Fraud detection using Python

Using Python and Benford analysis to detect the presence of fraud in the 2020 election

BIG DISCLAIMER

- I chose this because I think it's a cool way to show off Python, PyCharm and a basic plotting library (matplotlib)!
- AFTER I coded this up, I looked around and found out that since I originally messed around with Benford in 2016, it went mainstream for a while after the 2020 election. People shared pretty basic memes about this on Insta/Twitter/Reddit and Facebook trying to prove fraud. Maybe they should have coded it up themselves…?
- I don't want this to get any sort of political - the focus is on the cool stats and the code. I can't even vote.

Rudy W. Giuliani
@RudyGiuliani

The Biden selection by the Crooked Media is based on unlawful votes in PA, Mich, GA, Wisc, Nevada et al.

We will prove it all.

This claim about election fraud is disputed

1:03 AM · Nov 9, 2020 from Washington, DC

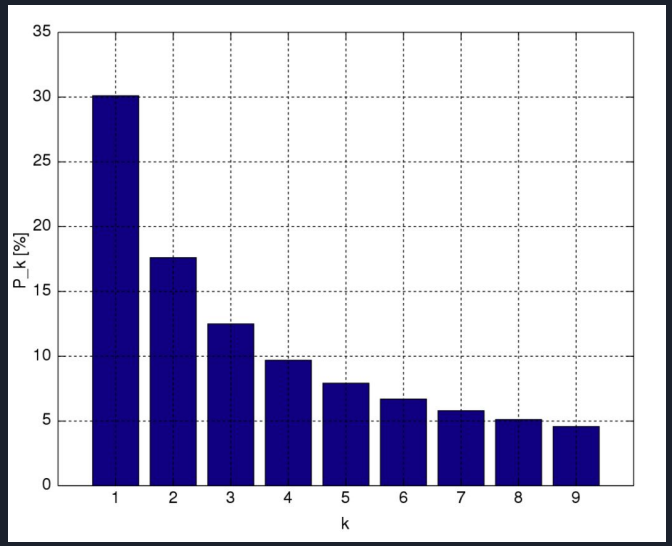
107K 16.2K Copy link to Tweet

Tweet your reply

What even is Benford’s law?

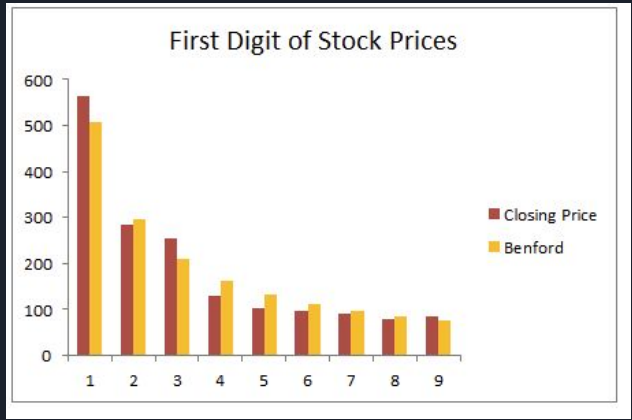
- Benford’s law was first observed by Astronomer Simon Newcomb by 1881.
- The fact it is called Benford’s law is due to Stigler's law of eponymy - no scientific discovery is named after its original discoverer (Hubble, Pythagorean theorem, and Stigler’s Law).
- Newcomb realised that, when using log tables, the pages in the front were more worn than those in the back.
- He realised that, for measurements and constants in nature, the leading digits were much more likely to be small than large.
- The frequency of occurrence of the leading digits in naturally occurring numerical distributions is predictable and nonuniform.
- Most people expect a uniform distribution - in fact a given number is 6x more likely to start with a 1 than a 9.

A useful tool for fraud detection.



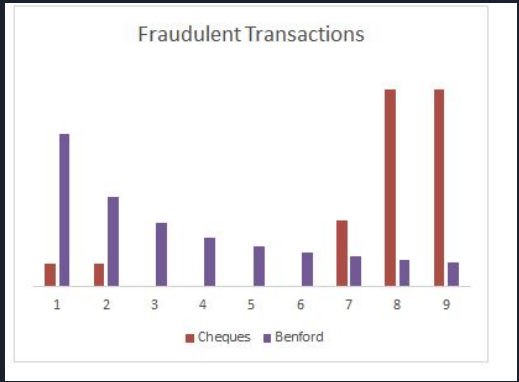
The distribution of first digits, according to Benford's law. Each bar represents a digit, and the height of the bar is the percentage of numbers that start with that digit.

- This is counter-intuitive - and because of this, it's pretty good for fraud detection. People often fail to account for it when fudging the numbers.
- Benford's Law is a 'distribution of distributions' - and therefore, financial datasets (and voting tallies) tend to conform to it.
- **Comparisons are legally admissible as evidence of tampering at federal, state and local levels.**
- State of Arizona v. Nelson 1993 - CV92-18841



Intuitive Example (used in the State of Arizona v. Nelson case)

- An intuitive explanation of Benford's law is to consider the total assets of a mutual fund that is growing at 10% per year. When the total assets are $100 million, the first digit of total assets is *1* . The first digit will continue to be *1* until total assets reach $200 million. This will require a 100% increase (from *100* to *200*), which, at a growth rate of 10% per year, will take about 7.3 years (with compounding).
- At $500 million the first digit will be *5* . Growing at 10% per year, the total assets will rise from $500 million to $600 million in about 1.9 years, significantly less time than assets took to grow from $100 million to $200 million. At $900 million, the first digit will be *9* until total assets reach $1 billion, or about 1.1 years at 10%. Once total assets are $1 billion the first digit will again be *1* , until total assets again grow by another 100%. The persistence of a *1* as a first digit will occur with any phenomenon that has a constant (or even an erratic) growth rate.



Let’s code this up!

- People keep telling me not to trust the results of elections. I want to prove to myself that the elections were legitimate.
- Load a dataset from the election
- Record the frequency of the occurrence of the first digits
- Use a Chi-Square test (goodness of fit test)
- Present the results!

So...did we find evidence of widespread election fraud?!

Not really.

Apparently, Reddit and Facebook beat me first, and memes disputing Biden's victory in 2020 have beaten me to the punch. That means that concerns have already been addressed...

Theodore P. Hill, Professor Emeritus of Mathematics at Georgia Tech ->
"Benford's Law can NOT be used to "prove fraud"," he told Reuters by email. "It is only a Red Flag test, that can raise doubts."

Not all data sets should follow Benford's law; for example, if most of the precincts have approximately the same population and each candidate's support is the same in each precinct, there will be a clustering of leading digits.

One solution is to look at second digits, which have their own distribution as observed by our old friend Newcomb (who observed this phenomenon in the first place).

What else can you do?

- Look at the second digit
- Count in base 3 rather than base 10 - numbers get spread out over several more magnitudes (81 is a two digit number in base 10, but in base 3 it is 100003, five digits, and this spreads out clumped data) -Alex E. Kossovsky and Steven J. Miller
- Do literally any other project in this book; I don't know why I picked this one.