped up the talk by discussing the implications of Diekmann's paper. The idea that large quantities of grant money are going to research, and that the idea that the standard for proof in research is only the ability of recreation, and that there should be a standard in which data needs to be vetted. The combination of Benford's Law and an Intrusion Detection System would allow researchers to be able to better detect fraudulent research data, and an improvement of said fraud detection system would vastly improve the reliability of research.