Benford's Law

Setup

My last tweet shared this YouTube video from Stand-up Maths, https://www.youtube.com/watch?v=etx0k1nLn78&t=213s, "Why do Biden's votes not follow Bendford's Law?", as a good in-depth analysis of using Benford's distribution as well as other distributions of digits. A comment, https://twitter.com/TIves1995/status/1326529491343187968, was posted that raises some concerns about the violation of Benford's Law.

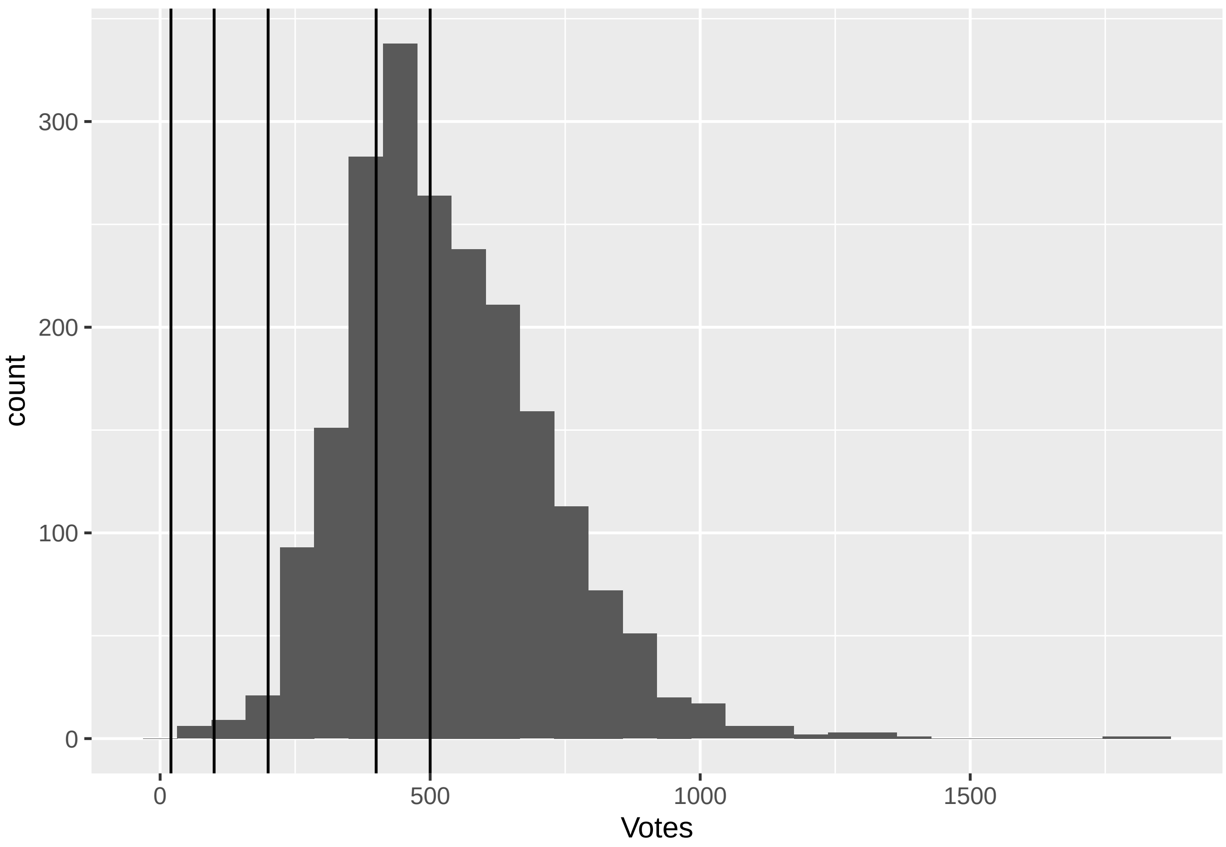
This comment makes a series of claims:

1) the Biden vote count violates Benford's Law
2) the Biden vote count is still in violation of Benford's Law when aggregating across multiple datasets
3) Benford's Law applies for a random sampling of a given event (always)
4) the aggregate of a separated dataset should follow Benford's Law
5) the datasets should mirror each other in Benford's Law
6) the aggregate of Trump's and Biden's votes should follow Benford's Law

I will do my best to interpret what these claims are mathematically, to see if the Chicago vote data exhibit these violations, and to run simulations that may explain why we see these violations.

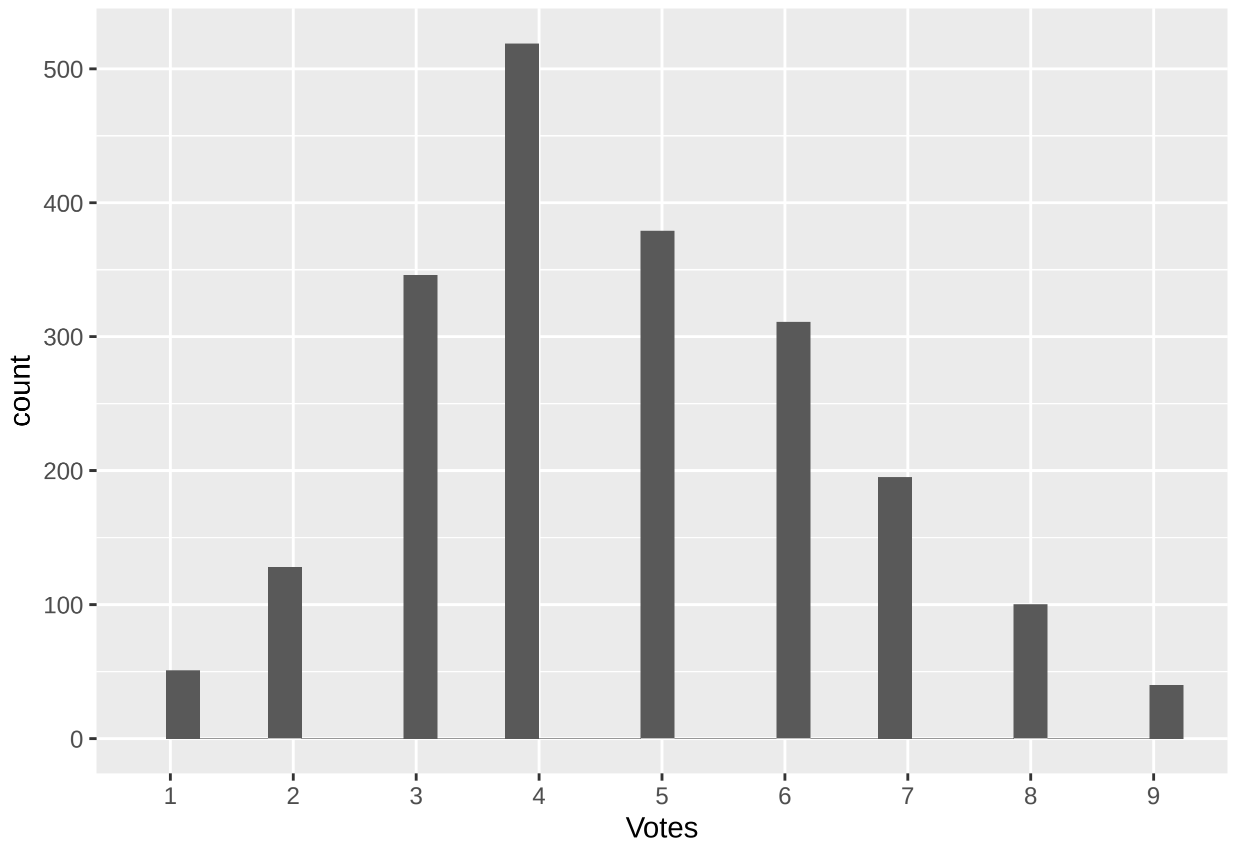
Chicago Vote Data

So, first, let's get an overall feel for the data. I use the Chicago vote data from this website: https://chicagoelections.gov/en/election-results-specifics.asp.



Above is the distribution of observed vote counts in each precinct. 20, 100, 200, 400 and 500 vote counts are highlighted by vertical lines from left to right. From this graph we would not expect to see many "1s" to be the leading digits since there are no precincts with votes in the range [10, 20) but we would expect to see some since there are precincts with vote counts in [100, 200). Many of the precincts have vote counts in the range [400, 500).

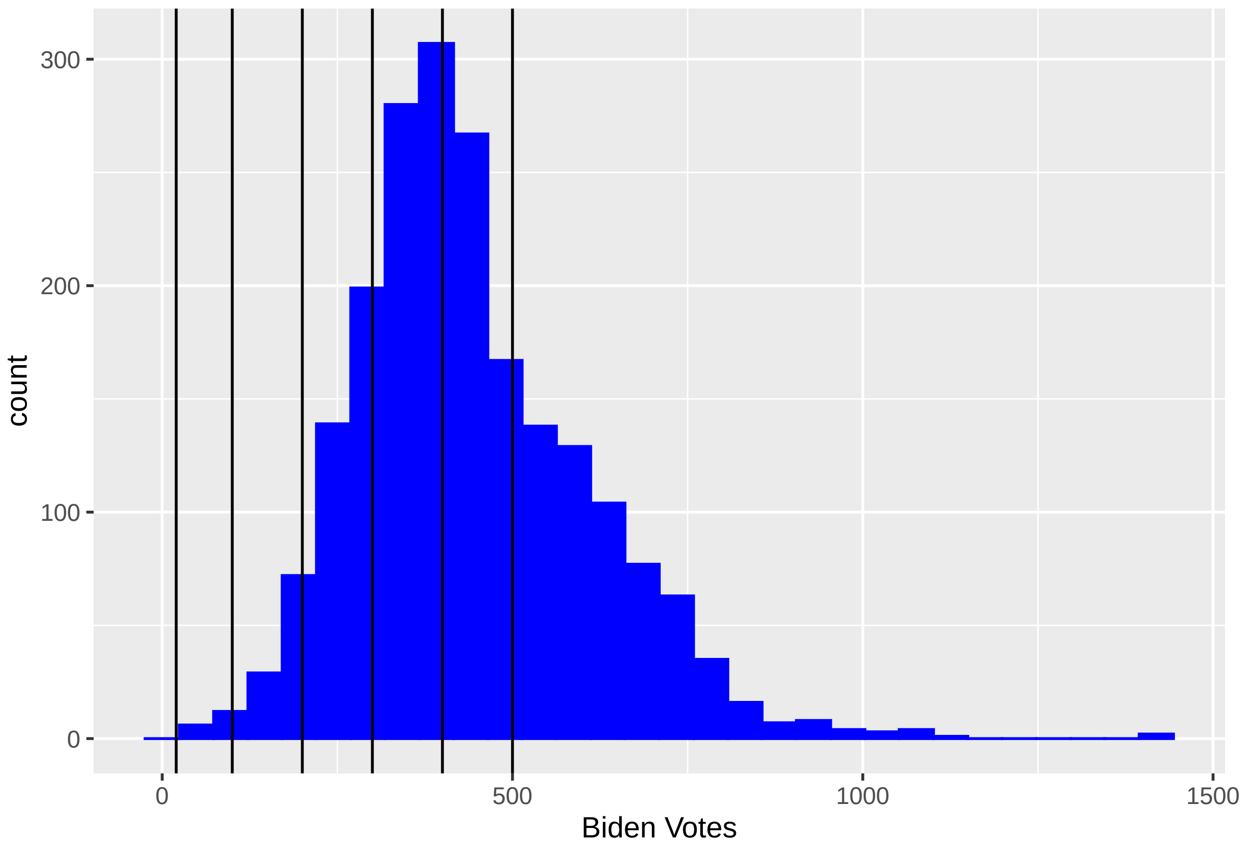
Benford's distribution can show us just that.



Above is the distribution of the leading digit of the total observed vote counts in each precinct (same dataset as above). As we expected we do not see many leading "1s" but instead see many leading "4s". This is because the vote has a peak at 400 votes. Here we can see that statement 3 above does not always apply. In fact this is actually highlighted in the video at time markers ~3:00, ~8:30, and ~10:00, this is evidence against claim 3) ("Benford's Law applies for a random sampling of a given event (always)").

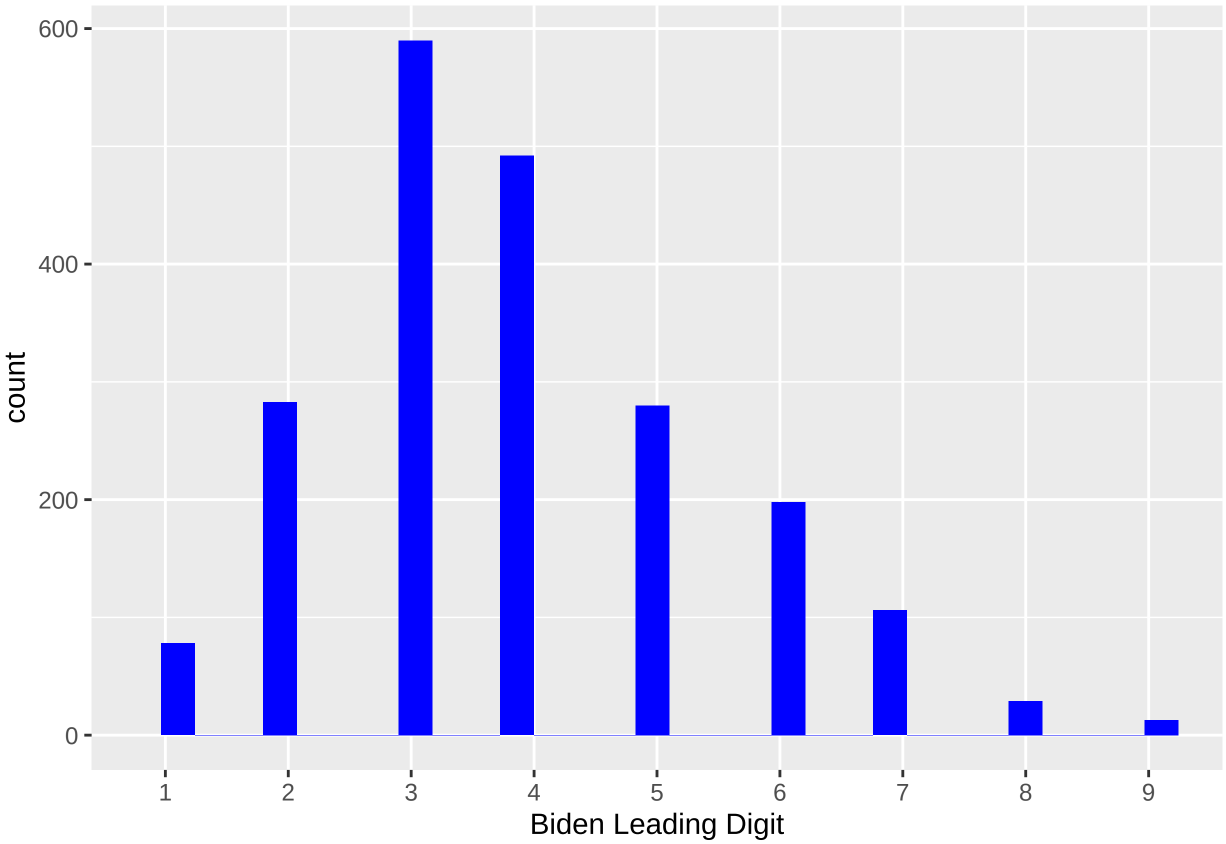
Biden's Chicago Vote Data

Let's investigate Biden's vote counts.



Above is the distribution of observed vote counts for Biden in each precinct. 20, 100, 200, 300, 400 and 500 vote counts are highlighted by vertical lines from left to right. From Biden's graph we would not expect to see many "1s" to be the leading digits since there are no precincts with votes in the range [10, 20) but we would expect to see some since there are precincts with vote counts in [100, 200). Many of the precincts have vote counts in the range [300, 500). This is in fact very similar to the entire vote distribution. We saw that the total vote counts in each precinct did not follow the canonical shape of Benford's distribution because the total votes counts do not span across several orders of magnitude like Biden's. This distribution is mainly grouped in the hundreds. The assumptions for this test are not present in this distribution.

Let's look at the leading digit distribution of Biden's vote counts.



Above is the distribution of the leading digit of the Biden's observed vote counts in each precinct (same dataset as above). As we expected, we do not see many leading "1s" but instead see many leading "3s" and "4s". This is because the vote has a peak around this many votes 300 to 400 votes. It seems as if Biden's vote violates Benford's law and is subject to fraud but, as we have shown above, Benford's Law is not appropriate for this distribution. This addresses claim 1) ("the Biden vote count violates Benford's Law").

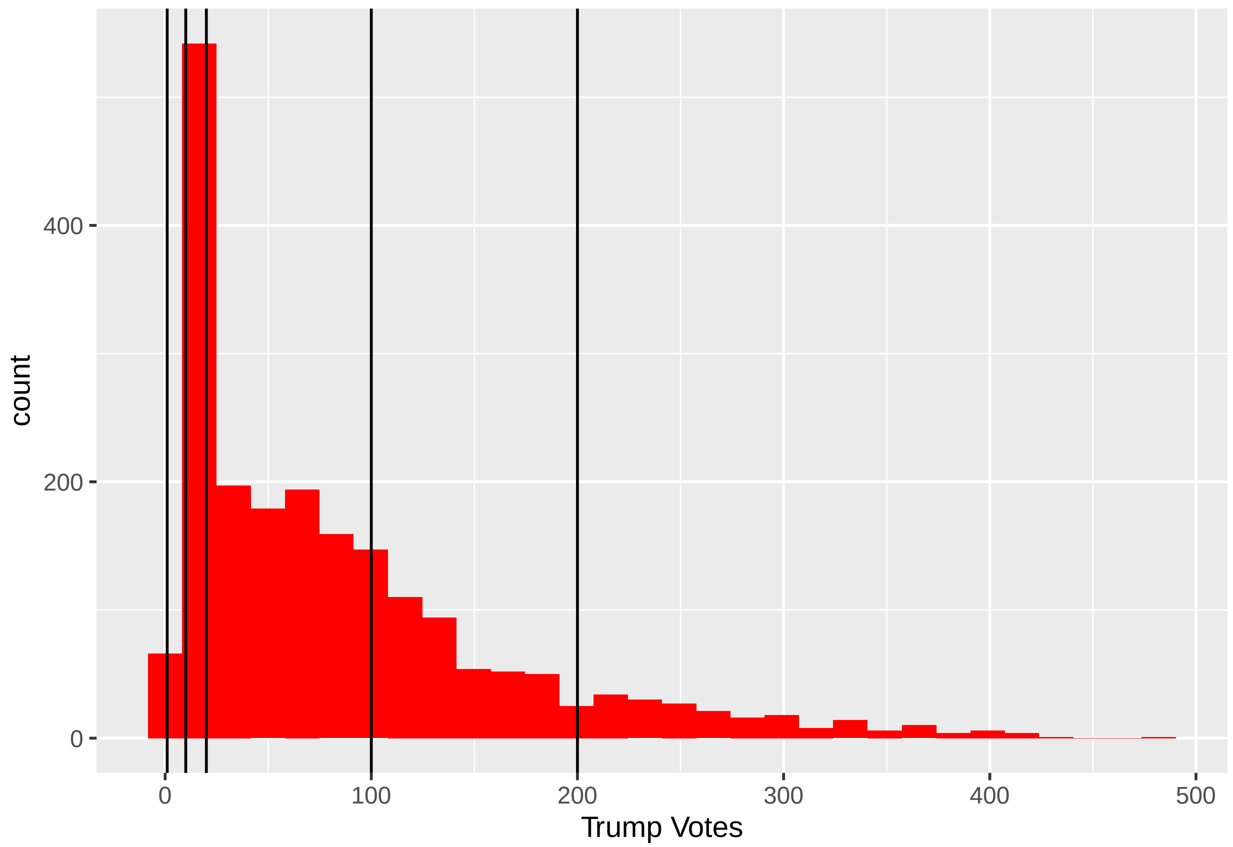
Aggregating Data

I have not heard of claim 2) ("the Biden vote count is still in violation of Benford's Law when aggregating across multiple datasets") and this statement is a little ambiguous. There are two interpretations I can imagine: A) the distribution of the leading digits of the sum two candidates vote counts should follow Benford's Law and B) the distribution of leading digits of both candidates vote counts should follow Benford's Law.

Interpretation A) is not likely. If we split the data set and recombine the dataset, we would expect to see the same distribution of leading digits as the entire dataset. This is simply because we are applying and function to the dataset and applying its inverse function which exactly undoes the the first function.

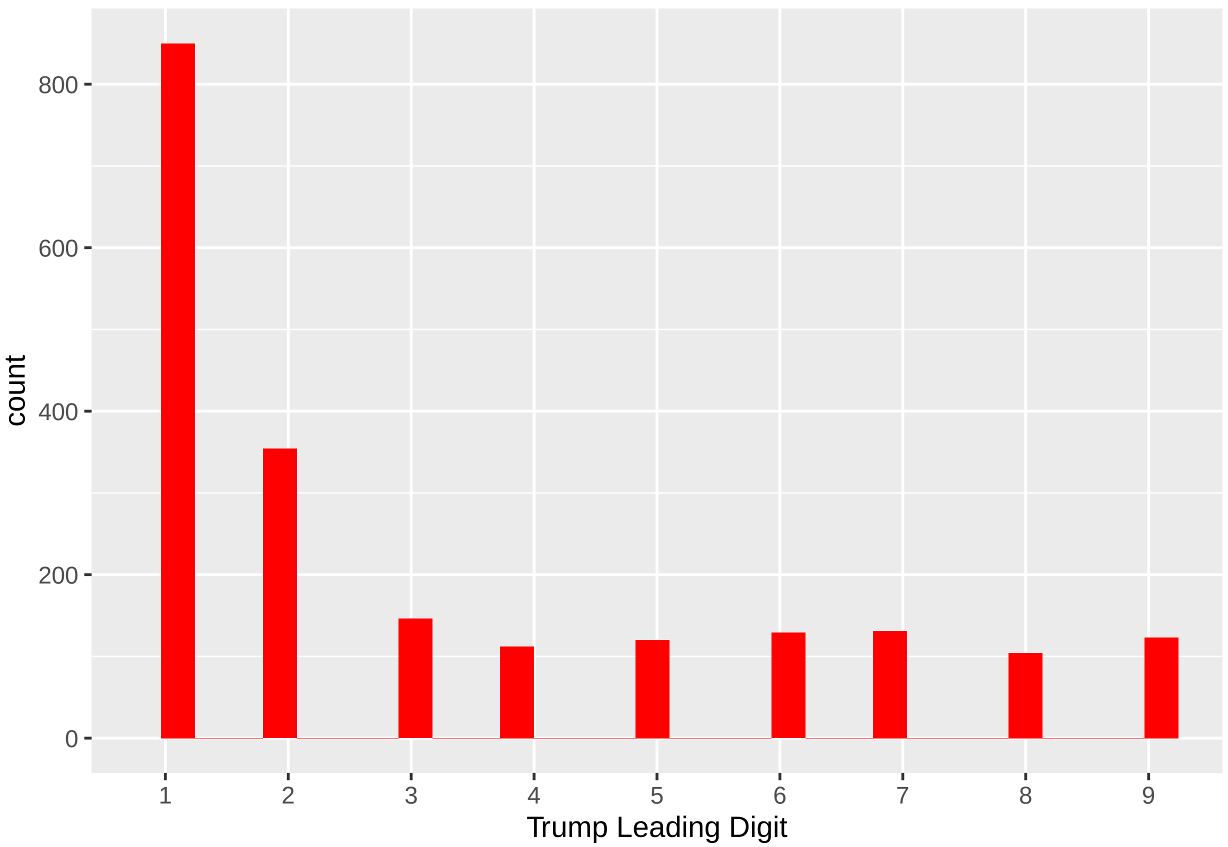
A Detour to Trump's Chicago Votes

First, we will look at the distribution of Trump's vote counts.



Above is the distribution of observed vote counts for Trump in each precinct. 1, 10, 20, 100, and 200 vote counts are highlighted by vertical lines from left to right. From this graph we would expect to see many "1s" to be the leading digits since there are many precincts with votes in the range [10, 20) but we would not expect to see less "3s" since there are very few precincts with vote counts in [300, 400).

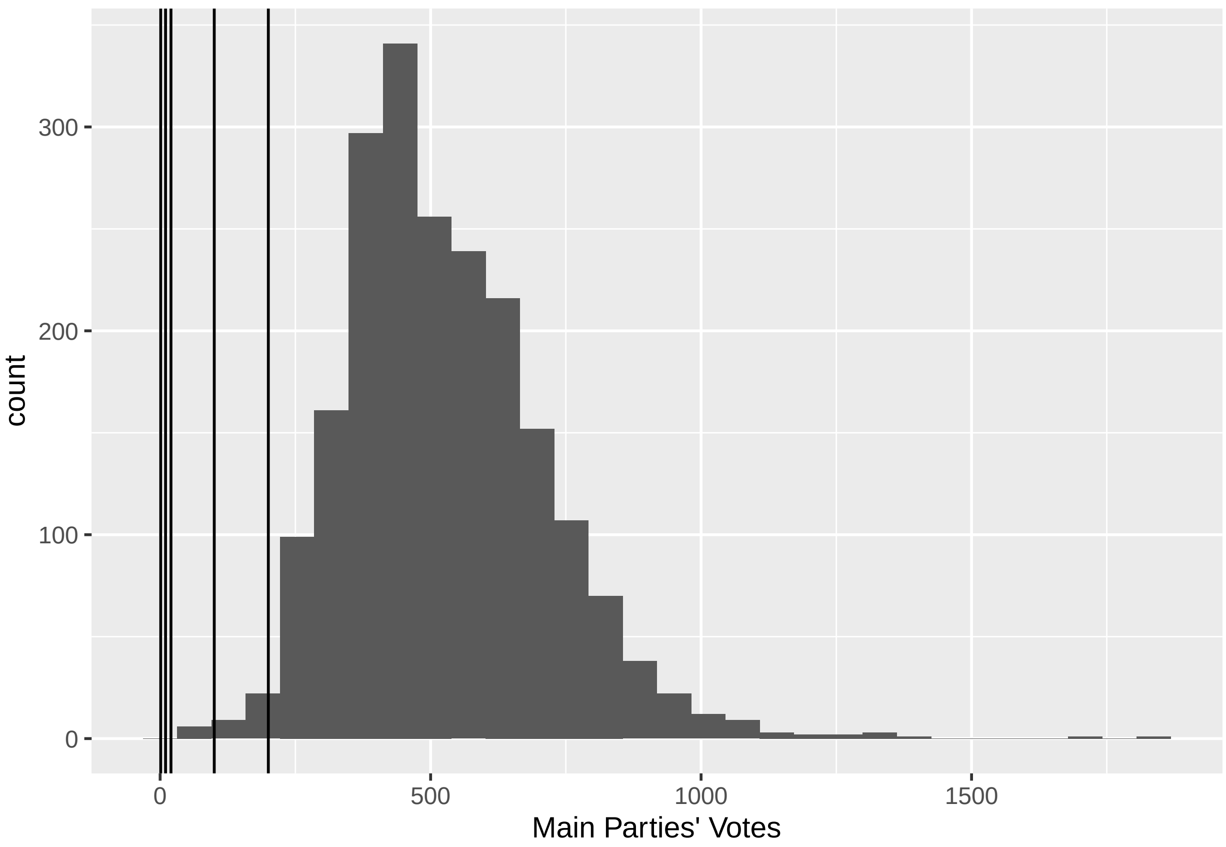
Now let's see Trump's leading digit distribution.



Above is the distribution of the leading digit of the Biden's observed vote counts in each precinct (same dataset as above). As we expected we do see many leading "1s" and few leading "3s" and "4s".

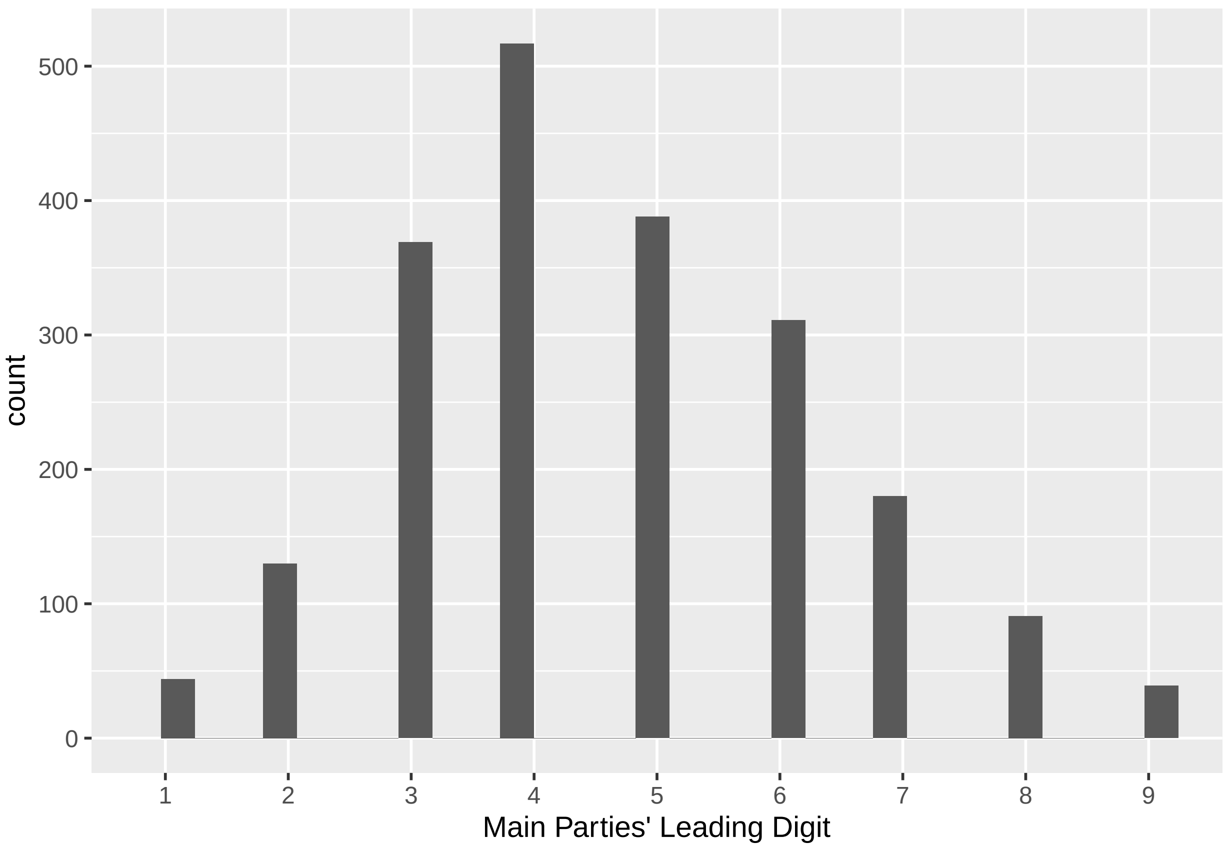
Aggregating Data Interpretation A)

Now let's first look at the distribution of the sum of Biden's and Trump's vote counts.



This distribution looks much like the distribution of the entire vote count for each precinct. This should be expected since the majority of people did not vote third party.

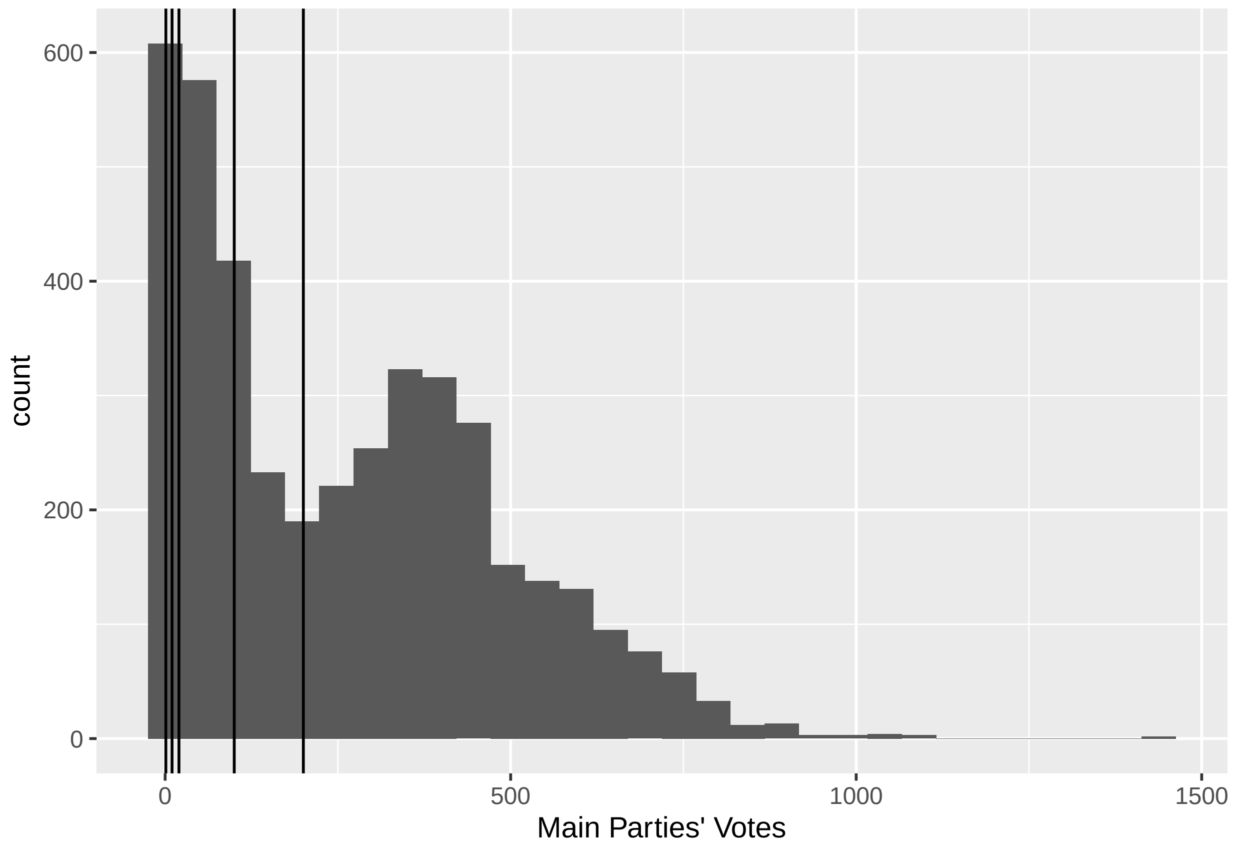
Again, we can look at the leading digit distribution of the sum of the two candidates.



This distribution looks much like the leading digit distribution of the entire vote count for each precinct. This should again be expected since the majority of people did not vote third party. So even if interpretation A) is correct, there is no reason for alarm.

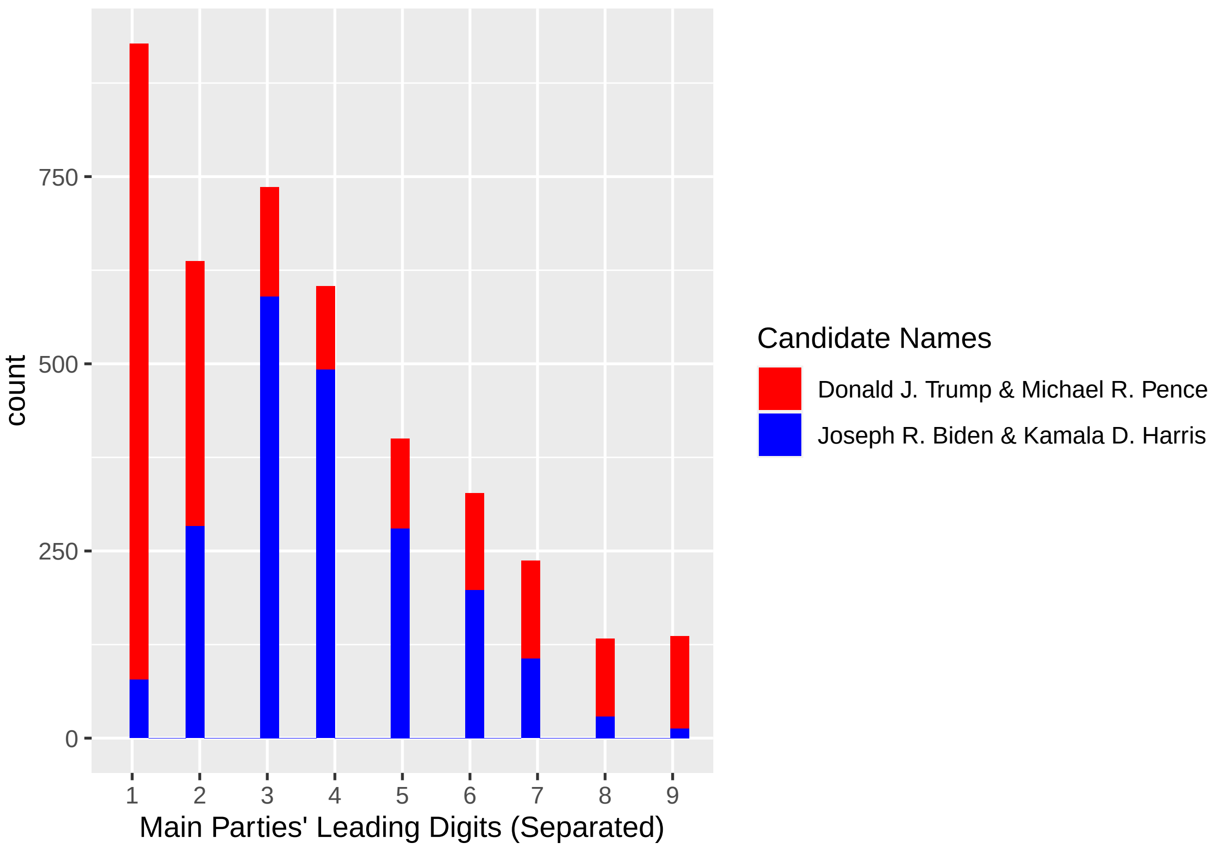
Aggregating Data Interpretation B)

Let's continue to interpretation B).



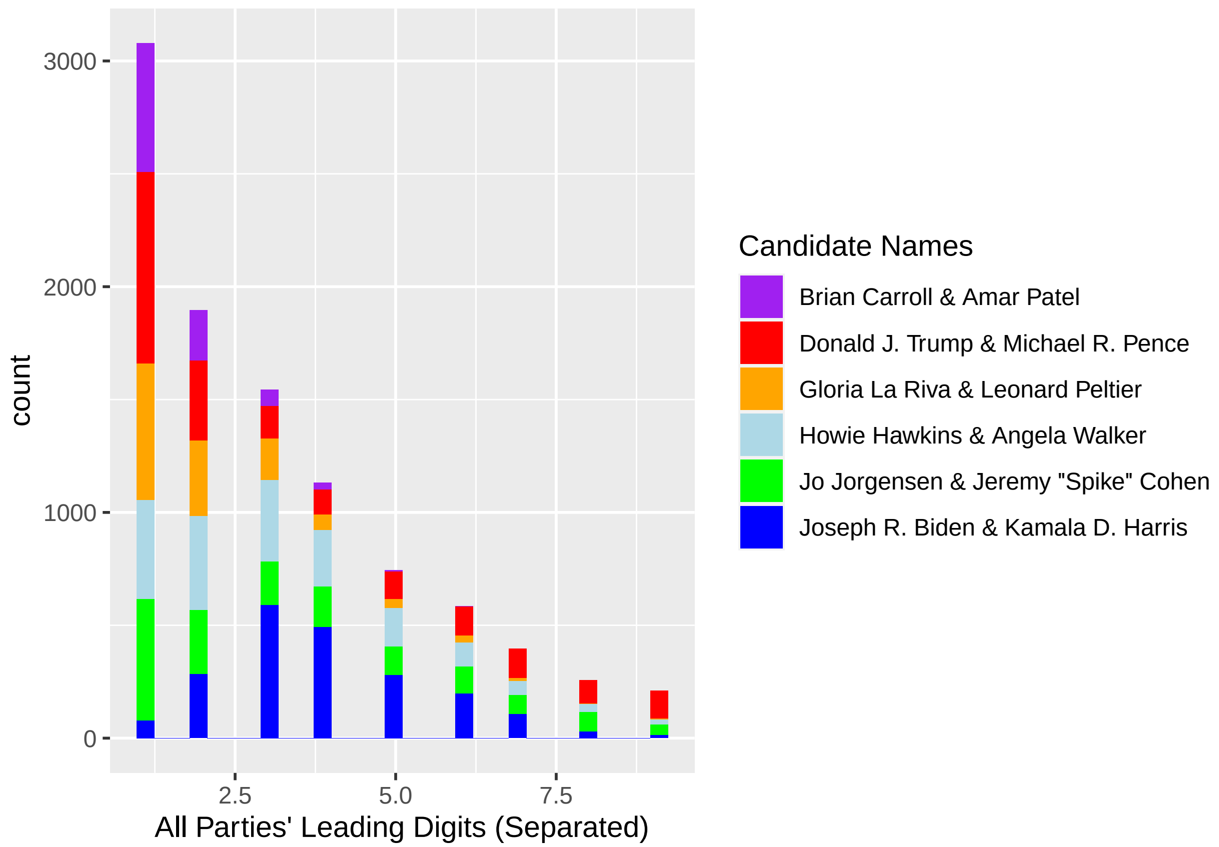
Above is the distribution of observed vote counts for Biden and for Trump in each precinct. 1, 10, 20, 100,and 200, vote counts are highlighted by vertical lines from left to right.

Let's inspect the leading digit distribution of the vote counts for both Trump and Biden.



Here we can see a bump in the counts at 3 and 4. This means that claim 2) ("the Biden vote count is still in violation of Benford's Law when aggregating across multiple datasets") seems to have some credence in this plot, specifically using the scenario in claim 6) ("the aggregate of Trump's and Biden's votes should follow Benford's Law").

However, there were actually 4 other parties on the ballot in Chicago: Howie Hawkins & Angela Walker, Gloria La Riva & Leonard Peltier, Brian Carroll & Amar Patel, and Jo Jorgensen & Jeremy ''Spike" Cohen. Let's see if the aggregate of all 6 of these candidates leading digit distribution follows Benford's Law.

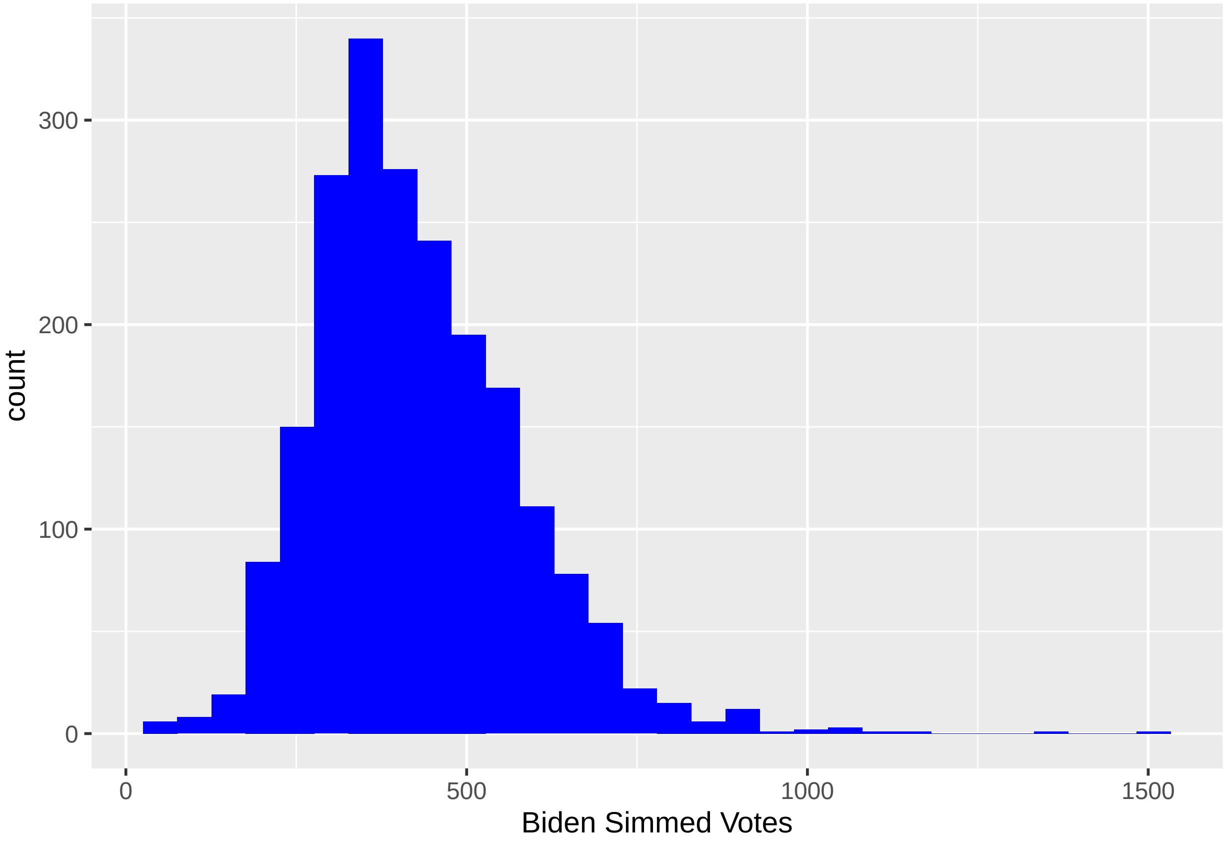


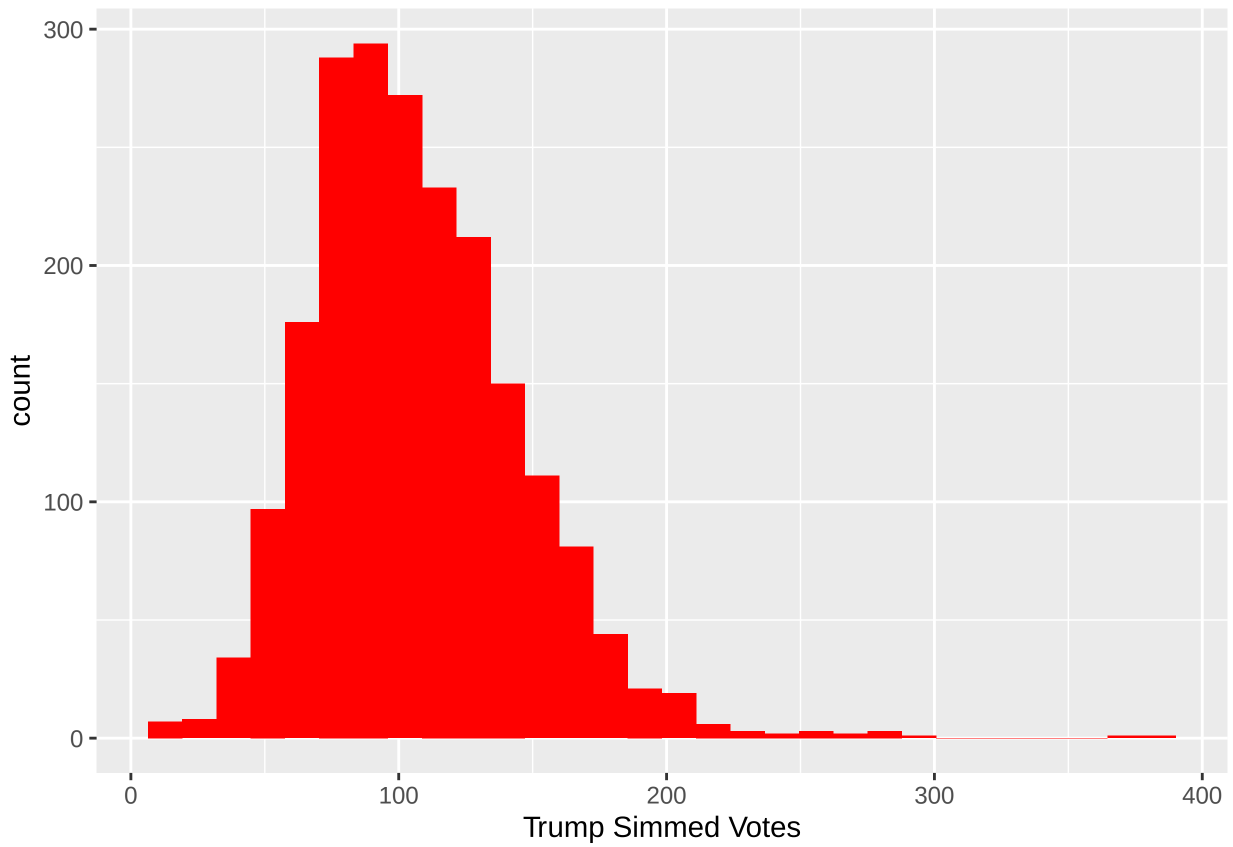
This aggregated set seems to be well predicted by Benford's Law. Then when can conclude that claim 2) ("the Biden vote count is still in violation of Benford's Law when aggregating across multiple datasets") has evidence against it.

Modelling Aggregating Data Interpretation B)

So let's now pretend that the dataset was split into two groups instead of 6 groups. (This is of course not the best simulation of the dataset since we ignore third parties but we can investigate some of the above claims using this model.) We can look at what the distributions should be when this is done to this dataset set. The probability that a Biden vote is cast in these precincts is about .80 and that means that we will give a Trump vote a .20 chance. We will first simulate how many votes Biden will take in each precinct then give the remain votes in the district to Trump.

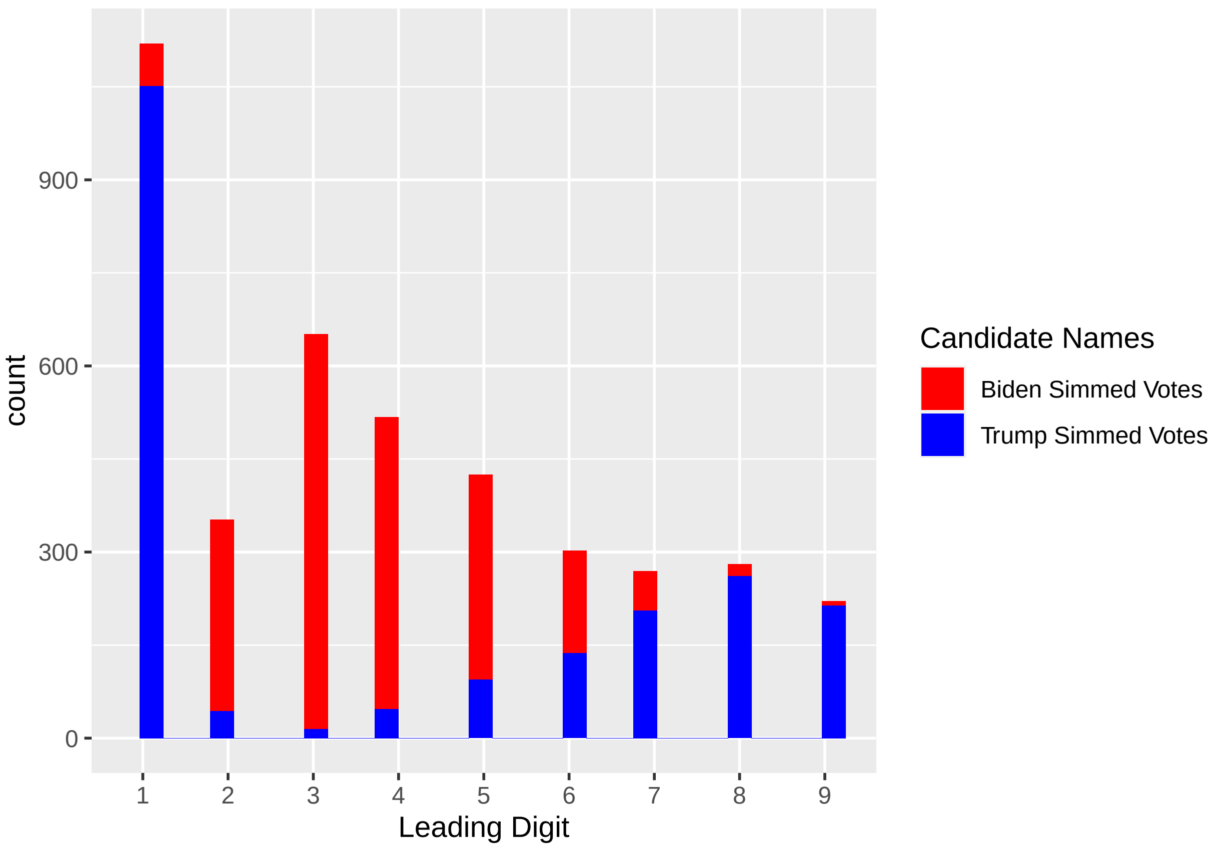
Let's look at Biden's and Trump's simmed votes.


count
300
200
100
0
0
500
1000
1500
Biden Simmed Votes



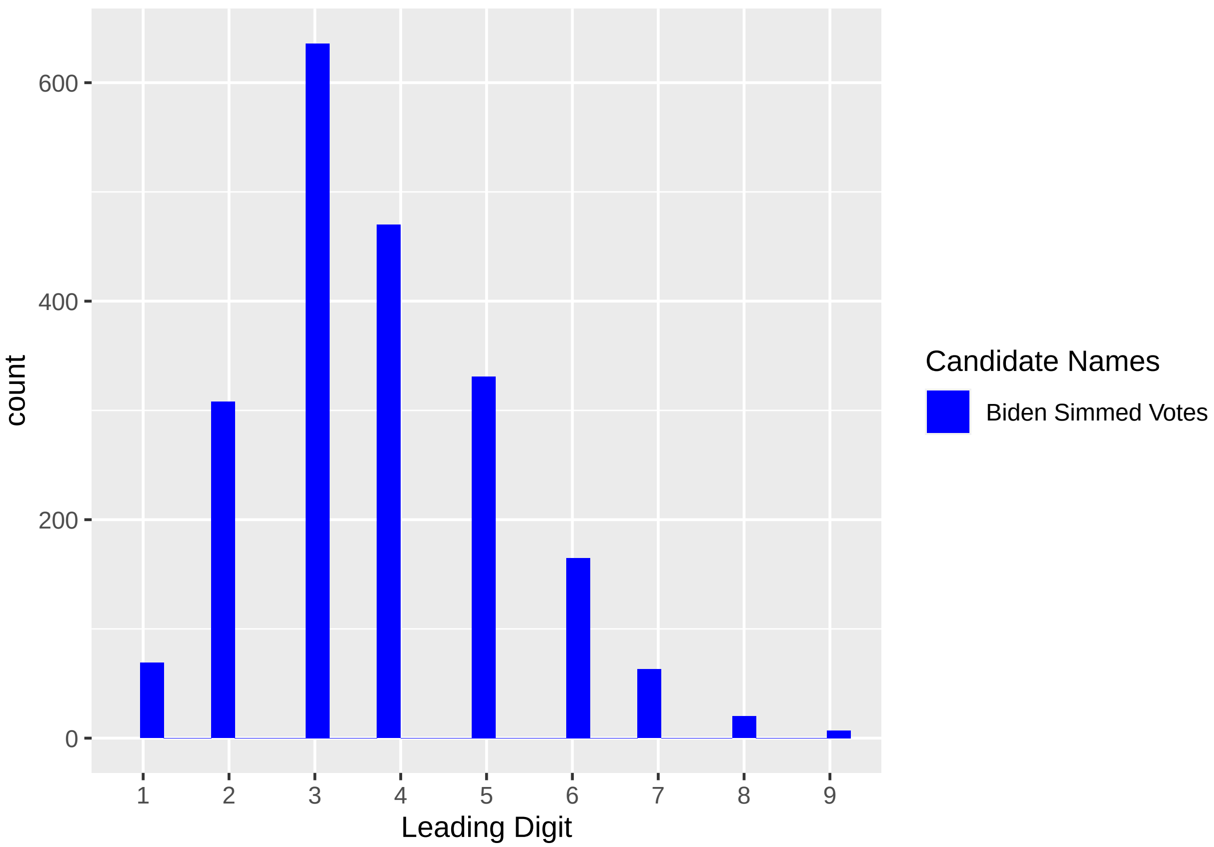
We can see that Biden's sim votes look similar to Biden's observed votes but we can not say the same about Trump's. Trump's simmed vote distribution is most different from his observed vote distribution because this model does not account for the other 4 parties and thus does not well reflect the probability of a Trump vote. Say the probability of a Biden vote is actually .8, then this model will over estimate the probability of a Trump vote because the remaining probability is given to Trump in this simulation.

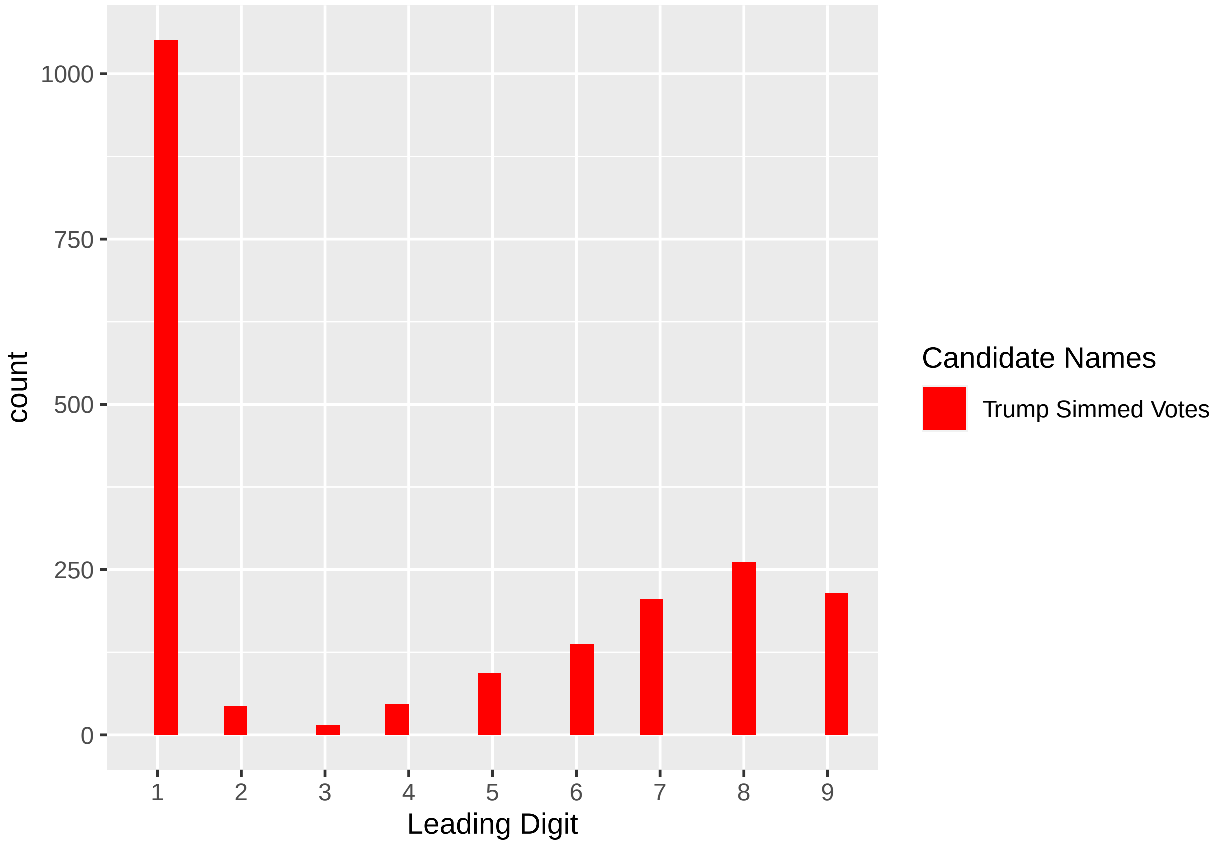
Let's look at Biden's and Trump's leading digit distribution under interpretation B).



This aggregate leading digit set is also in violation of Benford's Law. This seems to be due to the probability of voting for Trump or Biden and the number of votes counted in each precinct. This means that claim 4) ("the aggregate of a separated dataset should follow Benford's Law") is not always true. If we assume that claim 5) ("the datasets should mirror each other in Benford's Law") is actually that the two distributions should have the complement probability of observing each leading digit, then we can also conclude that this is false. However, claim 5) ("the datasets should mirror each other in Benford's Law") is a little ambiguous as well.

We can also look at the leading digit distribution for the simmed votes for each candidate.

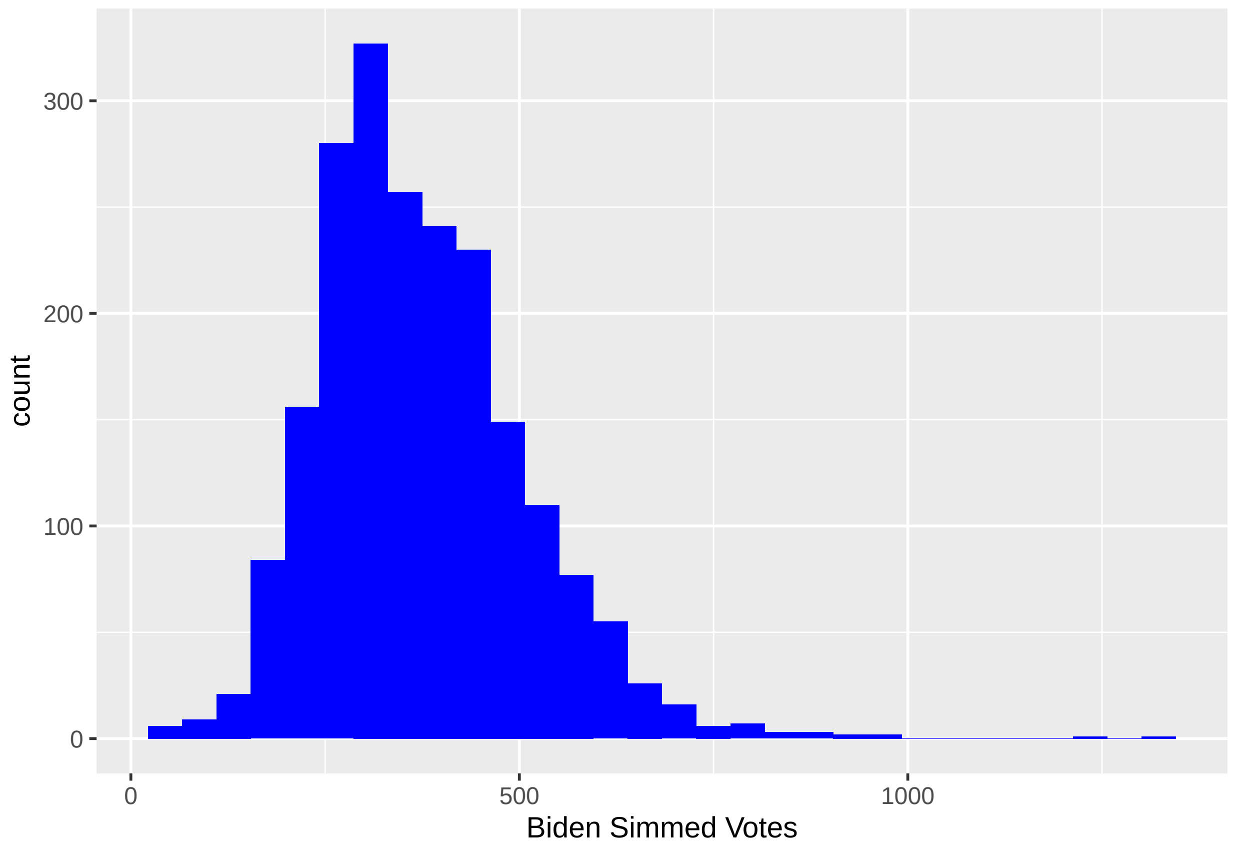

600
400
200
0
count
1
2
3
4
5
6
7
8
9
Leading Digit
Candidate Names
Biden Simmed Votes

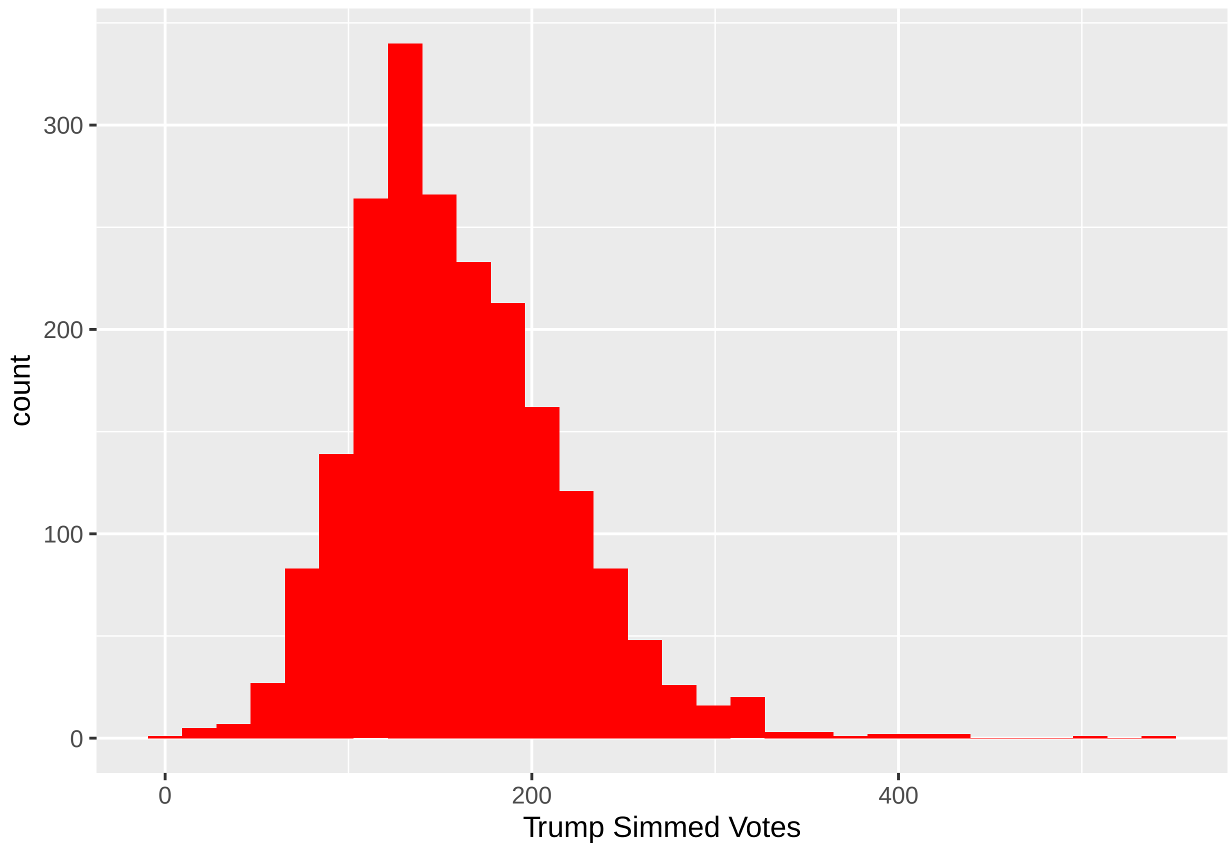


Claim 5) ("the datasets should mirror each other in Benford's Law") does appear to be correct that there are more "3s" and "4s" in Biden's simmed distribution and a lack in Trump's simmed distribution. Likewise, there appear to be more "1s", "7s", "8s", and "9s" in Trump's distribution and a lack in Biden's distribution. This highlights that claim 5) ("the datasets should mirror each other in Benford's Law") does not lead to claim 4) ("the aggregate of a separated dataset should follow Benford's Law)" and 6).

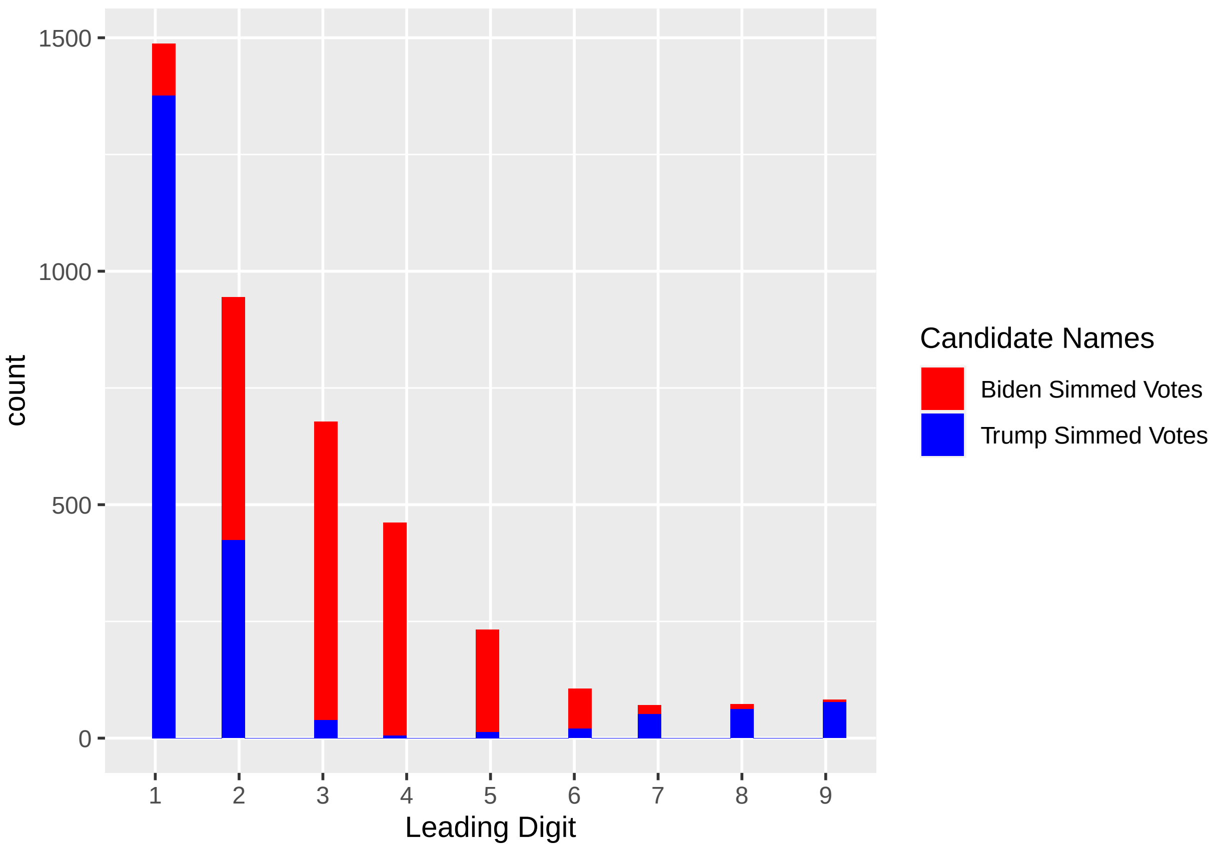
Let's look at another set of probabilities for each candidate, Biden .7 and Trump .3.

Let's look at Biden's and Trump's sim votes.


count
300
200
100
0
0
500
1000
Biden Simmed Votes



Let's look at Biden's and Trump's leading digit distribution under interpretation B).



This now seems to not violate Benford's Law in the aggregate.

We have walked though the claims made that Biden's vote counts still show irregularities. These irregularities may arise due to the size of the voter population in Chicago during the election and the probability of the voters voting for Biden. These characteristics of the dataset produce distributions that are noted exceptions of Benford's Law, well explained in the video, and further explained here in this analysis. We have also investigated other claims about aggregate distributions of the leading digits, and that the Trump and Biden vote counts are in violation of Benford's Law in this respect. These claims do not seem to have good grounds to refute that this dataset is not an exception of Benford's Law. As stated in the video, Benford's Law is a problematic forensic tool when applied to elections.