Analysis of Voting Data from the Recent Venezuela Referendum[1]

Edward W. Felten
Dept. of Computer Science
Princeton University

Aviel D. Rubin
Dept. of Computer Science
Johns Hopkins University

Adam Stubblefield
Dept. of Computer Science
Johns Hopkins University

September 1, 2004

This paper and related information are available online at http://www.venezuela-referendum.com/.

1. Introduction

On August 15, 2004, Venezuelans voted in a national referendum, to decide whether to recall President Hugo Chávez from office. Afterward, CNE, the Venezuelan electoral authority (which supports Chávez), announced that roughly 58% of voters had voted against the recall, so that Chávez would stay in office. The election was performed using Smartmatic electronic voting machines that tally votes electronically and produce a voter verifiable paper backup of each vote. Opposition figures have accused the CNE of election fraud.

Some of the fraud allegations can be evaluated only by a careful investigation and audit of election procedures and documents. We are not in a position to perform such an investigation, but we hope that others can do so. Other fraud allegations, however, would – if they were true – create statistical anomalies in the reported election results. This subset of fraud allegations can be evaluated by statistical analysis.

This paper describes our statistical analysis of the detailed election results reported by the CNE. The dataset includes the vote count from 19,055 electronic voting machines. We received these data from Sumate, an opposition group that asserts that the data match the official CNE results. The CNE had originally posted per-machine data on the web, but the per-machine data were removed, with only per-polling-place data remaining on the CNE website, on about August 18. (See http://www.cne.gov.ve/resultados/.) Our dataset is available at http://www.venezuela-referendum.com/.

Our analysis did not seek to produce any particular result, but was designed to shed light on the technical question of whether the asserted anomalies justify a claim of fraud. We take no position regarding Venezuelan politics in this paper, and we urge readers to accept or reject our statistical analysis on its merits rather than trying to second-guess our motives.

Opposition figures have claimed that the reported election results contain statistical anomalies that would not have arisen by chance in an honestly conducted election. This paper considers whether the patterns reported by the opposition really do indicate fraud.

We emphasize that there are types of election fraud that would not create statistical anomalies and hence could not be detected by statistical analyses such as ours. For example, if electronic voting machines were programmed to change each Yes vote into a No vote with 10% probability, statistical analysis could not detect such fraud. (It could probably be detected, though, by manually recounting paper ballots.) Our results, at most, can shed light on whether certain types of fraud occurred; but an analysis like ours cannot rule out fraud altogether.

[1] Este articulo esta disponible en espanol en http://www.venezuela-referendum.com/paper_sp.pdf.

2. Election Logistics

According to our data, the election was conducted at 4,582 polling places. Most polling places used electronic voting machines, but a few used manually counted paper ballots. Because the fraud claims we are evaluating involve voting machines, we ignore the polling places that used manual counting, and consider only the 19,055 electronic voting machines.

Each polling place had one or more tables, with one or more voting machines at each table. Voters were assigned to polling places based on where they lived, so we would expect the voters at different polling places to differ somewhat in demographics and political views. However, we are told that within a single polling place, a voter was assigned to a table and a machine in a fashion that was effectively random, independent of the voter's demographics and political views.

3. Allegations of Fraud

Opposition figures assert that the reported results contain various statistical anomalies that would not have arisen by chance in an honestly conducted election. Such anomalies, if they exist, could have arisen because of election fraud, or they could have arisen due to unintentional errors in the conduct of the election or bugs in the programming of electronic voting machines.

We have received numerous specific allegations of fraud. Among the ones circulating on the Internet, we have chosen to look most closely at two possibilities. The most prominent claim is that the number of Yes votes was capped on the voting machines at each polling place, so that once a certain number of Yes votes were tallied on a voting machine, that machine would count all subsequent votes as No, regardless of how the voters chose. If we assume that each machine in a polling place had the same cap, such caps could be detected by looking for polling places whose data are consistent with such a cap. In particular, we looked for polling places that had two machines with the same number of Yes votes and no machines with a higher number of Yes votes.

The other allegation that has been prominently mentioned is that two or three machines out of three machines at the same table had the same number of votes, either Yes or No in a statistically anomalous way. We examined the probabilities of these 'coincidences' of Yes or No votes within a table and compared them to the actual data from the CNE.

Our goal was to determine whether the supposed anomalies reported in the official results are consistent with the hypothesis that the election was conducted honestly.

4. Background on Statistical Analysis

Because some readers may be unfamiliar with statistical analysis, we will digress briefly to review how such analyses work and what conclusions can be drawn from them.

Hypothesis Testing

Statistical analyses like ours cannot establish the truth of the hypothesis being tested; they can only determine whether the observed data are consistent or inconsistent with the hypothesis. If the data are inconsistent with the hypothesis, this is evidence against the truth of the hypothesis. But if the data are consistent with the hypothesis, this does not establish that the hypothesis is true. If, for example, we observe that each morning a rooster crows and then the sun rises, this observation is consistent with the hypothesis that the rooster's crowing causes the sunrise; but we cannot infer that that hypothesis is correct.

To put it another way, analyses like ours operate by assuming that the hypothesis is true, and then determining the consequences that follow from that assumption. If these consequences are inconsistent with the observed facts, the inconsistency is evidence that the (assumed) hypothesis is incorrect.

In our case, we were testing the hypothesis that the election was conducted honestly. If the reported election data turn out to be inconsistent with that hypothesis, this will be evidence that something has gone wrong in the election.

Random Variation

Random processes naturally exhibit some degree of variation. For example, if we were to flip 1000 coins and count the total number of times heads came up, we know that on average we would see 500 heads. However, the odds are very strong that we would not see exactly 500 heads, but would instead see some other number close to 500.We can characterize this kind of random process by two numbers: the "mean" or average result, and the "standard deviation" which characterizes the typical amount of random variation in the result. For example, if our 1000-coin-flip experiment, the mean result is 500 heads, and the standard deviation is about 16.

A general rule of thumb, which is accurate in all of the situations encountered in this paper, says that the result of a random experiment will be more than one standard deviation away from the mean about 34% of the time, and more than two standard deviations away from the mean about 5% of the time. For example, in our coin-flip experiment, the result will be outside the range 484-516 about 34% of the time, and outside the range 468-532 about 5% of the time. If we widen the interval further, the chance that the result is outside our widened interval get ever closer to zero.

Suppose someone told us that they had flipped 1000 coins and gotten 644 heads. This result is nine standard deviations above the mean (it is 644-500 = 144 above the mean, and 144 is 9x16). A result this many standard deviations away from the mean would happen only 0.00000000000000002% of the time by chance. We could safely conclude that this person was lying.

Avoiding the Lottery Fallacy

In the New York State Lottery's Lotto game, a player chooses six numbers between 1 and 59. Then the state randomly draws six numbers from that same range. A player wins if all six of the numbers he chose match the numbers drawn by the state. A player's chance of winning is one in 45,057,474.

On August 28, 2004, the state drew the numbers 6, 9, 34, 44, 49, and 58. The drawing of these particular numbers was a very unlikely event, which would occur only 0.000002% (1 / 45,057,474) of the time in a fair lottery. Yet the fact that this very unlikely event occurred is not in itself reason to suspect fraud. The reason for this is that there is nothing special about this particular sequence of numbers – it is only one of 45,057,474 equally unlikely events that could have occurred. No matter what sequence had been drawn that day, we could have made the same argument.

If we argued (incorrectly) that the August 28 draw showed that the New York State Lottery was corrupted, we would be in error. Our error would lie in the fact that we started with a very large set of possible patterns in the data; and chose afterward which pattern to use in our argument, based only on the result of the random draw.

To avoid the lottery fallacy, we need to argue either that we chose in advance to look for that pattern as evidence of fraud, or that there is some reason to believe that that particular pattern would be caused by some likely mechanism of fraud.

5. What We Did

Since we were testing the hypothesis that the election was conducted honestly, we began by assuming that the election was in fact conducted honestly. (We emphasize that this assumption was made only hypothetically, for the purposes of analysis.)

Under this assumption, we then held a series of simulated elections. In each simulated election, we assumed that the same voters went to the same polling places and voted in the same ways as in the data from the real election. However, within each polling place, the voters were reassigned randomly to voting machines. In a simulated election, the same number of votes was cast on each machine as in the data from the real election, but different (randomly chosen) voters were assigned to each machine. We held a total of

1238 simulated elections. The data from the real election, the results of these simulated elections, and the Python code we used to generate and evaluate the simulated elections are all available at http://www.venezuela-referendum.com/.

Assuming that, in the real election, voters within a polling place were assigned to machines randomly, or at least independently of how they voted, we would expect the data from the real election to be statistically similar to our simulated elections.

6. Analysis

To evaluate the claim that there was a per-polling-place cap on the number of 'yes' votes that could be counted on each machine, we first counted the number of polling places in which two or more machines had the same number of Yes votes and no machine in the polling place had more Yes votes. This corresponds to a situation in which the cap has been reached on at least two machines; note that it is difficult to distinguish capped machines from non-capped machines unless at least two of the machines reach the cap. We will call such polling places "cap-consistent."

In the data from the real election, we found that 190 of the polling places were cap-consistent. In our simulated elections, there was an average of 163 cap-consistent polling places per election, with a standard deviation of 12.33, a minimum of 121, and a maximum of 204. This indicates that the data from the real election are 2.1 standard deviations from the mean, so that we would expect such a result about 4% of the time. It is important to note that this is not clear evidence of fraud since some of our simulated elections had even more cap-consistent polling places, even though no machines were capped in our simulations.

To further evaluate the claim that caps existed, we looked for polling places where three or more machines all had the same number of Yes votes and no machine had more Yes votes (i.e. three machines have reached the cap). In the data from the real election, five polling places meet these criteria; in the simulated elections the average is 5.2 polling places per election. This result is consistent with the machines not being capped.

To evaluate the claim that there was an unusually high number of machines at the same table with the same number of Yes or No votes, we computed the number of such "coincidences" in both the data from the real election and our simulated elections. These statistics are shown in Table 1. While the deviation of the Yes coincidences is much higher than the deviation of the No coincidences, this does not necessarily indicate fraud for the reasons explained below in Section 7.

	Yes	No
Data From Real Election	402	311
Simulated Elections Average	360.90	317.35
Simulated Elections Standard Deviation	17.75	16.99
Standard Deviations From Mean	2.3	0.37

Table 1: Comparison of coincidences from the real election and simulated election

7. Mechanisms of Possible Fraud

As mentioned in Section 3, we examined the allegations that there was a cap on the number of Yes votes on machines within a polling place, and that there was an unusually large number of coincidences within a table, where two or three machines had the same numbers of Yes or No votes. Clearly, a cap on the number of Yes votes could have been able to successfully bias the election towards the sitting government. However, our analysis did not uncover any evidence that such a cap existed. The number of caps fell within the range of numbers predicted by our simulations.

Our experiments did suggest that there was a slightly higher incidence of two Yes or No votes at each table. The question, then, is whether or not this related to any realistic fraud scenario. We were unable to think of a way in which someone could have cheated, to bias the election one way or the other, that would result in a slightly higher number of these coincidences. If the number of such coincidences were (say) 100 standard deviations from the expected mean, then we would have to conclude that there was something unusual going on. However, given that the slightly higher number of coincidences still falls within the values exhibited in our experimental runs, we attribute that to statistical variation rather than fraud. We would reevaluate that conclusion if someone could present us with a credible explanation for how the Yes and No coincidences could have been used to favor a particular outcome in the election. We note that there are many things one could have measured that would have resulted in statistics within a few standard deviations of the mean. The fact that this particular statistic deviates appears to be more a product of a search for arbitrary statistical anomalies (that is, a mild version of the lottery fallacy) than evidence that there was fraud in the election.

8. Summary

After the August 15 referendum in Venezuela on whether or not to recall president Chávez, opposition groups examined the polling data and made accusations of fraud due to statistical anomalies in the reported election results that they claim could not have occurred if the election were run fairly. However, our analysis of the same data, based on simulations, did not detect any statistical anomalies that would indicate obvious fraud in the election.

We emphasize that a lack of statistical evidence does not imply the absence of fraud. Rather, it rules out certain classes of fraud. In any case, the fraud that is alleged is not the type that we would expect a cheating government to employ. In particular, we believe that the forms of election fraud that are most likely to succeed, such as voting machines silently switching some fraction of Yes votes to No votes inside the computer, would not produce observable statistical anomalies.

Electronic voting is more susceptible to widespread fraud than less automated mechanisms. The fact that the opposition is highly suspicious of the outcome is due, in part, to the choice of electronic voting machines in a simple Yes/No election. While we did not find any statistical evidence for the claims of caps on the machines or other specific accusations of fraud, we are concerned that wide scale unobservable fraud is much easier to realize in electronic voting machines than in, for example, precinct based paper systems.