

In Search of the Black Swan:
Analysis of the Statistical Evidence of Electoral Fraud in
Venezuela

Ricardo Hausmann
Harvard University

Roberto Rigobón
Massachusetts Institute of Technology

September 3, 2004

*This study was requested by Súmate who also provided the databases we used. We appreciate the great information gathering effort carried out by this organization. We are equally indebted to a hard working collaborator who, because of institutional reasons, must remain anonymous. We thank Andrés Velasco as well for his useful comments. The opinions expressed in this report and the errors we may have incurred are our responsibility and do not compromise either Súmate, or the universities to which we belong.

Abstract

This study analyzes diverse hypotheses of electronic fraud in the Recall Referendum celebrated in Venezuela on August 15, 2004. We define fraud as the difference between the elector's intent, and the official vote tally. Our null hypothesis is that there was no fraud, and we attempt to search for evidence that will allow us to reject this hypothesis. We reject the hypothesis that fraud was committed by applying numerical maximums to machines in some precincts. Equally, we discard any hypothesis that implies altering some machines and not others, at each electoral precinct, because the variation patterns between machines at each precinct are normal. However, the statistical evidence is compatible with the occurrence of fraud that has affected every machine in a single precinct, but differentially more in some precincts than others. We find that the deviation pattern between precincts, based on the relationship between the signatures from the November 2003 Reafirmazo, and the YES votes on August 15, is positive and significantly correlated with the deviation pattern in the relationship between exit polls and votes. In other words, those precincts in which, according to the number of signatures, there are an unusually low number of YES votes, is also where, according to the exit polls, the same thing occurs. Using statistical techniques, we discard the fact that this is due to spurious errors in the data or to random coefficients in such relationships. We interpret that it is because both the signatures and the exit polls are imperfect measurements of the elector's intent but not of the possible fraud, and therefore what causes its correlation is precisely the presence of fraud. Moreover, we find that the sample used for the auditing done on August 18 was neither random nor representative of the entire universe of precincts. In this sample, the Reafirmazo signatures are associated with 10 percent more votes than in the non-audited precincts. We built 1,000 random samples in non-audited precincts and found that this result occurs with a frequency lower than 1 percent. This result is compatible with the hypothesis that the sample for the audit was chosen only among those precincts whose results had not been altered.

Introduction

This study presents a statistical evaluation of the results of the August 15, 2004 Recall Referendum on President Hugo Chávez's mandate. From the morning of August 16, 2004, when the CNE (Consejo Nacional Electoral) announced the results, opposition spokespersons expressed doubts about the validity of these results, and argued that an electronic fraud had been committed. These doubts have not been cleared up with the passing of time and the opposition has yet to acknowledge President Chavez's alleged victory.

In this context, Súmate requested that we do a statistical analysis to verify if the available information is compatible with the hypothesis of fraud or if, on the contrary, it rejects this hypothesis. Súmate provided the data used in this study but gave us complete autonomy over the conduct of our research.

We were informed that the presumption of fraud is based on the following elements:

1. A new automated voting system in spite of the fact that the opposition had requested a manual tally.
2. The voting machines left a paper trail by printing ballots that allowed each elector to verify that the machine had counted his vote adequately. These ballots were collected in boxes. However, the CNE did not allow the boxes to be opened and counted. Instead, it performed a so-called "hot" audit of 1 percent of the machines on the evening of the election. Moreover, the CNE decided that the number of boxes to be opened would be chosen by a random number generator program run on its own computer.
3. After a difficult negotiation, the CNE allowed the OAS and Carter Center to participate as observers in every phase of the process except for access to the central computer server that communicated with each machine in each voting precinct. No witness from the opposition was granted access to that room either. Only two people were allowed in that room until the results were ready.
4. The adopted technology allowed --in fact required-- bidirectional communication between the central servers and the voting machines. This bidirectional communication occurred. This is different from the information that was provided to opposition negotiators about the nature of the technology involved.
5. Contrary to what was initially stipulated, the voting machines communicated with the central server before printing the results in a document called Acta. This opens the possibility that the machines were instructed to print a result different from the one expressed by the voters.
6. On August 15, 2004, different organizations including Súmate, conducted exit polls in a number of precincts. To assure its quality, Súmate's poll was conducted with the assistance of the firm Penn, Shoen and Berland. Its results were radically different from official figures. The same thing occurred with the exit poll conducted by "Primero Justicia," a political party. The data-base of both surveys was given to us to conduct this study.

7. The “hot-audit” conducted at dawn on August 16, 2004 was not carried out to the satisfaction of either the opposition or the international observers. Only 78 of the 192 boxes stipulated were counted. The opposition only attended 28 counts, and the international observers were only present in less than 20.

8. As requested by the international observers, a second audit was conducted on August 18. The opposition did not participate in this audit because its conditions were not met; for example, the electoral materials were not delivered to a centralized location before choosing the boxes to be opened and there was no verification that the boxes selected had not been tampered with. Instead, the boxes were chosen 24 hours before they were opened, which in theory would give time for them to be altered. Notably, the CNE did not use the random number generator program proposed by the Carter Center, and instead insisted on using its own program run on its own computer and started with a seed defined by a pro-government member of the CNE. This raises doubts over whether the sample selected was truly a random one.

All these facts raise the possibility of an electronic fraud in which the machines printed outcomes different from the real count. This could in theory have been done through software alterations, or through electronic communications with the computer hub.

Our main findings are the following. First, the paper finds that the sample used for the audit of August 18, which was observed by the OAS and the Carter Center, was not randomly chosen. In that sample, the relationship between the votes obtained by the opposition on August 15 and the signatures gathered requesting the Referendum in November 2003 was 10 percent higher than in the rest of the boxes. We calculate the probability of this taking place by pure chance at less than 1 percent. In fact, we create 1,000 samples of non-audited precincts to prove this.

This result opens the possibility that the fraud was committed only in a subset of the 4,580 automated precincts, say 3,000, and that the audit was successful because it directed the search to the 1,580 unaltered precincts. This sheds new light on the fact that the Electoral Council did not accept the use of the random number generator proposed by the Carter Center and under these conditions one can infer why the Carter Center could not identify the fraud with the audit they observed.

In addition, we develop a statistical technique to identify whether there are signs of fraud in the data. To do so, we depart from previous work on the subject that was based on finding patterns in the number of votes per machine or precinct. Instead, we look for two independent variables that are imperfect correlates of the intention of voters. Fraud is nothing other than a deviation between the voters’ intention and the actual count. Since each variable used is correlated with the intention, but not with the fraud we can develop a test as to whether fraud is present. In other words, each of our two independent measures of the intention to vote predicts the actual number of votes imperfectly. If there is no fraud, the errors these two measures generate would not be correlated, as they each would make mistakes for different reasons. However, if there is fraud, the variables would make larger mistakes where the fraud was bigger and hence the errors would be positively correlated. The paper shows these errors to be highly correlated and the probability that this is pure chance is again less than 1 percent.

The first variable we use is the number of registered voters in each precinct that signed the recall petition in November, 2003. This clearly shows intent to vote yes in a future election but it does so imperfectly. Our second measure is the exit poll conducted by Penn, Schoen and Berland and complemented with an independent exit poll conducted by Primero Justicia. This is also an imperfect measure as it depends on potential biases in the sample, differences in the skill of the interviewer, etc. But this source of error should not be correlated at the precinct level with the one that affects the signatures. Therefore, it is very telling that in the precincts where the Penn, Schoen and Berland exit poll makes bigger mistakes is also where the number of petitioners suggests that the Yes votes would be higher.

This evidence is troubling because it resonates with three facts about the conduct of the election. First of all, contrary to the agreed procedure, the voting machines were ordered to communicate with the election computer server *before* printing the results. Secondly, contrary to what had been stated publicly, the technology utilized to connect the machines with the computer hub allowed two-way communication and this communication actually took place. This raises the possibility that the hub could have informed the machines what numbers to print, instead of the other way around. Finally, after an arduous negotiation, the Electoral Council allowed the OAS and the Carter Center to observe all aspects of the election process except for the central computer hub, a place where they also prohibited the presence of any witnesses from the opposition. At the time, this appeared to be an insignificant detail. Now it looks much more meaningful.

The structure of the paper is as follows. First we describe the evidence coming from the exit polls. We show that the difference between the exit polls and the actual vote is not caused by a sampling error, due for example, to an over-representation of anti-Chavez precincts, but instead to a generalized but variable difference, precinct by precinct. Second, we test the validity of the popular so-called “topes” hypothesis. According to this theory, machines were ordered not to surpass a certain maximum number of Yes votes. If this was the case, there should be an unusually large number of repeated Yes totals in each precinct and the repeated number should also be the maximum Yes vote total in the precinct. We do not check whether the number of repeats is unusually high, but we do show that the frequency with which the repeated number is also the maximum Yes vote of the precinct is consistent with a random event.

We then move on to study whether the variance of results at the precinct level is unusual. This would be the case if some but not all machines were manipulated at the precinct level. We find the variance at the precinct level to be if anything smaller than would be expected by pure chance.

The next section develops our test for fraud using our two independent but correlated measures of voters’ intent. We then move on to test whether the sample used for the audit of August 18 was random. The final section concludes.

Exit Polls vs. Votes: Analysis of the Differences

The first evidence of potential irregularities in the election count derives from the exit polls conducted independently by Súmate and Primero Justicia (PJ). As shown in Table

1, according to the CNE, 41.1 percent of voters supported the YES. On the other hand, in the Súmate and PJ surveys, the weighted projections were 62.0 and 61.6 percent respectively, a difference of more than 20 points.

We check whether this difference is due to the fact that the sample chosen by Súmate and Primero Justicia was not representative of the electoral universe. In other words, we check whether the problem arises because of an over-representation of precincts in favor of the YES vote in relation to those in favor of the NO. We show that this is not the source of the problem. As shown in Table 1, according to the CNE the percentage obtained by the YES in the precincts surveyed by Súmate was 45.0 percent, while in PJ’s sample the result was 42.7 percent. In other words, in the sample chosen by both organizations, the result reported by them differs from the official one by more than 17 percentage points. Hence, the difference in the results is not principally due to the sample composition but to a systematic difference across the sample where the exit polls were conducted.

Table 1: Comparison between Electoral Results and Súmate’s and Primero Justicia’s Exit Polls

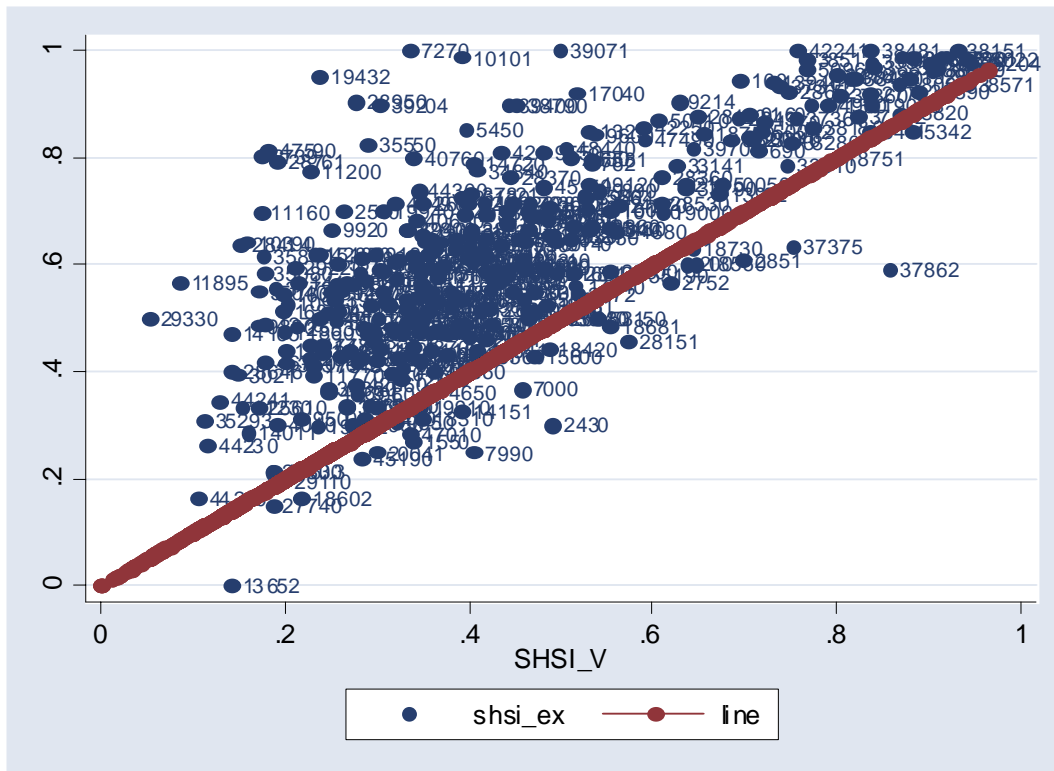
	Unweighted	Weighted
Percentage of YES votes at the precinct level	37.0%	41.1%
Percentage of YES in Súmate’s exit poll	59.5%	62.0%
Percentage of YES votes where Súmate did their exit poll	42.9%	45.0%
Percentage of YES in PJ’s exit poll	62.6%	61.6%
Percentage of YES votes where PJ did their exit poll	42.9%	42.7%
Percentage of YES in Súmate+PJ exit polls	61.3%	62.2%
Percentage of YES votes where Súmate+PJ did their Exit polls	43.1%	44.2%

To illustrate this problem more clearly, in Figure 1 we show the percentage of votes and the survey results for the 340 precincts surveyed by both groups. If the surveys were perfect, the points would align in a ray from the origin with a 45 degree slope (drawn in the graph). In other words, where the YES option received respectively 10 percent, 50 percent or 80 percent, the surveys would show the same result. If the points in the graph are above the 45 degree line, it means that the poll overestimates the result in that precinct. If the points are below, the poll underestimates it.

As Figure 2 clearly shows, the bulk of the 342 precincts polled are above the 45 degree line. Moreover, the graph indicates that the differences between the votes and the surveys are very variable among precincts. The distances to the 45 degree line are largest in places where the YES option garnered between 20 and 40 percent.

This analysis has the following implications. First, it indicates that the difference between the surveys and the votes is not due, in any important way, to problems in the selection of the precincts to be included in the survey. Second, the analysis implies that the difference may be due to one of the two reasons, or to a combination of both. It may be due to a generalized failure in both surveys in each precinct, or to a quite generalized and non-linear manipulation of the results. It will be a challenge of the statistical work to distinguish between these two hypotheses and investigate which is the right one.

GRAPH 1 Exit polls vs. Electoral result: percentage of the YES by Precinct



The Caps or “Topes” Hypothesis

The fraud hypothesis most discussed in Venezuela has been based on the idea that numerical caps were imposed on the amount of YES votes that could be allowed in a precinct and that the overflow of YES votes would be switched into NO votes. In this section we evaluate this hypothesis.

To analyze the feasibility of this hypothesis we examine how many times the number of YES and NO votes are repeated at the precinct level in the CNE’s database, which contains 19,062 automated machines.

Table 2 . Number of YES and NO total votes per machine that are repeated in the same precinct

Variable	Number of machines	Numbers repeated	Frequency
Si	19,062	1,875	9.8
No	19,062	1,472	7.7

The repetition of the YES count occurs with a frequency of 9.8 percent while that of the NO occurs with a frequency of 7.7 percent. We do not test whether this frequency is

unusually high or low¹. However, the relatively high frequency is at least in part due to the fact that the number of electors as well as the voting percentage tends to be very similar among machines in the same precinct. The fact that the repeated YES totals occurs with a slightly higher frequency than the NO is at least in part due to the fact that YES has a lower percentage of votes. Let us illustrate this point with an example. Suppose the preference for the YES in a single precinct is approximately 40 percent and the number of voters at each machine is 100. A 5 percent variation would imply 2 votes, so the expected result in each machine could be between 38 and 42. The result could be in some of the 5 numbers included in that interval. On the contrary, the same percent variation for the NO would yield a variation between 57 and 63 votes, which gives 7 possible numbers. Since the amount of possible numbers is higher for the NO than for the Yes, it is logical the latter would repeat less frequently.

More importantly, the cap hypothesis implies that the number that repeats itself is also the maximum from the precinct and that the difference is assigned to the NO. For this, it is necessary that the repeated number also be the maximum YES vote in the precinct. We study this hypothesis in Table 3.

If the repeated number was randomly distributed, it would occur with a frequency equal to $1/(\text{Number of machines} - 1)$. For example, in the case of precincts with 2 machines, the repeated number is simultaneously the maximum and the minimum, for there is only one number. In the case of three machines, the probability that the repeated number is the maximum is 50 percent. As we see in Table 3, 66 is not very far from being half of 124. In the case of 5 machines, 54 is not far from being one fourth of 198.

We conclude that if there was fraud, this was not done through the imposition of numerical caps to the YES votes in the machines of a precinct.

¹ Jonathan Taylor from Stanford University has argued that it is unusually high. See <http://www-stat.stanford.edu/~jtaylo/venezuela/>

Table 3. Maximum and non-maximum numbers repeated per voting tome at the precincts.

Machines per precinct	Non-maximum	Maximum	Total
2	0	64	64
3	58	66	124
4	161	80	241
5	144	54	198
6	230	46	276
7	221	46	267
8	197	14	211
9	151	4	155
10	97	8	105
11	85	2	87
12	52	2	54
13	36	0	36
14	18	0	18
15	20	0	20
16	7	0	7
17	6	0	6
18	6	0	6
Total	1,489	386	1,875

Variance Analysis of the Within-precinct Results

The caps hypothesis, if true, would also affect the percentage difference in the results of the machines belonging to the same precinct. This is due to the fact that the amount of voters per machine varies due to differences in the abstention rate or in the number of electors assigned to each machine. This variation would show in the number of No votes, and therefore would create a source of variation in the results across machines of the same precinct. This hypothesis and any other hypothesis that is based on the idea of altering some machine more than others at the precinct level can be tested.

In each precinct, voters are distributed to machines according to the last two digits in their identity card (*cédula*) number. This allows each machine to be a random sample of the precinct's voters because the last digits in their identity card are not correlated with any variable relevant to the voting decision. This limits the possible distance between the results from two machines from the same precinct. To illustrate this, consider how opinion surveys are done in any country. A random sample is chosen - usually of a thousand or two thousand people - and the outcome is used to predict the results of millions of voters. In other words, a representative sample composed of a miniscule fraction of the electorate is used to predict the outcome of the whole. In the case of a precinct we are taking a much smaller and homogeneous universe than a country and we are dividing the population randomly according to the number of machines in the precinct. For example, in the case of a precinct with five machines, each machine represents approximately 20 percent of the total population of the precinct. In addition, in the case of this referendum, the options were limited to two: YES or NO. This

imposes a condition for the standard deviation of the number of votes per machine. Suppose that in a machine, N- number of people vote and the probability that each one of them votes YES is p. Probability theory requires that the standard deviation follow a binomial distribution and be equal to:

$$\text{Standard Deviation} = \sqrt{p(1-p)N}$$

To illustrate this, take the case in which p is the probability that an elector will vote YES in a given precinct, is equal to 50 percent and N is 400. In this case, the standard deviation would be 10 votes. The coefficient of variation (or the standard deviation of the percentage vote) would be 10 divided by 400, meaning, 2.5 percent. Given this, the typical deviation among machines in the same precinct must be compatible with this rule.

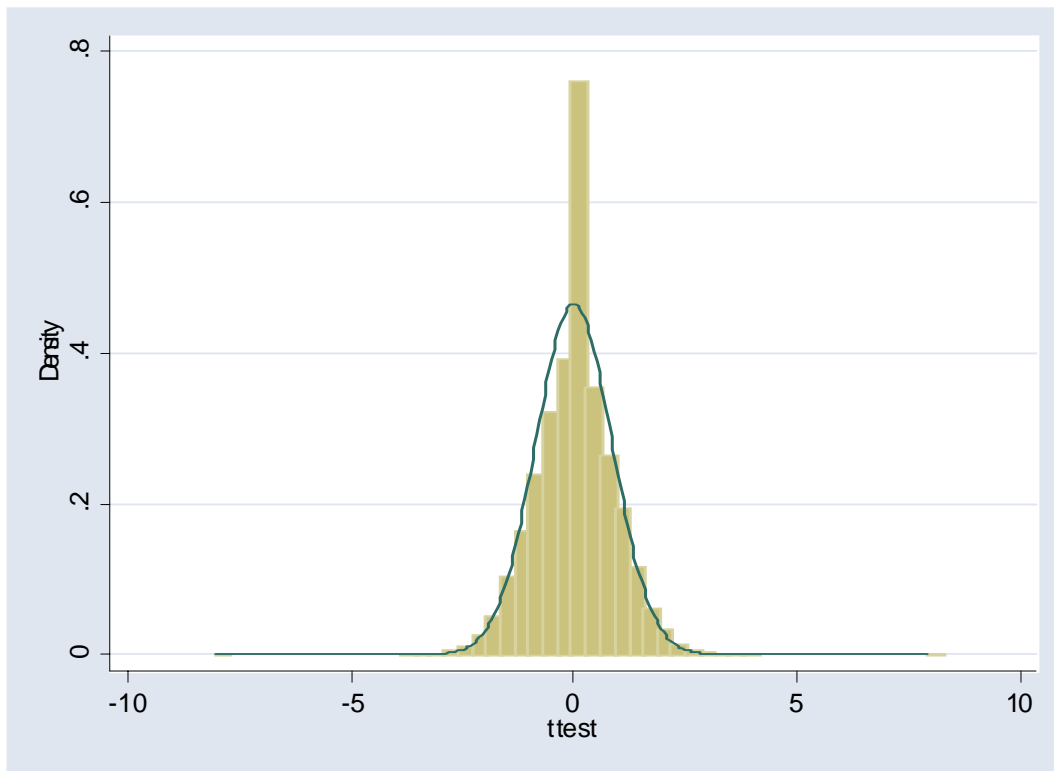
If for example within a precinct the results of some machines were changed by 10 percent while the others were left unaltered, then we would see an increase in the deviation among all machines that would be 4 times the expected standard deviation of 2.5 percent. This would be abnormal.

One implication of this result is that the caps or “topes” theory would also violate the expected distribution of a binomial. If numerical caps were assigned to each machine in a precinct, the variation of the number of voters per machine would affect the number of NO votes and therefore alter the percentage results in a manner that would increase the dispersion of the results and cause these to violate the binomial rule.

To verify if the CNE vote data complies with the standard deviation predicted by probability theory, we calculate each machine’s deviation with respect to the average of its precinct. Moreover, we divide this number by the standard deviation that would correspond to a precinct with the actual number of voters and machines. Figure 2 presents our results. It shows a histogram of the percent difference among machines of the same precinct with respect to the standard deviation expected by the binomial distribution. The curve reflects the expected theoretical distribution. The bars are the frequency calculated with the actual data. As can be seen, the coincidence is quite substantial. The graph indicates that only close to 1 percent of the machines have deviations higher than 2 times the expected standard deviation. This frequency is consistent with the theoretical distribution. In fact, if there is anything surprising about the graph, it is that the deviations of the results are if anything too small, as can be seen by the large concentration of results near zero variation.

This result has two possible interpretations. One is that there was no fraud. The other is that if fraud was committed, it must have been done by changing every machine in the precinct by a similar percentage. In fact, a fraud of this kind would not be detected with the analysis done so far for it would not alter the variance results among machines. Any hypothesis of fraud that does not comply with this condition would violate the restriction imposed on the deviation of the results by the binomial distribution.

Figure 2. Distribution of the deviation of results of machines relative to the precinct mean
(relative to the predicted standard deviation)



A Statistical Strategy to Detect the Presence of Fraud

To detect if the data is compatible with the presence of fraud we need to develop a model and put it to the data. We define fraud as the difference between the voters' intent and what the electoral system registered about his decision.

$$\text{Votes} = \text{Intent} + \text{Fraud} = I + F$$

We will take as our null hypothesis the assumption that there was no fraud, i.e. that $F = 0$. We will then develop a test to see if the null hypothesis can be rejected. The problem is that we cannot observe the voters' intent directly. The statistical strategy we adopted begins with finding two sets of independent variables that are correlated to the voters' intent, but not with the fraud. For our purposes, it is not too important that our variables do not predict the voters' intent perfectly. Even if they do so imperfectly, it may still give us a chance to reject the hypothesis of no fraud. Notice that the worse the quality of the data, the harder it will be to reject the null hypothesis meaning that bad information makes it harder, not easier, to reject the hypothesis of no fraud.

To illustrate what we do, we start with a simplified presentation of our approach. In practice, things are a bit more complicated, but explaining the sources of complexity will be easier after the fundamental intuition is presented.

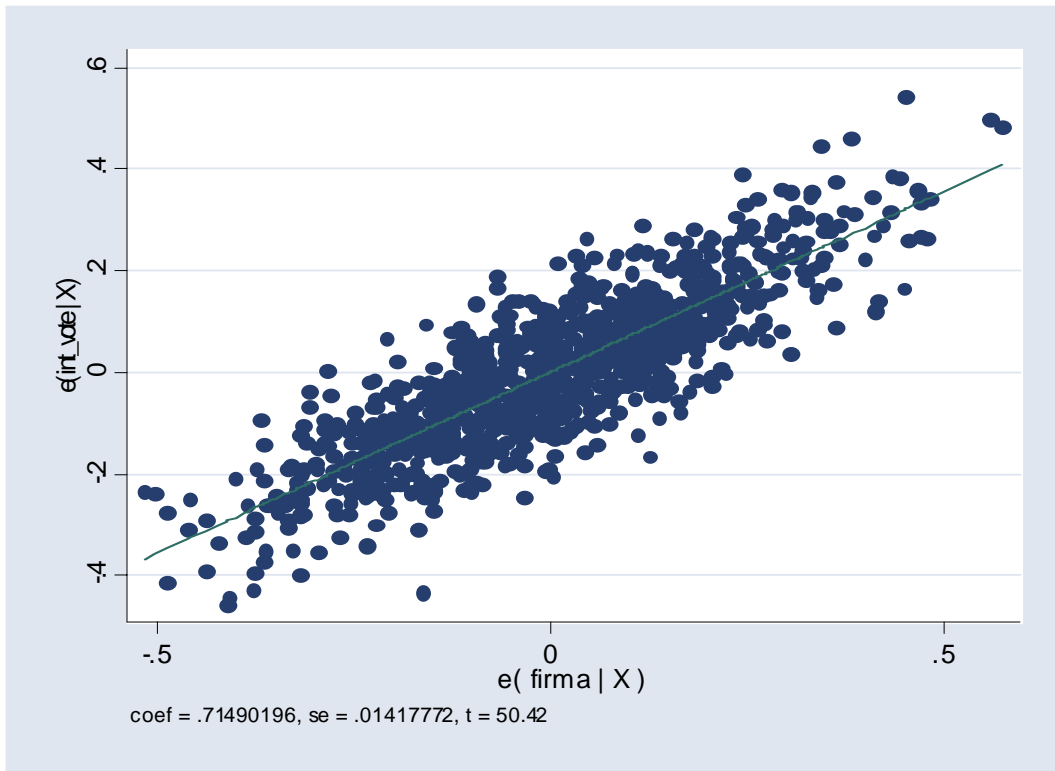
Let us take two variables that are correlated to the elector's intent: the November-December 2003 Reaffirmazo drive for signatures for the Recall Referendum petition and the exit polls. Each one of these variables is an imperfect measure of the voters' intent on August 15, 2004. Some people that signed the petition may have changed their opinion in the intervening months. Others might have decided not to sign because it was not secret, but may have decided to vote given its secrecy. Others may not have been registered in November and hence could not sign, but were registered by August and hence could vote. The lines in the August election were particularly long and slow and that may have reduced the number of voters, etc.

Equally, exit polls are an imperfect measure of the voter's intent. Pollsters may have, consciously or unconsciously, gathered a biased sample. People may have had more or less willingness to cooperate with the interview, etc. However, these errors are of a quite different nature from the errors generated by the relationship between signatures and votes and hence should not be correlated.

Suppose we have an imperfect measure of the voter's intent in each precinct and we build a graph relating this variable – say the signatures— and the voters' intent. As the signatures are an imperfect measure of the voters intent, the graph will look like a cloud of dots around some basic relationship (Figure 3-a)². Regression analysis can identify the line that relates the signature with the voters' intent. The real relationship is 0.7, because that is how we built the data. The estimated relationship using the simulated data 0.71 +- 0.014, as is indicated by the graph.

² This graph was built with simulated data using a random number generator. The data was created supposing that each signature generates 0.7 votes with an error normally distributed between +0.1 and -0.1.

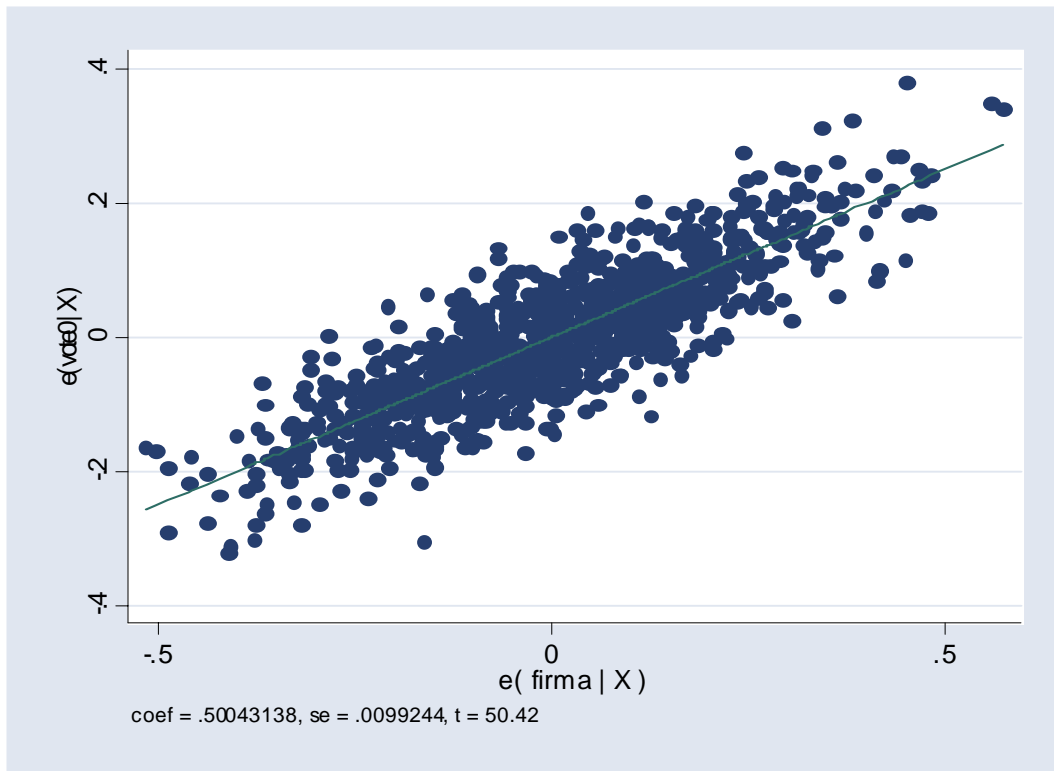
Figure 3a Simulated relationship between signatures and voters' intent



We cannot observe the voters' actual intent but the votes registered, and these, in theory, could be influenced by fraud. Suppose fraud takes place and it is directly proportional to the numbers of votes in that precinct. For example, let us suppose that fraud is committed by multiplying the total number of Yes votes in a machine by 0.7 and the difference added to the No votes.

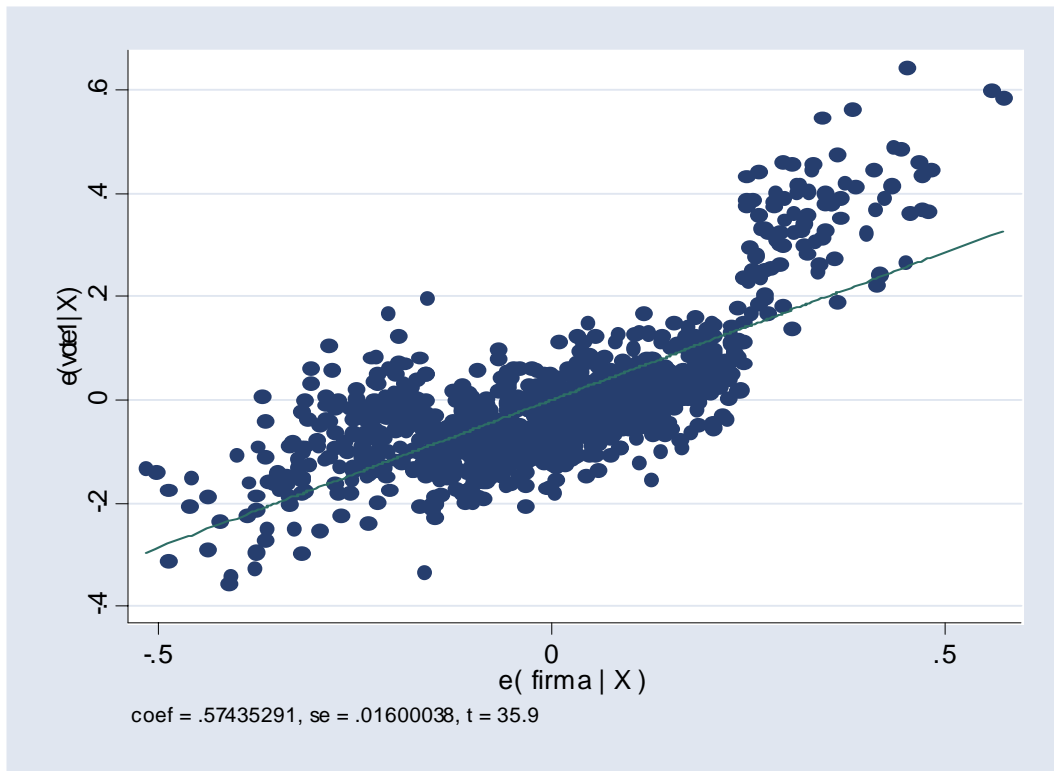
Figure 3b illustrates this case. In this case, the estimated slope is no longer 0.7 but 0.5. In addition, the pattern of errors – that is to say, the distance with respect to the regression line — looks similar. It reveals no evidence of fraud. If fraud were committed this way, we would be unable to detect it with our method. In fact, a fraud that reduces a fixed percentage of Yes votes across all machines would practically be impossible to detect by purely statistical methods. It could only be detected using another source of information such as counting the paper ballots.

Figure 3b Simulated ratio between signatures and votes with fraud proportional to 30 percent of the Yes votes.



Now, suppose the fraud was not committed in a proportional manner. For example, suppose it was committed in some precincts and not in others. Specifically, suppose fraud consists of eliminating 30 percent of the Yes votes in precincts where signatures were less than 30 percent or more than 70 percent of the registered voters. In this case, the pattern of errors will have a peculiar shape, as shown in Figure 3c. This peculiarity is not due to the imperfect nature of the number of signatures as predictor of votes, but in the fraud.

Figure 3c. Non proportional fraud



What happens if we now use a second measure of the voters' intended vote, for example the exit polls? This is also an imperfect measure of the voters' intended vote and as such when doing a regression analysis, this will generate some errors. Nevertheless, if there is a non-proportional fraud, this will also generate an irregularity in the errors which will look similar, i.e. will be correlated with the errors in the other relationship. A positive and significant correlation would identify non-proportional fraud.

Note that each measure - signatures and exit polls - is imperfect. Nevertheless, what make each of them imperfect are factors different and independent from each other. The exit poll is not influenced by the abstention rate, as people are interviewed after they vote. The signatures do not depend on the ability or bias of the interviewer. People could have changed their minds between November and August, but people do not change their minds for the same reason between the act of voting and the exit interview. Signing is a public act and voting is secret, etc. Therefore, errors made by each measure may be larger or smaller but they should not be correlated. Nevertheless, if there is non-proportional fraud, it will influence each of these measures in the same way. Hence, the errors made by both should be positively correlated. This is the essence of the method we used.

Instrumental Variable Approach

In this section we derive formally the technique we use. In particular, we show that for a variety of increasingly complex assumptions about the nature of the fraud the

covariance between the errors of the instrumental variables regression is an appropriate test of the absence of fraud.

Assume that the fraud is defined as the difference between the votes for SI actually collected and an unobservable variable that is the intention of voting of the voters that showed up. We define the first one as V_i , the intention of voters as X_i , and the fraud as F_i .

$$V_i = X_i + F_i$$

There are also two additional measures of the intention of voters: the exit polls (E_i) and the signatures (S_i) in each of the precincts. These measures, however, are imperfect. We assume a very general form of that imperfection – a random coefficient model. However, to make the point clear we start with a simpler form of errors and then generalize them. Assume that

$$\begin{aligned} E_i &= a * X_i + \text{epsi} \\ S_i &= b * X_i + \text{eta}_i \end{aligned}$$

Where we are assuming that the exit polls are possibly a biased estimate of the intention to vote: a can be smaller than one. The signatures (S_i 's), as well, could be a biased measure. Both equations have an error (epsi and eta_i) that take into account the fact that both the exit polls and the signatures are very imperfect measures of the voter's intentions – even the biased measured intentions. We assume that these errors are uncorrelated among themselves and with the fraud.³

How can we detect the fraud? The fraud can only affect the actual votes, not the exit polls, nor the signatures. In other words, the fraud is a displacement of the distribution of votes that is not present in the other two measures. Statistically, this means that the fraud could be detected by using the exit polls and the signatures as predictors of the voting process and analyzing the correlation structure of the residuals. Under the assumption that all residuals are uncorrelated – which makes sense given the definitions we have adopted – then the correlation of residuals is an indication of the magnitude of the fraud.

The particular procedure used to detect the fraud is the following:

1. Estimate the regression of V_i on E_i plus controls and recover the residual. This residual has two components: the fraud and the errors in variables residual due to the fact that the exit polls are noisy.
2. Estimate the regression of V_i on S_i plus controls and recover the residual. This residual has two components: the fraud and the errors in variables residual due to the fact that the signatures are an imperfect measure of the intention of voters.

Notice that these two residuals are correlated. First, because both have the fraud

³ This is a reasonable assumption considering that the signatures were collected at different times and conditions than the exit polls.

as an unobservable component, and second, because the right hand side variables are correlated and there are errors in variables in the regression.

3. Estimate the regression of V_i on E_i plus controls using S_i as an instrument. Recover the residual. Notice that in our model, because ϵ_i is uncorrelated with ϵ_{Si} and F_i , we can use S_i as an instrument to correct for the error in variables.
4. Using the same logic estimate V_i on S_i plus controls, and using E_i as the instrument. Recover the residual.

In this case, because the two coefficients are supposed to have solved the problem of error in variables the residuals can only be correlated if there is a common component – which in our case is the definition of the fraud.

This procedure actually detects how important the fraud is. The next section first explains why this procedure indeed is able to identify the fraud. After that we also analyze the possibility that the fraud is correlated with the signatures – which is likely given what we have argued about the stochastic properties of the votes per machine and precinct. Finally, we present evidence.

OLS estimation with no correlation between fraud and intention to vote

Running the OLS regression of Votes on Exit Poll is:

$$\begin{aligned}V_i &= X_i + F_i \\E_i &= a \cdot X_i + \epsilon_i\end{aligned}$$

Where

$$X_i = (1/a) \cdot E_i - (1/a) \cdot \epsilon_i$$

Substituting in the voting equation

$$V_i = c_1 \cdot E_i + \psi_1$$

Where

$$\psi_1 = F_i - (1/a) \cdot \epsilon_i$$

In this model, the OLS coefficient is

$$c_{1ols} = a \cdot \text{var}(X_i) / (a^2 \cdot \text{var}(X_i) + \text{var}(\epsilon_i))$$

which is always smaller than $1/a$ which is the true coefficient. This means that the residual from the regression (ψ_1) is

$$\psi_1 = F_i + (1/a - c_{1ols}) * E_i - (1/a) * \epsilon_i$$

We can do the same thing for the signatures. Notice that everything is symmetric so the equations are almost identical.

$$\begin{aligned} V_i &= X_i + F_i \\ S_i &= b * X_i + \epsilon_i \end{aligned}$$

Which means that

$$X_i = (1/b) * S_i - (1/b) * \epsilon_i$$

Substituting X_i in the V_i equation

$$V_i = c_2 * S_i + \psi_2$$

Where

$$\psi_2 = F_i - (1/b) * \epsilon_i$$

In this model, the OLS coefficient is

$$c_{2ols} = b * \text{var}(X_i) / (b^2 * \text{var}(X_i) + \text{var}(\epsilon_i))$$

which is always smaller than $1/b$ – it is only equal to $1/b$ when the variance of ϵ_i is zero. The outcome is that the residual will be

$$\psi_2 = F_i + (1/b - c_{2ols}) * S_i - (1/b) * \epsilon_i$$

Notice that the two residuals are correlated. Under the assumption that ϵ_i and ϵ_i are uncorrelated, and also uncorrelated with the fraud there are two components that create the correlation among these residuals: the fraud, and the errors-in-variable bias.

$$\text{cov}(\psi_1, \psi_2) = \text{var}(F_i) + (1/a - c_{1ols}) * (1/b - c_{2ols}) * \text{cov}(E_i, S_i)$$

The first term is the variance coming from the fraud, while the second term comes from the variance due to the error-in-variables that is present in both E_i and S_i . Notice that we are assuming that the errors in variables are independent. The covariance arises because the error-in-variables downward biases both coefficients ($c_{1ols} < 1/a$ and $c_{2ols} < 1/b$) and because the exit polls and the signatures are correlated.

Instrumental variables with no correlation between fraud and intention to vote

Under our assumptions, we have an easy solution to the error in variables in both regressions. Notice that ϵ_i and ϵ_i are uncorrelated and that ϵ_i is uncorrelated with F_i . Additionally, E_i and S_i are correlated because both measure the same factor (X_i). This means that S_i can be used for instrumenting E_i and E_i for instrumenting S_i . The outcome is as follows:

$$V_i = c_1 * E_i + F_i - \epsilon_i$$

The IV estimate is

$$\begin{aligned} c_{1iv} &= \text{cov}(S_i' V_i) / \text{cov}(S_i' E_i) \\ c_{1iv} &= b * \text{var}(X_i) / a * b * \text{var}(X_i) \\ c_{1iv} &= 1/a \end{aligned}$$

which means that the residual is

$$\psi_{1i} = F_i - (1/a) * \epsilon_i$$

Notice that now the errors-in-variable component has disappeared. Similarly, if we run the regression for votes on signatures and using the exit polls as instrument we find:

$$V_i = c_2 * S_i + F_i - (1/b) * \epsilon_{2i}$$

The IV estimate is

$$\begin{aligned} c_{2iv} &= \text{cov}(E_i' V_i) / \text{cov}(E_i' S_i) \\ c_{2iv} &= \text{var}(X_i) / b * \text{var}(X_i) \\ c_{2iv} &= 1/b \end{aligned}$$

which means that the residual is

$$\psi_{2i} = F_i - (1/b) * \epsilon_{2i}$$

The correlation between the residuals of the two IV regression is

$$\text{cov}(\psi_{1i}, \psi_{2i}) = \text{var}(F_i)$$

So, a simple test is to compare these two covariances, and determine if they are statistically different. Furthermore, if the covariance of the IV residuals is different from zero, then we have an estimate of the importance of the fraud.

Correlated Fraud

OLS estimation with correlation between fraud and intention to vote

The previous exercise has assumed that the fraud is uncorrelated with the signatures, but as we have argued in the previous section, this is unlikely. In fact, most probably, the fraud is correlated with the signatures. Let us repeat the previous exercise allowing for any covariance structure between the fraud and the signatures. Running the OLS regression of Votes on Exit Poll we obtain the same result as before:

$$\begin{aligned} V_i &= X_i + F_i + f * S_i \\ E_i &= a * X_i + \epsilon_i \end{aligned}$$

Where the residual of the voting equation has the independent term F_i and the part of the fraud that is correlated with the signatures ($f \cdot S_i$). In this model, the OLS coefficient is the same as before

$$\begin{aligned} c1_{ols} &= \text{cov}(V_i, E_i) / \text{Var}(E_i) \\ &= (a \cdot \text{var}(X_i) + a \cdot f \cdot \text{cov}(X_i, S_i)) / (a^2 \cdot \text{var}(X_i) + \text{var}(\epsilon)) \\ &= (a \cdot \text{var}(X_i) + a \cdot f \cdot b \cdot \text{var}(X_i)) / (a^2 \cdot \text{var}(X_i) + \text{var}(\epsilon)) \end{aligned}$$

which now we can't be sure is smaller than $1/a$ as before. This depends entirely on the sign of f . However, if f is negative (as we will mostly argue in this paper), then the bias downward is stronger than in the pure case. This means that the residual from the regression (ψ_1) is

$$\psi_1 = F_i + (1/a - c1_{ols}) \cdot E_i - (1/a) \cdot \epsilon_i + f \cdot S_i$$

For the signatures model

$$V_i = c2 \cdot S_i + F_i - (1/b) \cdot \eta_i + f \cdot S_i$$

Where all the three terms in the right hand side are part of the residual. The OLS coefficient is

$$\begin{aligned} c2_{ols} &= \text{cov}(V_i, S_i) / \text{Var}(S_i) \\ &= (b \cdot \text{var}(X_i) + f \cdot \text{var}(S_i)) / (b^2 \cdot \text{var}(X_i) + \text{var}(\eta)) \end{aligned}$$

where the additional term in the numerator is coming from the correlation structure between the signature and the fraud. A plausible assumption is that the fraud is usually a negative variable (in our specification) which we could expect to be larger the larger the signature is. This means that the coefficient, f , is likely to be negative. This means that the bias in $c2_{ols}$ is downward and stronger than just with errors in variables.

The residual is

$$\psi_2 = F_i + (1/b - c2_{ols} + f) \cdot S_i - (1/b) \cdot \eta_i$$

Notice that the two residuals are correlated as before, but now there are two additional terms.

$$\begin{aligned} \text{cov}(\psi_1, \psi_2) &= \text{var}(F_i) + (1/a - c1_{ols}) \cdot (1/b - c2_{ols}) \cdot \text{cov}(E_i, S_i) \\ &\quad + (1/a - c1_{ols}) \cdot f \cdot \text{cov}(E_i, S_i) + f^2 \cdot \text{var}(S_i) \end{aligned}$$

Instrumental variables estimation with correlation between fraud and intention to vote

What is the implication for this correlation when we do the instrumental variables approach? For the first equation we have:

$$V_i = c1 \cdot E_i + F_i - (1/a) \cdot \epsilon_i + f \cdot S_i$$

The IV estimate is

$$c1iv = cov(Si'Vi) / cov(Si'Ei)$$

$$c1iv = (b*var(Xi) + f*var(Si)) / b*a*var(Xi)$$

$$c1iv = 1/a + (f/ba)*var(Si)/var(Xi)$$

which means that the residual of the IV regression is

$$psi1 = Fi - (1/a)*epsi + f*Si + (1/a - c1iv)*Ei$$

It is easy to show that $c1iv$ is closer to $1/a$ than $c1ols$, which means that the residual, $psi1$, has a component coming from the error in variables that is smaller than from the OLS regression.

The IV regression of the Si specification is as follows:

$$Vi = c2*Si + Fi - (1/b)*etai + f*Si$$

The IV estimate is

$$c2iv = cov(Ei'Vi) / cov(Ei'Si)$$

$$c2iv = a*var(Xi) / a*b*var(Xi)$$

$$c2iv = 1/b$$

Notice that in this case the estimate of the IV in the voting on signatures is exactly the true coefficient. Exactly what we obtained in the previous section. Why? Simply because the exit polls do not have the error component coming from the fraud. The fraud – which is the residual in the voting equation – cannot be correlated with the exit polls or its innovations. Signatures, on the other hand, are correlated with the fraud. This is indeed the assumption of how the fraud was performed.

This makes exit polls a good instrument for signatures, but signatures is not a good instrument for exit polls. The residual in the second IV regression is:

$$psi2 = Fi - (1/b)*etai + f*Si$$

Comparison of the covariance

Let's compare the two covariances: the covariance for the OLS residuals with the covariance with the IV residuals. The OLS residual have

$$cov(psi1, psi2)_{OLS} = var(Fi) + (1/a - c1ols)*(1/b - c2ols)*cov(Ei, Si)$$

$$+ (1/a - c1ols)*f*cov(Ei, Si) + f^2*var(Si)$$

while the covariance for the IV estimates is

$$psi1 = Fi - (1/a)*epsi + f*Si + (1/a - c1iv)*Ei$$

$$psi2 = Fi - (1/b)*etai + f*Si$$

$$cov(psi1, psi2)_{IV} = var(Fi) + f*(1/a - c1iv)cov(Ei, Si) + f^2*var(Si)$$

First, notice that as before, if there is no fraud the covariance of the IV residuals should be zero. Furthermore, this last covariance reflects different forms of fraud. If the fraud is a random variable shifting the distribution (or equivalently that $f=0$ and $F_i < 0$) the covariance is the same as before:

$$\text{cov}(\psi_1, \psi_2)_{IV} = \text{var}(F_i)$$

if the fraud is not introduced as a random variable but as a shift in the distribution correlated with the signatures ($f < 0$ and $F_i = 0$) then the covariance of the IV residuals is

$$\text{cov}(\psi_1, \psi_2)_{IV} = f(1 - c_{1iv})\text{cov}(E_i, S_i) + f^2 \text{var}(S_i)$$

Only if $F_i = 0$ and $f = 0$ will produce a zero covariance of the IV residuals. In reality, if there is a fraud, probably both aspects will enter and the covariance is a combination of the two.

The next question is, what is the direction of the change in the covariance, from OLS to IV?

$$\text{cov}_{OLS} - \text{cov}_{IV} = (1/a - c_{1ols}) * (1/b - c_{2ols}) * \text{cov}(E_i, S_i) + (c_{1iv} - c_{1ols}) * f * \text{cov}(E_i, S_i)$$

where the two terms are easily signed. Let's start with the first term. We know that the error in variables together with a negative f implies that both OLS estimates are downward biased. We also know that a reasonable set of assumptions imply that signatures and exit polls are positively correlated. Hence, the first term is a multiplication of three positive elements. Let us turn our attention now to the second term. We know that c_{1iv} is closer to $1/a$ than c_{1ols} . This means that the term in brackets is negative, and we have been analyzing only the case in which f is negative. Hence, the covariance of the OLS residuals has to be larger than the covariance of the IV residuals.

Notice that if f were positive we could not have made this claim. And there would be circumstances in which the covariance actually goes up after instrumenting.

In the empirical section we implement this strategy by calculating the covariance both using OLS and instrumental variables and observe a reduction in the covariance, as would be consistent with the presence of f being negative (i.e. against the Yes vote). In addition, we find that cov_{IV} is positive and statistically significant, which is informative of the fact that f is significantly different from zero.

It is important to remember that in this procedure we are allowing the exit-polls and the signatures to be imperfect measures of the actual votes. Not only do we allow them to be noisy, but we also allow them to be biased. So, our results will NOT depend on the fact that the mean of the exit polls is different from the mean of the votes. Our test is one in which the errors in the predictions should be different.

Results

We ran a regression between votes and signatures (together with other variables which we will discuss shortly). That is to say, we calculated the best line that runs through the cloud of dots between votes and the explanatory variables that we used resembling graphic 3. Then we recovered the errors that this regression or line makes. We did the same with the ratio between votes and exit polls and recovered these errors. Then, we analyzed (to see) if these two sets of errors compared favorably.⁴

In actual fact, we do things in a slightly more complex way. We included other variables in our analysis that also influence the number of votes. These are the number of new voters and the rate of voting abstention in each precinct. The new voters were unable to take part in the Reafirmazo as they had not been previously registered with the REP (Permanent Voters Register). The more new voters there are, the greater the number of votes there should be. Now then, the percentage of Yes votes could increase or diminish according to the difference in political preferences of the new voters with respect to those registered previously. As in the previous case, the abstention rate obviously reduces the number of votes and is able to do so in a differentiated manner between the Yes and No options.

Furthermore, we have to decide which kind of line to use in the regression. There are several options: a straight line, a geometric ratio or a ratio between percentages. In other words, we can compare votes with signatures (linear); the votes logarithm with the signatures logarithm (geometric); or the percentage of Yes votes against the percentage of signatures divided by voters. Although for technical reasons, we preferred the logarithmic method,⁵ we nevertheless ran our analysis according to all three (above) methods to see if our results depended on the functional method we applied.

An example of the estimates we made using the logarithmic method is shown in Schedule 4. The estimated equation is:

$$LSI = a + b * LFIRMA + c * elc_now * d VEL + error$$

where LSI is the logarithm of the number of Yes votes; LFIRMA the logarithm of the number of signatures in each precinct; elc_now is the percentage of new voters; VEL is the percentage of voters participating; and where a, b, c and d are parameters to be estimated. Schedule 4 shows the results of our estimates for the 342 (voting) precincts for which we also have exit polls, using the most conventional method: the squared minimums.

⁴ An error or difference is defined as the difference between the regression line at a certain dot and the observed value which corresponds to it. Graphically, this is the vertical distance between the dots in graphics 3a, 3b and 3c and the charted regression line.

⁵ Given that the voting precincts are of very different sizes the linear method creates problems of heteroelasticity (i.e. absolute errors tend to be much larger in the larger precincts which implies that they are not normally distributed).

Table 4. Estimate of the equation between votes and signatures, new voters and voters participating

Source	SS	df	MS			
Model	185.800888	3	61.9336295	Number of obs =	342	
Residual	5.84296339	338	.017286874	F(3, 338) =	3582.70	
				Prob > F =	0.0000	
				R-squared =	0.9695	
				Adj R-squared =	0.9692	
				Root MSE =	.13148	
Total	191.643852	341	.56200543			

LSI	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LFIRMA	.9942821	.0099034	100.40	0.000	.974802	1.013762
elc_now	.4604462	.0375	12.28	0.000	.3866834	.5342089
VEL	.3311808	.0813913	4.07	0.000	.1710835	.4912781
_cons	.3059669	.0782436	3.91	0.000	.1520611	.4598727

The estimate allows us to explain 97 percent of the variation in votes among (voting) precincts. It estimates parameters a, b, c and d with great precision. Specifically, a is the constant, estimated at 0,306. Parameter b is the elasticity between signatures and votes and is estimated at almost 1 (in reality it is 0.994). This implies that if a precinct has twice as many signatures as another, it obtains on average twice as many votes. Parameter c is the elasticity of the Yes votes with a view to variations in the percentage of new voters. It is estimated at 0.46, which means that if the number of voters in a precinct increases by 100 percent, the Yes votes would increase by 46 percent. Parameter d is the elasticity of the number of Yes votes compared to a change in the voters participating and is estimated at 0.306, which indicates that a 10 percent increase in the rate of voters participating would cause a 3.06 percent increase in the number of Yes votes.

This equation does not indicate the actual ratio between the voters' intended vote and its explanatory variables, but between the latter and the votes recognized by the CNE. As in Graphic 3b, the possible presence of fraud influences the estimated coefficients, biasing the slopes downward, and in part is found in the error term.

The second equation we estimated was the ratio between votes and exit polls also for the 342 precincts for which we have data. The equation we estimated is:

$$LSI = f + g * lex_si + h * VEL + j + error$$

Where lex-si is the number of Yes votes which the poll for this precinct predicts. The letters f, g, h and j are parameters while LSI and VEL have already been defined. The results appear in Table 5.

Table 5. Estimate of the relationship between votes and the exit polls

Source	SS	df	MS	Number of obs = 342		
Model	157.862978	3	52.6209927	F(3, 338)	=	526.51
Residual	33.7808737	338	.099943413	Prob > F	=	0.0000
				R-squared	=	0.8237
				Adj R-squared	=	0.8222
Total	191.643852	341	.56200543	Root MSE	=	.31614

LSI	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lex_si	.9701892	.025357	38.26	0.000	.9203118	1.020067
elc_now	-.6612884	.0868377	-7.62	0.000	-.8320987	-.490478
VEL	.4244489	.1957766	2.17	0.031	.0393549	.8095429
_cons	.0722736	.2086177	0.35	0.729	-.3380789	.4826261

Again, the equation explains a large part of the variance of the votes logarithm (82%). The estimated elasticity of the voting intentions according to the polls is 0.97.

These estimates could also be biased downwards by the presence of fraud. However, the error term would reflect in part not only the imperfection of the instruments used but also the presence of fraud.

The strategy then is to analyze the correlation between the errors in both equations. This correlation is 24%, which is surprisingly high. This does not permit us to reject the fraud hypothesis. In other words, in those places where the signatures are proportionally wrong in the sense of predicting more YES votes than those obtained, the exit polls also overestimate relatively more the obtained votes. Since both measurements are independent, the implication is that what they have in common is the fraud.

Table 6 Analysis of the relationship between the errors in the equations using minimum squares.

Covariance	$9.3 * 10^{-3}$
Covariance Typical Deviation	$2.8 * 10^{-3}$
T-Student on the covariance	4.1
Probability different from zero	0.999
Correlation	0.24

This is the first result consistent with the fraud hypothesis. Formally, we can say that we cannot reject the hypothesis that fraud was committed. The presence of this correlation indicates that there is something in common between the errors committed by the exit poll and the errors committed by the signatures and this is consistent with a difference between the elector's voting intent and the registered votes.

However, it is possible to argue that the observed correlation might be generated by two sources. One is the fact that our measurements of the voter’s intent are very noisy or imperfect and that the errors in such variables might generate problems. The second is that we suppose fixed coefficients between signatures and votes or between exit polls and votes, and that these coefficients might be random. This opens the possibility that the correlation we are finding may have been generated by other factors and not by fraud.

To discard this possibility, we applied a statistical technique called “Instrumental Variables.” The idea is that both the signatures, as well as the exit polls have errors or noise. However, this noise is independent of each other. What the variables have in common is the fact that both are related to the voter’s intent. The technique begins by using in the regression not the signatures directly, but that component in them that is related to, or in line with, the exit polls. In statistics jargon, we would say that we use the exit polls as an instrument to correct or clean the signature errors before studying their correlation to the votes. Symmetrically, we use the signatures to clean the exit polls before relating them to the votes. After having done these two regressions with instrumental variables, we take the errors of each of them and study their correlation. If the errors are positively correlated we cannot reject the hypothesis that there was fraud. The theoretical justification on this methodology for the analysis of fraud that we are conducting is discussed in detail in the technical version of this document.

Table 7 shows the same equation as Table 4, but this time it uses the instrumental variables method, using exit polls as an instrument.

Table 7. Regression between votes and signatures using exit polls as an instrumental variable.

```

Instrumental variables (2SLS) regression
-----
Source |          SS      df      MS                Number of obs =      342
-----+-----+-----+-----+-----+-----
Model   | 185.741458      3    61.9138192      F( 3, 338) = 3013.34
Residual|  5.90239422    338   .017462705      Prob > F      =  0.0000
-----+-----+-----+-----+-----+-----
Total   | 191.643852    341   .56200543      R-squared     =  0.9692
                                           Adj R-squared =  0.9689
                                           Root MSE     =  .13215
-----+-----+-----+-----+-----+-----
LSI     |          Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
LFIRMA  |  1.012645      .0110631     91.53  0.000      .9908834      1.034406
elc_now |  .4792798      .0380142     12.61  0.000      .4045055      .5540541
VEL     |  .3067645      .0820558      3.74  0.000      .1453602      .4681688
_cons   |  .1718993      .0861817      1.99  0.047      .0023791      .3414194
-----+-----+-----+-----+-----+-----
Instrumented:  LFIRMA
Instruments:   elc_now VEL lex_si
-----

```

Note that the coefficient of the signatures now slightly increases: from 0,994 in the estimate in Table 4 to 1,013 in Table 7. This is normal, as the existence of errors or noise in the data tends to lower the coefficients estimated with the method of Table 4. On cleaning or lowering the problem of errors in the data, higher coefficients are obtained usually.

Table 8 re-estimates the same equation as Table 5 but using instrumental variables. This time, the coefficient of the exit poll (lex-si) increases from 0,97 to 1,17. This is to be expected as the data of the exit polls, given their nature, have more noise than the data of the signatures, which is why the method in Table 5 skews the coefficient, lower than in the case of the signatures.

Table 8. Regression between the votes and exit polls using the signatures as an instrumental variable

```

Instrumental variables (2SLS) regression

-----+-----+-----+-----+-----+-----+-----+-----+-----+
Source |          SS          df          MS          Number of obs =      342
-----+-----+-----+-----+-----+-----+-----+-----+-----+
Model  | 151.228444          3  50.4094815  F( 3, 338) = 517.96
Residual | 40.4154074        338  .119572211  Prob > F      = 0.0000
-----+-----+-----+-----+-----+-----+-----+-----+-----+
Total  | 191.643852        341  .56200543  R-squared     = 0.7891
                                           Adj R-squared = 0.7872
                                           Root MSE     = .34579

-----+-----+-----+-----+-----+-----+-----+-----+-----+
LSI |          Coef.      Std. Err.      t      P>|t|      [95% Conf. Interval]
-----+-----+-----+-----+-----+-----+-----+-----+-----+
lex_si | 1.176787      .030827      38.17  0.000      1.11615      1.237424
elc_now | -.6829967     .0949936     -7.19  0.000     -.8698498     -.4961437
VEL | .1627794     .2148175      0.76  0.449     -.2597683     .5853271
_cons | -1.523351     .250735     -6.08  0.000     -2.016549     -1.030153

-----+-----+-----+-----+-----+-----+-----+-----+-----+
Instrumented:  lex_si
Instruments:  elc_now VEL LFIRMA
-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

On studying the relationship between the errors in the variables generated by these two equations, we obtain the data presented in Table 9. The analysis shows that even after using the method of instrumental variables to correct problems of errors in variables and random coefficients, the correlation between errors generated using signatures and those generated using exit polls diminishes only from 0,24 to 0,17 and remains significantly different from zero.

Table 9. Analysis of the relationship between errors in the 2 equations used to estimate the number of votes: Minimum squares for vs. (against) Instrumental Variables

Item	Minimum squares method.	Instrumental Variables Method
Covariance	$9.3 * 10^{-3}$	$7.7 * 10^{-3}$
Typical distortion	$2.8 * 10^{-3}$	$2.5 * 10^{-3}$
Probability different from zero	0.999	0.991
Correlation	0.24	0.17
T-Student on covariance	4.1	3.1

Our initial hypothesis was that if there had been a not perfectly proportional fraud, this would have generated a pattern of errors that would have caused a positive correlation between both variables. We found this positive correlation, which allows us to reject the hypothesis that there was no fraud. The exit polls, in spite of their imperfections, tend to have a greater margin of error in those places where the signatures also have a margin of error. On using the instrumental variables method, we have discarded that the correlation is due to problems of error in the variables or in the random coefficients. This would be the type of imprint left by a not perfectly proportional fraud.

Our strategy has consisted in utilizing two sources of information related to the voters' intended vote but not to the possible fraud. If we use these sources or variables in estimating the votes imperfectly, then the residue or term of error will contain not only the imperfections of our sources but also a component associated to the fraud. Our interpretation is that as the imperfections are independent one from another and the residuals are correlated, this is because of the presence of a common factor, i.e., fraud.

The Audit

Any hypothesis of fraud requires an explanation of why the audits that took place did not find any foul play. While the first audit carried out in the wee hours of the morning of August 16 failed, the audit conducted on August 18, if it was well carried out, should have settled the issue. The audit was based on opening 150 randomly selected ballot boxes, which contain the original paper ballots checked by the voters and which thus reflect their real intended vote. If these boxes were not tampered with and if they really are a random sampling of the universe of precincts, the audit should rule out any presumption of fraud. So, how could fraud have taken place, if the audit did not find it? It should be pointed out that any hypothesis of fraud, which involves changing hundreds of ballot boxes would constitute a conspiracy involving a large number of participants and hence would be more likely to be revealed through disloyalty.

One hypothesis is that fraud was not committed in all precincts but only in a fraction of them. To give an example, suppose that out of the 4,580 automated precincts used in the election, 3,000 precincts were altered but the rest were not. Let us further suppose that the unaltered 1,580 precincts were picked at random. This implies that they would represent a balanced sample of the country from a regional and social point of view. The same would be true of the 3,000 precincts in which the results supposedly had been altered. One reason to do things this way is that it was known (beforehand) that ex post

audits would be carried out and that a number of precincts would be checked. To accommodate this, they would have to be unaffected by fraud.

Note that if fraud is committed in some precincts and not in others then it will not be perfectly proportional and the method used in the previous section would detect it.

If the selection of the precincts left unaffected was done in this way, this creates an important complication but also opens up a great opportunity. The complication is that the selection of the boxes to be audited could not really be random. It is critical that the selection be made among the 1,580 un-tampered precincts and not among the 3,000 tampered ones. This is only possible if one has control over the random number generator that selects the boxes to be audited. In this sense, it has to be pointed out that the National Electoral Council refused to make use of the random numbers-generating program proposed by the Carter Center and insisted on the use of their own program installed in their own computer.

The opportunity generated by this form of solving the problem of the audit is that any sample taken of the 1,580 un-tampered precincts is a representative sample of the country in the social and regional sense. This makes it more difficult to know if the sample taken was really random, as it resembles the country in all the dimensions usually associated with representativeness, such as regional or social.

To solve this problem we must develop a methodology that allows us to test if the sample taken for the audit on August 18th really is a random sample. To understand the problem more clearly, let us call the un-tampered precincts "fat" and the tampered ones "thin." The sample taken for the audit must be a sample of only "thin" precincts, while the rest of the precincts are a mixture of "fat" and "thin." If we could "weigh" the audited precincts, we would be able to see that on average the un-audited precincts are "fatter."

The problem is that we need to develop a methodology that can test whether the audited precincts weigh as much as the others do or whether on the contrary they have a statistically different body frame.

The method we suggest is as follows. There exists a theorem in statistics that states that if a ratio applies to an entire unit, any random sampling of the same must have the same properties. If we estimate the ratio of the universe of un-audited precincts and estimate another for the audited ones, the second cannot be statistically different from the first. Otherwise, it would not be a random and representative sample.

To implement this strategy we again made use of our model that correlates signatures, voter participation rates and new voters with the number of actual votes cast. We estimated this ratio based on the universe of 4580 precincts and we looked at the obtained coefficients. We then estimated them separately between the audited precincts and the un-audited ones and determined if the coefficients are statistically different.

To see if the results are different and to calculate the statistical significance of the difference, it is useful to estimate the equations in the following way:

$$\text{Votes} = a + b * (\text{vector of explanatory explaining variables}) +$$

$c * d *$ (vector of explanatory explaining variables)

where a, b and c are parameters to be estimated and d is a "dummy" variable worth 1 if we are dealing with audited precincts and 0 if we are dealing with un-audited ones. The boxes belong to the same random distribution if the parameters c are not different from zero. The explainable variables utilized are the number of signatures, the number of voters registered in the REP (Permanent Election Register) at the time of the Reafirmazo (petition re-signature collection drive), the number of new voters registered after the Reafirmazo and the number of voters who did not vote. We estimated the equation in logarithms.

Theoretical Considerations

Sophisticated frauds are hard to detect. According to Rubin if all the machines are affected using the same procedure then the fraud is statistically undetectable. The advantage of the Venezuelan case is that there was the possibility of audits and therefore, some of the CDV's were required to be untouched. Because the government was the one that would choose which precincts would be audited, it is imaginable that they knew which precincts were not touched and lead the auditors toward those precincts.

First, lets discuss the form of the fraud (the theory). Then, we provide the evidence.

The theoretical idea of a voting process is that it is an imperfect but unbiased measure of the popular intention

$$V_i = X_i + F_i$$

Where X_i is the voting intention in each CDV, V_i are the total votes, "i" indicates the CDV, and F_i is the fraud. It is important to indicate that if F_i is a shift with the same distribution as X_i (meaning both are normally distributed) then V_i is the sum of two normal distributions, which is also a normal distribution and it would be hard to identify any anomalies in the data. The fraud would be undetectable.

Under this assumption, in the case of Venezuela, the fraud could not be normally distributed because the system had to allow some precincts to be untouched so that they could be audited. Interestingly, the government did not allow the Carter Center to choose the precincts randomly. It was the government the one that chose the precincts. So, if there is fraud, it is imaginable that we could detect the shift in the population by comparing the audited population with the non-audited population.

We have a variable that is correlated with the intention of voters – the signatures. We assume that the signatures follow

$$S_i = b(i)*X_i + \epsilon_{i}$$

where S_i are the number of signatures, $b(i)$ is a random coefficient mapping the intention of voters to the total number of signatures, and ϵ_{i} is a random disturbance indicating the noise involved in the measurement of the intentions.

This specification allows the signatures to be a biased estimator of the voters intentions ($b(i)$ on average could be less or larger than one).

Additionally, F_i could be correlated or not with the number of signatures. We will adopt, then the following specification: $F_i + f(i)*S_i$ to encompass both possibilities.

What is the OLS estimate of V_i on S_i ?

$$V_i = c*S_i + \psi_i$$

Notice that the reduced form model is the following:

$$\begin{aligned} V_i &= X_i + F_i + f(i)*b(i)*X_i + f(i)*\eta_{i1} \\ S_i &= b(i)*X_i + \eta_{i2} \end{aligned}$$

The OLS coefficient is the covariance between these two variables divided by the variance of the signatures. These variances include not only the variances of the shocks but also the variances of the coefficients. For notational convenience assume

$$\begin{aligned} b(i) &= b + b_i \\ f(i) &= f + f_i \end{aligned}$$

where b_i and f_i conditional on X_i have mean zero and finite variance. Under these assumptions the OLS coefficient is

$$cols = \left(b(1+fb)*var(X_i) + f*var(\eta_{i1}) \right) / \left(b^2*var(X_i) + var(\eta_{i2}) \right)$$

which obviously is different from $1/b$ – which is the limit if $f=0$ and $var(\eta_{i1})=0$. There are several biases in this coefficient worth highlighting: first the error-in-variables which is the result of $var(\eta_{i1})$ being different from zero. This bias is the attenuation bias and reduces the coefficient. So $cols < 1/b$. Second, there is the bias introduced by the fraud component that is correlated with the signatures (f). we will assume throughout the paper that f is negative, if that is the case, notice that the denominator is reduced by this bias, which means that $cols$ is farther from $1/b$. In summary, both biases are working in the same direction.

Let us see how the residuals look

$$\psi_i = F_i + (1+(f(i)-cols)*b(i))*X_i + (f(i)-cols)*\eta_{i1}$$

Notice that even if the structural shocks are homoskedastic ($var(X_i)$, $var(\eta_{i1})$ and the variances of the random coefficients) the variance of the residuals is going to be heteroskedastic. There is a term multiplying the residuals that is related to the random coefficient model.

Therefore, it is not surprising that indeed the standard deviation of the residuals of this regression at the parroquia level are correlated with the predicted residuals of the regression. Part of that correlation is coming from the fraud (obviously), but also part of that correlation is the result of the random coefficient model. Let us see.

If $f(i)$ and F_i are zero, then

$$\begin{aligned} \psi &= (1 - \text{cols} * b(i)) * X_i - \text{cols} * \eta_i \\ \text{cols} &= \left(b * \text{var}(X_i) \right) / \left(b^2 * \text{var}(X_i) + \text{var}(\eta_i) \right) \end{aligned}$$

notice that unambiguously cols is smaller than $1/b$ which means that the expected value of $\text{cols} * b(i)$ is smaller than $E(b(i)/b) = 1$. Therefore the term in the first bracket of the residuals has a positive expected value and the variance of ψ will depend on X_i . This source of correlation is not interesting (in terms of fraud) and therefore, we require a better test to differentiate between a random coefficient model and one with fraud.

The idea is to compare a “supposedly random” sample with the total sample. If the sample is truly a random coefficient model, and the innovations to the coefficients (b_i and f_i) are truly orthogonal to the other shocks, then any sub-sample – any sub-sample – should have the same properties as the full sample.

This is exactly what is done when agencies collect surveys on consumption, industry production, etc. If the sub-sample is a random draw it is representative of the population, or full-sample. For instance, assume that we are interested in studying consumption patterns. We know that consumption depends on the level of income, education, race, gender, age, religion, etc. Furthermore, there is no particular reason why we have to assume that a one percent increase in income will imply the same increase of consumption to all the individuals in the sample. In other words, it is reasonable to assume that the coefficients from income to consumption are random. However, if the model is well specified (meaning that all the controls that have to be in the right hand side are there), then the coefficients are truly random and independent of everything else. If we pick a random sub-sample of the population – a representative sample – the behavior of those individuals is a good proxy of the behavior of the population. This is standard in all micro data models where always we make inference about the population by looking at a smaller sample. This is exactly what we do here.

Any sub-sample of precincts should have the same behavior as the whole. This does not mean that the coefficients estimated are going to be the same. What it does mean is that the differences cannot be statistically significant. We can, for example, choose as the random sub-sample, the precincts that the government allowed the Carter Center to audit. Why is this a good sample? Well, because we know that a shift of the full distribution is statistically undetectable, the fraud cannot only be found if we concentrate in the sub-sample that, ex-ante, has a lower likelihood of being tainted.

Therefore, we split the sample between those that were audited and those that were not audited. As is discussed in the paper, we find them to be statistically different.

Differences in the samples - in terms of their OLS estimates

As was mentioned above, the fraud introduces a bias in the regression coefficient if it is correlated with the signatures in the precincts. To clarify the exposition we show the OLS coefficient of the bivariate model with and without fraud:

$$\text{cols_fraud} = \left(b(1 + f_b) * \text{var}(X_i) + f * \text{var}(\eta_i) \right) / \left(b^2 * \text{var}(X_i) + \text{var}(\eta_i) \right)$$

$$\text{cols_Nofraud} = \left(b \cdot \text{var}(X_i) \right) / \left(b^2 \cdot \text{var}(X_i) + \text{var}(\eta) \right)$$

which implies that

$$\text{cols_fraud} = \text{cols_Nofraud} + f \cdot \left(b \cdot \text{var}(X_i) + \text{var}(\eta) \right) / \left(b^2 \cdot \text{var}(X_i) + \text{var}(\eta) \right)$$

which under our assumptions that $f < 0$ implies that the OLS coefficient of the fraud sample is smaller than the OLS coefficient of the no-fraud sample. Additionally, it should be the case that these two coefficients are statistically different from zero, because otherwise the changes in the coefficients are mainly explained by the small sample properties of OLS and not by fraud.

To test for this possibility we estimate our preferred estimation allowing for interactions of all the right hand side variables with a dummy that takes value of one when the precinct was one of the ones assigned to be audited. The regression is

$$V_i = c_2 \cdot S_i + c_3 \cdot S_i \cdot D + c_4 \cdot \text{NewElectors} + c_5 \cdot \text{NewElectors} \cdot D + c_6 \cdot \text{Participation} + c_7 \cdot \text{Participation} \cdot D + c_0 + c_1 \cdot D$$

Where we are predicting the total votes by the signatures (S_i) and the signatures interacted with the dummy for audited centros (D). We introduce several controls, obviously allowing for different constant terms in the two sub-samples ($c_0 + c_1 \cdot D$) and controlling for the increase in the universe of voters (NewElectors) and for the participation in the precinct (Participation).

The coefficients of interest are c_3 . If there is fraud (and hence there is a shift in the distribution), then c_3 should be positive and statistically different from zero. Why? As was shown before, under the assumption that the fraud reduced the number of votes for “ S_i ” the OLS coefficient in the full sample (c_2) is smaller than the true one ($c_2 + c_3$). Notice that this is what we find in the empirical results.

Results

The results are very clear, as is indicated in Table 10. The interaction term $D \cdot \text{Signatures}$ shows that the elasticity of the signatures in votes is 10.5 percent higher in the audited precincts than in the un-audited ones, i.e., the signatures collected in the audited precincts on August 18th generate 10 percent more YES votes than the rest of the precincts. The statistical value of Student's t is 2.73. The probability that this is by chance is less than 1 percent (shown with the three asterisks in the Table). The coefficient on new voters is also different with a level of confidence of 1 percent whereas the coefficient with regard to the abstaining voters is different with a level of confidence of 10 percent.

To illustrate what is unusual with this result we constructed 1,000 random samples of 200 precincts based on the universe of un-audited precincts. We estimated the same equation and calculated the statistic value of Student's t for the term $D \cdot \text{signatures}$. The result is shown in Table 11. As the Table shows, a value of said statistic higher than 2.48 occurs less than 1 percent of the time. In the sample of the audit on August 18, this value is 2.73.

Table 10. Do the audited precincts represent the (entire) universe of precincts ?

	Log SI
Log FIRMA	0.958 (129.46)***
D * LFIRMA	0.105 (2.73)***
Log Electores Reafirmazo	0.043 (4.89)***
D * Log Electores Reafirmazo	-0.126 (3.06)***
Log Electores Nuevos	0.595 (23.64)***
D * Log Electores Nuevos	0.118 (1.30)
Log Electores no votantes	-0.459 (11.47)***
D * Log Electores no votantes	0.312 (1.89)*
AUDIT	0.171 (1.51)
Constant	0.254 (9.14)***
Observations	4580
R-squared	0.97

Robust t statistics in parentheses

* significant to 10 %; ** significant to 5 %; *** significant to 1 %

Table 11. Frequency distribution of the statistic of Student t value on the parameter of signatures in 1000 regressions estimated on the basis of 1000 samples randomly taken from the un-audited precincts universe.

Percentiles		Smallest		
1%	-2.60853	-3.342794		
5%	-1.832646	-3.233441		
10%	-1.425525	-3.053542	Obs	1000
25%	-.8046502	-3.053519	Sum of Wgt.	1000
50%	-.0189599		Mean	-.0191664
		Largest	Std. Dev.	1.104314
75%	.7440667	3.232639		
90%	1.360018	3.658616	Variance	1.219509
95%	1.770322	3.975739	Skewness	.0747199
99%	2.48632	4.010863	Kurtosis	3.049892

We conclude that the data indicate that the audited precincts are statistically different from the unaudited precincts. This implies that they do not form a random sample of the entire universe of precincts (audited and unaudited). In the audited precincts, the signatures are transformed into a larger number of votes than in all of the precincts (audited and un-audited) taken together. The probability that this occurs by coincidence

is less than 1 percent. This result tends to confirm the doubts expressed as regards the reliability of the audit.

Intuition

In this section, we would like to illustrate both our theory of fraud, as well as how we test for it.

Assume that in Florida half the precincts are Republican and half Democrat. How do we know this? Well, first we have the results of the previous presidential election in each precinct, which should be a good predictor of today's preferences, and we also know how many Republicans and Democrats are registered in each precinct. Obviously, these measures are not perfect, and they are possibly biased, but they should be related. Also assume that on election day there are exit polls. Assume that these polls are extremely noisy and biased.

Assume that a fraud is going to be committed – in favor of the Republicans (just an example). How can we have a perfect fraud? In the absence of an audit, the electronic fraud is simple – when the machines connect to the central computer, the central computer sends a program that makes the machine to report 10 percent less Democratic votes, and 10 percent more Republican votes. This does not change the total number of voters but changes the proportion.

This is undetectable, statistically speaking. The exit poll and the signatures will show that there is a change in the public opinion in favor of the Republicans. The exit polls will give a different answer, but in the end, because the exit polls are so noisy, the blame will be given to the imperfection in the collection of the polls rather than use them as evidence of fraud.

The only deterrent of the fraud in this case is to have an audit, and the question is how can we achieve the fraud and at the same time pass the audit. Assume that the machines have a paper trail of each voter and some of the machines will be audited.

This is the procedure of the fraud that would be undetectable using standard statistical methods. Assume that for the fraud, half of the Democratic precincts will be converted to Republican. Let us say that 10 percent of the votes will be shifted. The result of the election is that $\frac{3}{4}$ of the centers are Republican and only $\frac{1}{4}$ is Democrat. Now, to pass the audit, the machines that will be recounted cannot belong to the set of fraudulent machines. This is why it is important to be able to control the choice of the machines to be audited. Therefore, to make the fraud pass the audit the authority draws random numbers that have $\frac{3}{4}$ weights in the Republican precincts and $\frac{1}{4}$ in the Democrat ones. Assume this is done in the morning of the election, so the authority knows in advance which precincts to leave unaffected by fraud. In the end, the audit is passed, $\frac{3}{4}$ are Republicans and $\frac{1}{4}$ are Democrat.

This simple procedure – which only requires observing the results of previous elections, or in the Venezuelan case, to observe the number of signatures and compare them to the universe of voters – will hide fraud from an audit if the precincts are not chosen in a truly random fashion. Here, the exit polls would give a different result, but again, most of the discrepancy would be blamed on the exit polls.

Two properties worth emphasizing are satisfied by this data. First, the mean of the audited sample and the whole sample would be similar. Second, the correlation between votes and the prior information is the same in the two samples. This, at a first glance could look as if this is evidence of no fraud, but that is incorrect. Remember that the correlation between two variables is unaffected if one of the variables is multiplied by a positive number. Hence, the correlation between the signatures and the votes in the Venezuelan case is exactly the same as the correlation between the signatures and 90 percent of the votes. So, a fraud of, say 10 percent, would not affect the correlation between signatures and fraud. It would however affect the coefficient, which is what we do.

Therefore, how can we detect fraud? In the audited sample, the information that existed before – the estimates of the preferences of the voters – is a better predictor of the actual votes than in the non-audited sample. For example, in the audited sample, if the precinct was Democratic in the past it has a high likelihood of being Democratic today, similarly if it was Democrat it has a high likelihood to be Democrat today. But in the non-audited sample, the problem is that this relationship is weaker. In other words, we can detect if the conditional behavior between the two samples is different, and therefore, argue that something strange is happening in the data. This is the second test we run.

The first test is one in which we compare the predicted error of the votes using the two different measures of the preferences of voting. For example, the reasons why the exit polls is an imperfect measure of the votes are different to why the results on the previous election are a bad measure. For instance, one is affected by the participation, while the other one is not; one took place several months before the other one; one is collected by the electoral committee and the other is collected by the private sector – which could possibly have a vested interest in a particular outcome, etc. The important aspect of our test is that the reasons why one of the measures is imperfect are different to the reasons why the other one is also imperfect. On the other hand, if there is a fraud, then both measures have a common reason why they are failing. This is our first test.

Conclusions

This report rejects certain hypothesis about fraud in the Venezuelan referendum of August 15 2004, but not others. We did not find empirical validity for the much-discussed hypothesis of numerical caps. We were also unable to prove any hypothesis that implies differentially tampering with the (voting) machines of the same precinct. A manipulation of this kind would alter the percentile differences in such a way as to violate the expected variance at the precinct, and would have been detected by this analysis.

All hypotheses of fraud must presuppose a similar tampering in all the machines of a precinct. If this had been done in a homogeneous manner in all of the country's precincts, none of the methods applied in this study - as well as other statistical methods - would have been able to identify it. What allows us to develop a test of the possible existence of fraud is precisely the heterogeneous treatment of the different precincts.

To carry out this test we used two imperfect, random and independent indicators of the intent to vote. Our definition of fraud consists of the existence of a difference between the voters' intended vote and the votes registered by the CNE. Our two indicators, as imperfect as they might be, are correlated with the intended vote of the elector, but not with the fraud. If both are used independently in regressions to estimate the relationship between them and the vote count, the error term or deviation will reflect not only the imperfection of the instrument applied but also the fraud. If both deviations are correlated, it shows there is a common element of deviation in both. This element is our evidence of fraud. Furthermore, to be consistent with the hypothesis of fraud, this correlation has to be positive: i.e. in those precincts where there fraud was larger, both measures would project more votes than were actually registered.

This is precisely what we find. Our two indicators are the number of registered voters in each precinct, who signed in the Reafirmazo of November 2003 and the exit polls held by Súmate and Primero Justicia on August 15th, day of the Recall Referendum. The result holds if we control for the changes in the electoral register and the abstention rate. Furthermore, the result holds up well to changes in the functional form of the ratio (linear, logarithmic, percentile). The result is not due to spurious statistical effects (errors in the variables or the possible presence of random coefficients), as they hold up when we correct for these factors using estimators based on instrumental variables.

We note that this technique identifies fraud in so far as it is carried differentially across precincts. It allows us to test for the presence of fraud, but it does not allow us to estimate its magnitude as the average fraud will be reflected in the parameters of the relationships we estimate while the differential fraud will be reflected in the error terms. We use the error terms in the identification of fraud, not the estimated parameters.

Again, any hypothesis of fraud must presuppose that the results of all the machines in the same precinct were tampered proportionally. This requires some coordination mechanism. In theory, this coordination could be in the software or in the communication with the central computer hub. For these reasons, it is useful to point out the following precedents:

- The machines had the capacity to communicate bi-directionally with the central computer server or hub and this communication took place.
- The machines communicated with the hub before printing the Certificates, which opens the possibility that they were instructed to print results different from the real ones.
- The entrance of witnesses from the opposition or of the international observers to the computer hub during election day was not allowed.

The voting system implanted in Venezuela generates voting ballots that are checked by the voter and placed in boxes, which are subject to audit in a random manner. A fraud scheme must take into account how to avoid detection during an audit.

One possibility is to leave some precincts unaffected and to direct the audit to those precincts. The choice of which precincts to affect can be done systematically or at random. This generates two kinds of precincts: those that were tampered with and those

that were not. Now, if the program that selects the boxes to be opened in an audit process can be controlled, then it will be possible to select the boxes of those precincts that were not tampered with and this sample might seem random in all aspects except as to the question of fraud.

Our analysis shows that the sample selected to carry out the audit on August 18, 2004 was not random nor representative of all the precincts. In this sample, the elasticity of the signatures compared to the votes is 10 percent higher and the possibility that this is random is significantly less than 1 percent. We repeated our analysis randomly selecting 1,000 samples from un-audited precincts and this result.

One important fact is that the CNE refused to use the random number generating program offered by the Carter Center for the August 18th audit and instead used its own program installed in its own computer and initialed with their own seed.

In conclusion, this study rejects certain hypotheses of fraud, but indicates others that are compatible with the statistical data.

In statistics, it is impossible to confirm a hypothesis, but it is possible to reject it. As Karl Popper said when observing 1,000 white swans: this does not prove the accuracy of the thesis that all swans are white. Nevertheless, observing a black swan does allow one to reject it.

Paraphrasing Popper, our white swan represents no fraud. The results we obtain make up a black swan. The alternate hypothesis that there was fraud is consistent with our results, which is why we are unable to reject it.