Kristian Lum is a research assistant professor in the Department of Computer and Information Sciences at the University of Pennsylvania.

How to mislead with statistics

Lessons from recent attempts to subvert the US election with data analysis. By **Kristian Lum**, **Naim Kabir** and **Joe Bak-Coleman**

"Lies, damned lies, and statistics" is a refrain every statistician has heard. But *how* does one effectively use the tools of statistics and data science to lie and mislead? Studying examples of misleading data analysis can help us spot intentionally misleading analyses and avoid unintentionally misleading others ourselves.

The 2020 US presidential election was a perfect breeding ground for such examples. Extreme partisanship with abundant, variable-quality data laid the groundwork for shoddy statistical analyses that were misused to undermine faith in the electoral process. While the examples below are specific to the 2020 election, the tactics are generalisable to other arenas, from future elections to public health crises and climate denial.

Obscure the data pre-processing

On election night and in the days following, the *New York Times* pulled data from official election sources to display real-time vote statistics for its readers on its website. Rather than piping in raw, granular data from official sources onto its web pages, the *New York Times* displayed derived statistics such as the proportion of votes for each candidate to three decimal places. Tables 1 and 2 show hypothetical examples of underlying vote counts and corresponding processed data.

These data were intended solely for producing human-readable charts and figures. However, amateur data scientists discovered that these data were available for download (via an application programming interface) and treated them as official tallies. Using the proportions supplied by the *New York Times*, these internet sleuths "reconstructed" the total vote counts for each candidate by multiplying the rounded proportion for each candidate by the total number of votes at each time point, and then rounding again to recover integer vote counts. Table 3 gives a hypothetical example of this, showing reconstructed vote counts based on data from Table 2.

Table 1: Hypothetical example of underlying true data.

Time-stamp	Trump votes	Biden votes	Other votes
10:33 a.m.	500,000	500,000	7,000
10:54 a.m.	500,050	500,500	7,500

Table 2: Corresponding processed data for display on website.

Time-stamp	Total votes	Rounded proportion Trump votes	Rounded proportion Biden votes
10:33 a.m.	1,007,000	0.497	0.497
10:54 a.m.	1,008,050	0.496	0.497

Table 3: Corresponding reconstructed vote counts, similar to those used in blogs to claim that votes had been stolen from Donald Trump.

Time-stamp	Reconstructed Trump votes	Reconstructed Biden votes	Change in Trump votes	Change in Biden votes	Net change
10:33 a.m.	500,479	500,479			
10:54 a.m.	499,993	501,001	–486	522	–1,008

Because rounding operations can exaggerate downward shifts in vote shares over time, attempts to reconstruct vote counts from the processed data made it appear as if the total number of votes for a candidate decreased at certain points. For instance, in our hypothetical example in Table 3, Donald Trump seems to have fewer votes at 10:54 than he does at 10:33.

Without disclosing the form of the data they had pulled and the reconstruction they had done, bloggers presented "results" like these as if they were the real-time series of votes. They used these apparent decreases to suggest that votes had been stolen from Trump. This fed into the broader narrative pushed by Trump that votes had been programmatically stolen in favour of Joe Biden.

Create confusing metrics

Other sources of disinformation used (to the best of our knowledge) real election data but created derived metrics that appear reasonable but are misleading representations of the quantity they claim to measure. One such example used precinct-level vote counts in states that allow "straight-ticket" voting. In these states, voters can select all candidates from one party on a straight-ticket (ST) ballot, rather than voting for each office on a candidate-by-candidate basis using an individual-candidate (IC) ballot. Using this data, for each precinct, they calculated the "excess" votes for Donald Trump using the formula:

$$\begin{aligned}&\text{Excess Trump votes}\\&= \text{proportion of IC ballots for Trump}\\&- \text{proportion of ST ballots for Republicans}\end{aligned}$$

This metric attempts to capture whether a candidate over- or underperformed within their own party. The analysts present the proportion of ST ballots that went to

Naim Kabir is a software engineer and data scientist at Even Responsible Finance.

Joe Bak-Coleman is a postdoctoral fellow at the University of Washington Center for an Informed Public.

Republicans as a metric of how Republican a precinct is. The difference between this and the proportion of IC ballots for Trump, then, represents the excess (or deficit, if negative) votes for Trump in that precinct. However, deeper scrutiny reveals just how strange and misleading this metric can be.

Consider the example illustrated in Figure 1 in which a precinct consists of 10 Republican (R) and 10 Democrat (D) voters, each of whom may choose to vote ST (represented by squares) or IC (represented by triangles). Voters who vote for Republicans are coloured red and those who vote for Democrats are coloured blue. In this case, the excess Trump vote would be calculated as 6 (red triangles) / 8 (all triangles) – 6 (red squares) / 12 (all squares) = 0.25 or 25%. This, despite the fact that there are only two extra votes for Trump (the two red triangles from the Democrat IC voters) in this hypothetical precinct of 20 voters.

Perhaps even more telling is what happens if the proportion of voters who choose to vote ST is not equal by party. This is illustrated in Figure 2, where two of the Republican ST voters are now IC voters but their votes have not changed (they still vote for Trump).

In this scenario, we can see that the metric would indicate that the excess Trump vote was 8 (red triangles) / 10 (all triangles) – 4 (red squares) /10 (all squares) = 0.4 or 40%. So, even in cases where no votes are changed, this metric can give vastly different measures of the "excess vote" for Trump.

Use an inappropriate null distribution or do not consider one at all

Plotting data in different ways can result in all sorts of patterns, some of which may appear strange or even "anomalous" at first. However, the fact that a pattern is unexpected to its maker does not mean the underlying reason for the pattern is something out of the ordinary. Unexpected patterns can arise when completely ordinary data are plotted in unintuitive ways. If the pattern arises even in "null" cases when nothing out of the ordinary is occurring, it is not logical to conclude that the existence of the pattern indicates a deviation from the ordinary.

One such pattern is the presence of a relationship between the above-described "excess Trump vote" metric and the percentage of ST Republicans (or Democrats) in a precinct. A former US Senate candidate, Shiva Ayyadurai, used the apparent negative correlation between these quantities to argue that it showed that Trump underperformed more in "more Republican" precincts. A reproduction of his original analysis using updated data for Oakland County, Michigan, is shown in Figure 3. Ayyadurai argued that the existence of this pattern was indicative of widespread vote tampering, claiming that this negative, linear trend in the right-hand side of the figure could only arise if votes were being transferred from Trump to Biden in precincts that are highly Republican.

This claim raises the question of what we might expect this plot to look like in the absence of vote tampering. Informally, what is our null? We can explore that in two ways. First, it suffices to notice that this analysis is effectively plotting $y - x$ against x, where x and y are both percentages (i.e., x and y are constrained to values between 0 and 100). In this setting, the range of possible values the function can take is constrained to the grey area shown in Figure 4(a), page 32. Although one could create many different patterns within this area, the constraints imposed by displaying the data in this way will force either a downward trend line over some portion of the region or a constant line at $y = 0$.

Another approach is to use a simulation model to get a sense of the range of typical curves under a simplified no-vote-tampering scenario. We present one such model here. Consider a simple setting in which precincts have p_R per cent Republicans and $p_D = 100 - p_R$ per cent Democrats. Suppose that in precinct i, the proportion of Republicans who choose to vote ST Republican is s_{Ri} and the proportion of Democrats who choose to vote ST Democrat is s_{Di}.

Of those Republicans and Democrats who choose to vote for individual candidates, the proportion who "defect" and vote against their party is given by d_{Ri} and d_{Di}, respectively. Figure 4(b)–(d) shows the patterns that emerge when plotting data simulated from this model as was done in Figure 3. The red lines show the pattern if the proportion of each party's voters choosing ST and the proportion defecting are exactly the same in all precincts, namely, $s_{Ri} = s_R$, $s_{Di} = s_D$, $d_{Ri} = d_R$, and $d_{Di} = d_D$. The black points are drawn such that $s_{Ri} \sim N(s_R, v)$, $s_{Di} \sim N(s_D, v)$, $d_{Ri} \sim N(d_R, v)$, $d_{Di} \sim N(d_D, v)$, truncating such that none of the draws are less than 0 or greater than 1, as they represent proportions. When v is small, as it is here, this corresponds to a simulated scenario in which all precincts are similar with respect to the ways in which each party's voters choose to vote, but all precincts are not exactly the same. Figure 4(d) uses parameter settings that are roughly consistent with the percentage of Democrats and Republicans

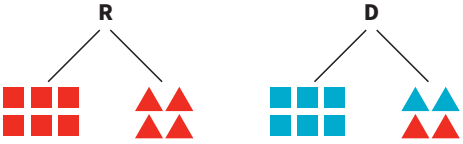

Figure 1: Hypothetical example of excess voting metric. Democrats (blue) and Republicans (red) have equal numbers of straight-ticket voters (squares). Among individual-candidate voters (triangles) Trump has a two-vote lead. The proposed excess Trump metric would score this precinct as +25% (= 75% IC Trump – 50% ST Republican).

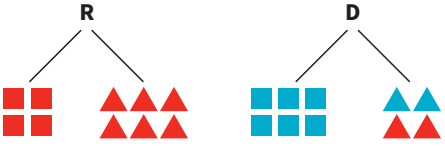

Figure 2: As in Figure 1, except two voters for Trump no longer vote straight-ticket. Despite the same vote totals for each candidate and proportion of Republican/Democrat voters, the excess Trump metric here would indicate +40% rather than +25%.

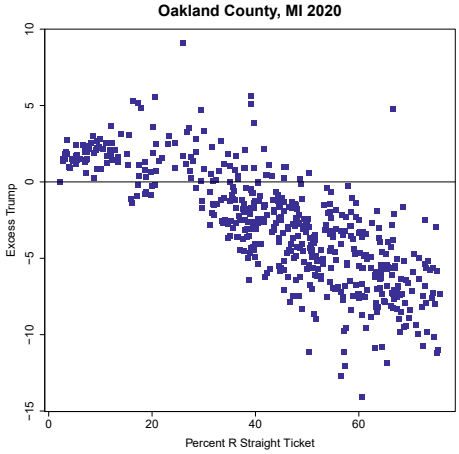

Figure 3: Plot of the "excess Trump vote" metric against the percentage of straight-ticket voters who voted Republican (i.e., voted for Trump) in Oakland County, Michigan. Each point represents one precinct.

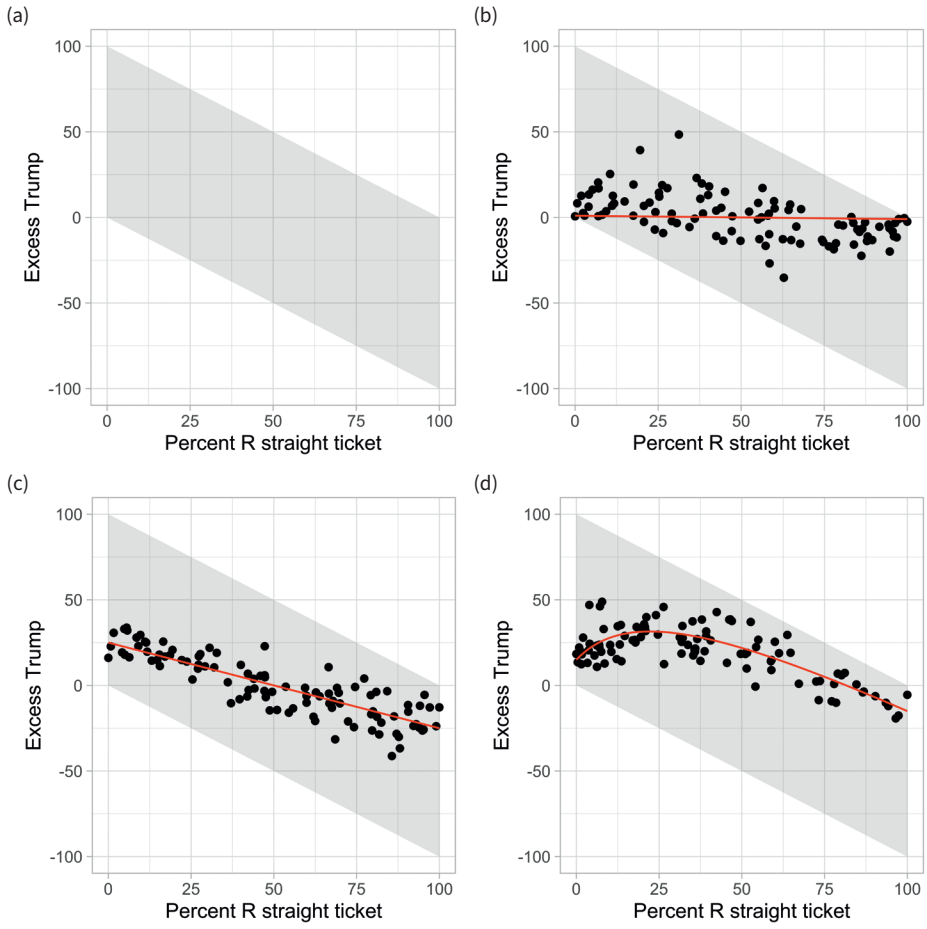

Figure 4: (a) The area to which data are constrained. (b) A realisation of our model where the proportion of straight-ticket voters for each party is 40% ($s_R = s_D = 0.40$) and the proportion of defecting voters among individual-candidate voters in each party is 1% ($d_R = d_D = 0.01$). (c) Another realisation, this time where the proportion of straight-ticket voters for each party is again 40% ($s_R = s_D = 0.40$) but the proportion of defecting voters among individual-candidate voters differs between the parties: Republicans 25%, Democrats 20% ($d_R = 0.25$, $d_D = 0.20$). (d) A final realisation where the proportion of straight-ticket Republican voters is 40% ($s_R = 0.40$), the proportion of straight-ticket Democrat voters is 65% ($s_D = 0.65$), and the proportion of defecting voters among individual-candidate voters in each party is 15% ($d_R = d_D = 0.15$).

who indicated they would "defect" in a national poll, 8% and 11%, respectively (bit.ly/3iHW333). Notably, the pattern that emerges in Figure 4(d) closely resembles the real data shown in Figure 3.

Under this simple model in which, proportionately, only small numbers of people defect from their party, we find that this method for displaying the data can lead to curves of different types (a concave-down parabola, a line with a negative slope, a flat line, etc.). What these analysts have claimed to be anomalous patterns of electoral fraud are not only plausible outcomes but expected ones. In fact, qualitatively, the patterns uncovered in the real election data are particularly unremarkable – they indicate precincts that are fairly homogeneous in their propensity for ST voting and small differences in the rate at which individuals choose to defect from their party's candidate.

Notably, the parameters d_R and d_D could represent the proportion of either defecting voters or stolen votes. These two possibilities are inherently confounded, making this analysis fundamentally unable to provide evidence of vote-switching. Based on this analysis alone, there is simply no reason to claim the more unlikely scenario of vote tampering (rather than defections) generating this pattern. All other evidence outside of this type of misleading data analysis points towards the absence of widespread voter fraud or vote rigging. Further, given that 95% of ballots in the USA leave a paper trail, any switching of votes would be caught with standard auditing (nyti.ms/3cKIVGn).

The combination of high-quantity and low-quality data results in a treasure trove of raw material from which to craft misinformation

Conclusion

Ensuring votes are counted fairly is crucial to maintaining trust in our democratic systems and accepting the outcomes of our elections. Data analysis to test and verify the integrity of our elections is of the utmost importance. However, elections produce vast quantities of data, with official tallies lagging behind estimates produced by news organisations and counties for their election night reporting. The combination of high-quantity and low-quality data results in a treasure trove of raw material from which to craft misinformation. Even earnest statistical analysis of these data sets would be challenging, as it requires a careful approach with well-reasoned expectations and metrics that incorporate a tapestry of complex events and demographic patterns that give rise to the data. Myriad sources of uncertainty (epistemic and aleatoric) threaten to generate apparent anomalies.

For less scrupulous data scientists, however, these irregularities can be held up as evidence of widespread fraud. Unconstrained by the need to carefully consider the generative process, they can use inappropriate but familiar distributions and models alongside whatever convoluted metric they prefer. Paired with a nation of precinct-level electoral data, researcher degrees of freedom become limitless. Anyone looking for evidence of electoral fraud will find it if they sufficiently relax their definition of evidence. ■

Disclosure statement
The authors declare no competing interests.