Statistical analyses attempting to determine election fraud are flawed without a causal framework

Norman Fenton[1] and Martin Neil[2], 13 November 2020

In the last week there has been much discussion about whether statistical analysis alone can establish if there was fraud in the US election. Notable examples include:

1. The claim that the first digit total votes for Biden in Chicago districts defy Benford's Law and hence must be fraudulent. This video by Matt Parker https://youtu.be/etx0k1nLn78 provides a description of Benford's law along with an explanation of why it is not relevant in this case. In other words, the 'statistical analysis' based on Benford's Law does not establish fraud.
2. The claim that larges batches of postal votes (in one case a batch[3] of over 23,000) all of which were votes for Biden must represent fraud because it is a statistical impossibility otherwise. This claim only works if 'fraud' or 'luck' were the ***only*** possible causal explanation for 23,000 consecutive votes all cast for Biden. If these were the only possible causal explanations then this would certainly prove fraud even if the ballots came from a district where, say 90% of people really are Biden supporters. The probability that all 23,000 votes would be for Biden purely by chance given that each vote has a 90% probability of being for Biden is 0.9 to the power of 23,000. That is a number much smaller than 1 divided by the number of atoms in the observable universe. People saying it is 'as likely as being struck by lightning' are massively understating how unlikely it would be. More like being struck by lightning on several consecutive days. However, the argument is flawed if there is another plausible causal explanation for the bag contents other than fraud and luck. For example, it may be possible that these ballots were part of a set that had already been counted and sorted. Or, perhaps, this was a deliberate hoax or set-up. So, the focus needs to be on whether any of these alternative explanations is feasible rather than on the statistical analysis. The statistical analysis only proves that the batch ***cannot*** have come from a random set of ballots.
3. The claim that the sudden large swings to Biden which started happening in key swing states after the counting stopped at 3.00am on election night (as in this analysis[4] by an anonymous data scientist) prove fraud. Assuming the data here are accurate, this does indeed look like convincing evidence of fraud. However, because it is at a State level, there could still be a causal explanation other than fraud. For example, it may be possible that large numbers of ballots that came in late were primarily from Biden-supporting areas.

If there really was fraud, the simplest and most efficient way of identifying it statistically would be a variation of what was done by the anonymous data scientist above but ***at a much more local/granular level*** and focused ***only on postal ballots***. in other words, districts sufficiently small such that there is less chance of a systematic or random interference in the natural process by which ballots are collected (no mail sorting, no sorting at the centre into for/against bundles etc. i.e. the "draws" come naturally as close as possible on a per household basis). The more granular we get, the closer we are at detecting anomalies that are not explainable by anything other than fraud. If there is some model of causal interference, then the normal and hypothesized abnormal process need to be tested against each other i.e. against patterns from previous elections.

[1] Professor of Risk and Information Management, Queen Mary University of London n.fenton@qmul.ac.uk

[2] Professor of Computing and Statistics, Queen Mary University of London, m.neil@qmul.ac.uk

[3] https://townhall.com/columnists/joshhammer/2020/11/06/the-election-battle-is-just-beginning-n2579568

[4] https://www.zerohedge.com/political/it-defies-logic-scientist-finds-telltale-signs-election-fraud-after-analyzing-mail-ballot

We hypothesise that districts with a total of no more than 5,000 postal votes may be a suitable level of granularity to analyse. In other words at this level of granularity there seems to be no reasonable causal explanation for the distribution of votes in postal ballots counted before and after the 3.00am 'cut off' point on election night to be significantly different.

So, let's consider a hypothetical example of how we would undertake the necessary analysis if we had the relevant district level data. Consider, a district with say 4,000 postal ballots. Suppose that 2000 ballots are counted before the cut-off and candidate *A* has, say, 55% of these (i.e. 1100). Using a Bayesian analysis[5] (which assumes that the 'true proportion' of people favouring candidate *A* before we see any votes cast can be anything between 0 and 100%) observing the 1100 votes out of 2000 means that we can update the 'true proportion' of people favouring *A* as shown in Figure 1. Specifically, the true proportion is still quite uncertain – but there is only a 5% chance it will be less than 53.2 and only a 5% chance it will be more than 56.8.

As also shown in the Figure, we can use this revised probability distribution of the true proportion of *A* voters to predict the expected number of ballots for *A* out of 2000 counted after the cut-off – since we are assuming that these come from the same population of voters. This also enables us to calculate how unlikely any observed 'swing' is.

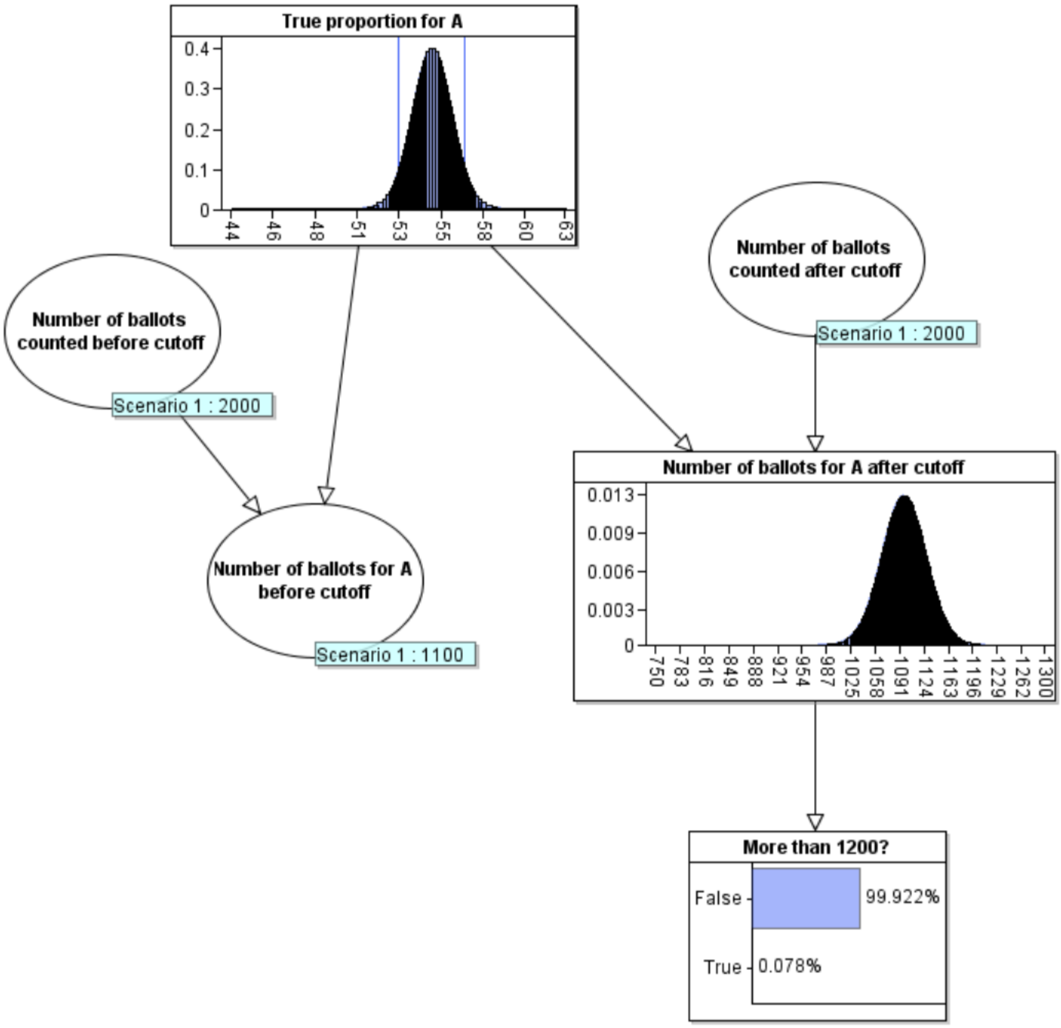

Figure 1 Bayesian analysis

[5] We also assume a Binomial distribution for the number of ballots cast for candidate *A*

For example, we can calculate the probability that there will be a swing of more than 5% between the before and after proportion of votes for *A* (i.e at least 1200 votes after, meaning 60% after compared to 55% before) as shown in Figure 1. The probability is extremely low (0.078% is less than 1 in 1000). Even a swing of just 2% in favour of *A* is unlikely (less than 10% probability).

Now, assuming we have the before and after ballot count data for a large number of districts in the same state – say 100, then if there is just one district with a swing larger than 5% to candidate *A*, this would not be so unusual that it cannot have happened by chance. There is a probability of about 7% that at least one district would have such a swing without some other causal explanation.

If, however, 4 out of the 100 had swings of more than 5% and all were in the direction of candidate *A* then this would also be so unlikely (about 1 in a million probability) that it would almost certainly require some other causal explanation. The same applies if there is even a relatively small number of instances of smaller swings all in the same direction. For example, if there are 10 swings above 2% which are ***all*** in the direction of candidate *A* then this would also be so unlikely (about 1 in 100,000 probability) that it would require some other causal explanation.

Hence, the data needed to establish fraud in the swing states are the postal ballots for a reasonable number of small areas separated into those counted before and after the night of the election.