**Benford’s Law**

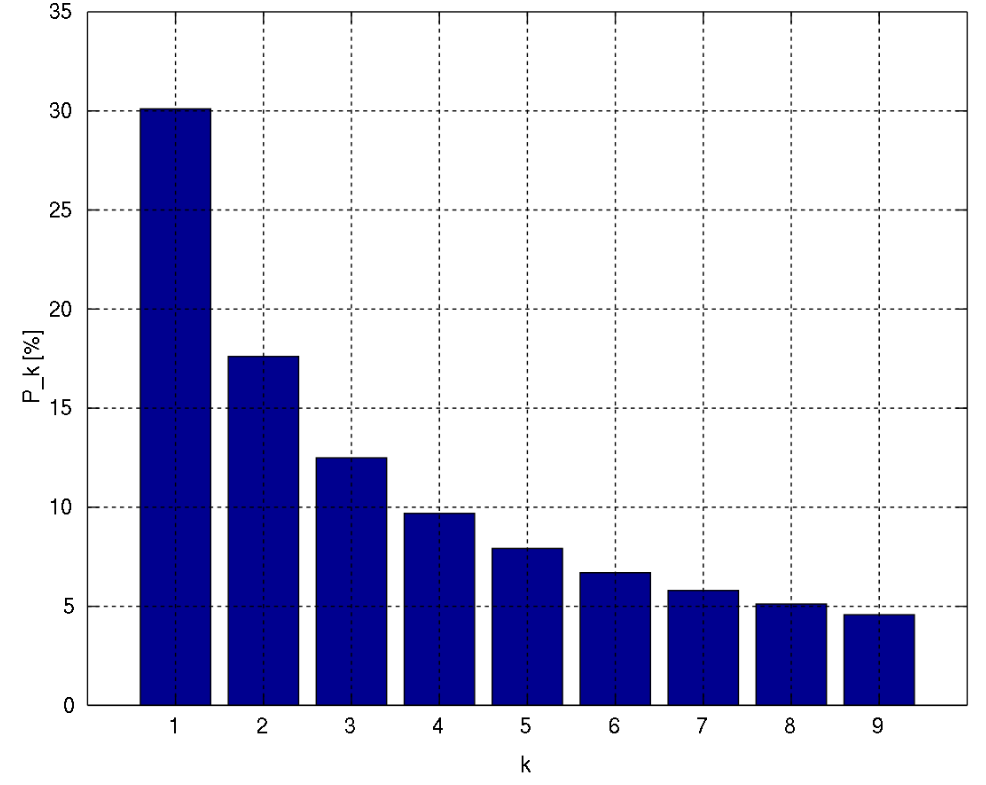
**Nick Miller**

**Bellevue University**

**Summer 2020**

**Abstract**

I recently saw a TV show that featured Benford’s Law and decided to perform a project about it. I am interested in all things math related, and will hopefully be able to change careers after this degree program is completed into something math related. Benford’s Law is the widely regarded principle that the first digit in large collections of natural numbers have certain rates of occurrence. The chart below outlines the target rates of each digit. This project is going to test this theory in two modern segments.



**Introduction**

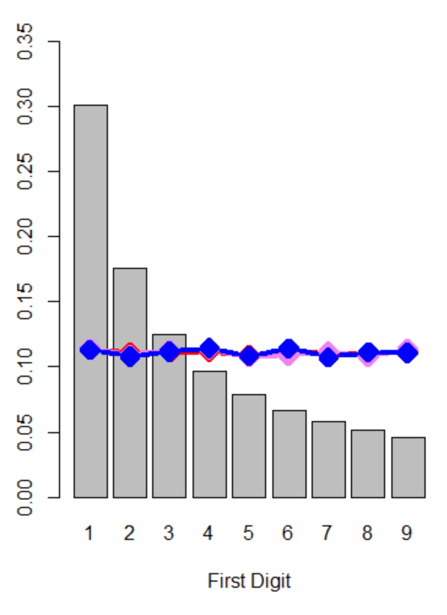
In order to test this theory, I wanted to examine 3 sets of data. One set would be a random set of numbers. The second set would be various statistics about YouTube videos, including the amount of views, likes, dislikes, and comments of each video in the dataset. The third dataset is video game sales in North America, Europe, and Japan. I picked these datasets at random from a list of popular data used for data science projects. Each dataset has a large amount of unique entries, so they should be good enough to analyze for this project. I am going to use R for the analysis and visualizations of the data, and the code will be available in the Appendix.

**Data**

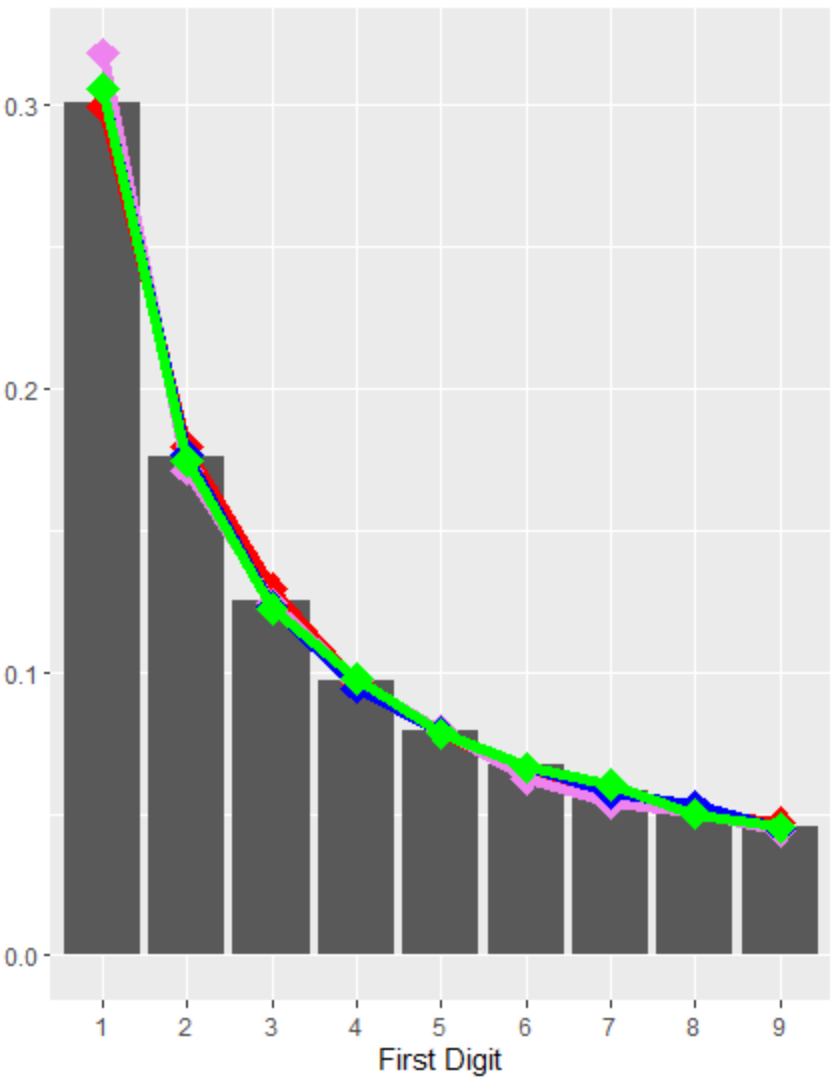
The data will be split up into 3 sections. First, the random numbers will be 10000 randomly generated integers, from 1 to 100, 1 to 1000, and 1 to 10000, to see if the range of the data matters. Next, the YouTube videos will be analyzed using the amount of views of each video, and the number of likes, dislikes, and comments that each video has received. And finally, video game sales will be analyzed, over all popular platforms (NES, Wii, Playstation and XBOX) for the past 30 years, and they will be grouped by region (North America, Europe, and Japan). These amounts are reported in millions, and therefor some of the data will be smaller than 1 (in millions of copies sold).

**Analysis**

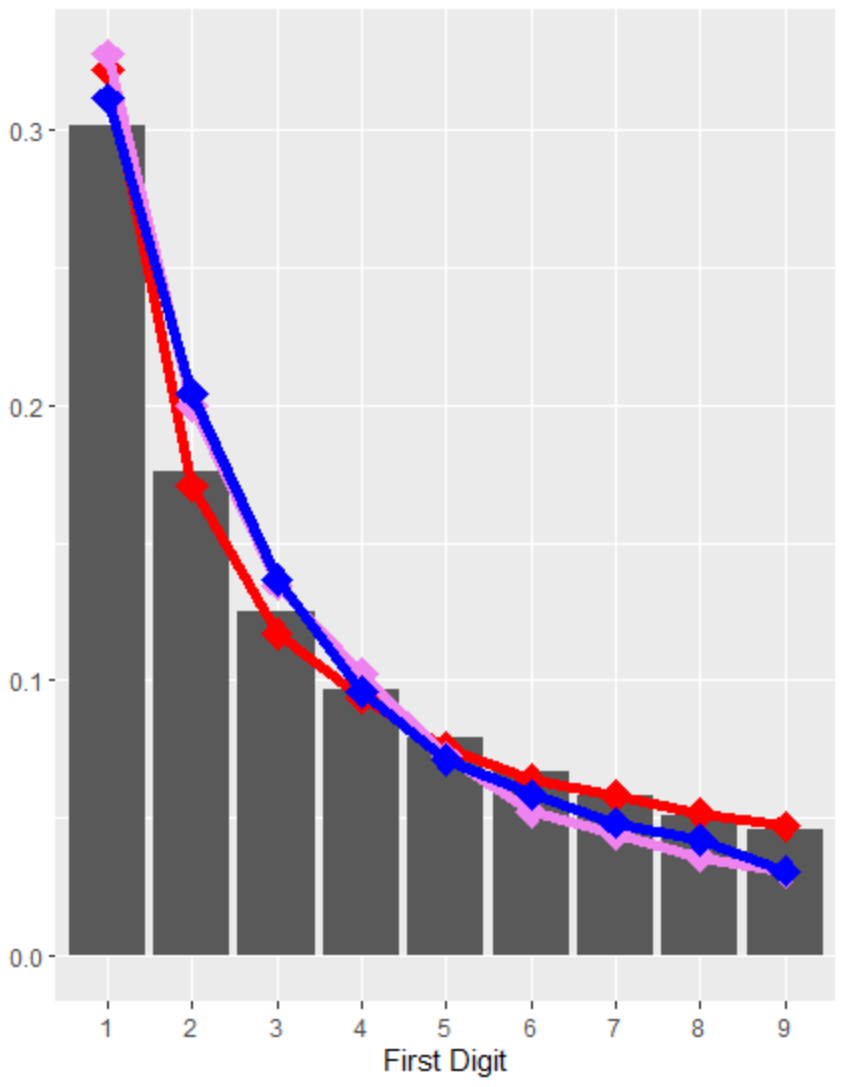
First, the random numbers were created and tested. I tried 3 tests, all using 10000 integers. The tests were for the numbers ranging from 1 to 100 (shown in red), 1 to 1000 (shown in violet), and 1 to 10000 (shown in blue). All of the tests show similar, and predictable, results. Each digit shows itself at the beginning of each number about one-ninth of the time, or around 11.1%. Random number generators do not follow Benford’s Law, as we can clearly see below.



Next, we ran the analysis of the Youtube videos. All 4 of the variables (number of video views, likes, dislikes, and comments) follow Benford’s Law!



Moving onto the video game sales figures, we see again that all 3 regions follow Benford’s Law in regards to sales of different video games. The North American market follows the pattern more accurately than the other 2, but they all are relatively close.



**Conclusion**

Both datasets follow Benford’s Law…but why? The answer seems to be that for large data sets of random numbers, mathematical logarithms of those numbers align in this manner, so that the first digit occurs most frequently at approximately 30%, and the higher the number, the lower the frequency. Tax and financial institutions, among many others, have used this law to find suspicious activity in records or transactions. It is very interesting to see that something as arbitrary and random, such as sales figures or YouTube views, can fall into this realm of “random values”.

**References**

<https://en.wikipedia.org/wiki/Benford%27s_law>

<https://mathworld.wolfram.com/BenfordsLaw.html>

<https://en.wikipedia.org/wiki/American_Journal_of_Mathematics>

<https://www.scribd.com/document/209534421/The-Law-of-Anomalous-Numbers>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2866333/>

**Appendix**

<https://www.kaggle.com/datasnaek/youtube-new>

<https://www.kaggle.com/gregorut/videogamesales>

<https://github.com/nickmiller1023/Projects/tree/master/Benford’s%20Law>