Forensic Analysis of the Venezuelan Recall Referendum

Raúl Jiménez

Abstract. The best way to reconcile political actors in a controversial electoral process is a full audit. When this is not possible, statistical tools may be useful for measuring the likelihood of the results. The Venezuelan recall referendum (2004) provides a suitable dataset for thinking about this important problem. The cost of errors in examining an allegation of electoral fraud can be enormous. They can range from legitimizing an unfair election to supporting an unfounded accusation, with serious political implications. For this reason, we must be very selective about data, hypotheses and test statistics that will be used. This article offers a critical review of recent statistical literature on the Venezuelan referendum. In addition, we propose a testing methodology, based exclusively on vote counting, that is potentially useful in election forensics. The referendum is reexamined, offering new and intriguing aspects to previous analyses. The main conclusion is that there were a significant number of irregularities in the vote counting that introduced a bias in favor of the winning option. A plausible scenario in which the irregularities could overturn the results is also discussed.

Key words and phrases: Election forensics, Venezuelan presidential elections, Benford's Law, multivariate hypergeometric distribution.

1. INTRODUCTION

The statistical controversies surrounding the outcomes of the Venezuelan referendum, convened to revoke the mandate of President Chávez on August 15th of 2004, generated a long spate of articles in newspapers and occupied significant television time. A Google search with the exact phrase "Venezuelan recall referendum" shows more than 100,000 hits in English. Several reports, commissioned by different organizations, reached opposite conclusions. Roughly speaking, a fraud may have occurred during the referendum or, on the contrary, was statistically undetectable. A good example of this is the work of Hausmann and Rigobon (2011), where the authors claimed to have found statistical evidence of fraud. According to experts consulted by *The Wall Street Journal*, "the Hausmann/Rigobon study is more credible than many of the other allegations being thrown around" (Luhnow and De Cordoba,

2004). However, their early claim (Hausmann and Rigobon, 2004) was later rejected by The Carter Center [(2005), Appendix 4] and by Weisbrot et al. (2004).

The first peer-reviewed article devoted to the statistical analysis of the referendum data (Febres and Marquez, 2006) concluded that there is statistical evidence for rejecting the official results. This article, in *International Statistical Review*, made no mention of the paper by Taylor (2005) which concluded explicitly that there is no evidence of fraud. Taylor's paper is the best known reference on the subject, widely covered by media; in part because he was asked to investigate the allegations of fraud on behalf of The Carter Center. Another well-known reference is a paper by Felten et al. (2004), which did not detect any statistical inconsistency that would indicate obvious fraud in the election. However, three papers in this issue of *Statistical Science* (Delfino and Salas, 2011; Prado and Sansó, 2011; Pericchi and Torres, 2011) support the claim of fraud. Who is right?

The statistical papers on the referendum can be grouped into two classes: those that only use vote counting and those that use related additional data. Five

Raúl Jiménez is Associate Professor, Department of Statistics, Universidad Carlos III de Madrid, C./ Madrid, 126 – 28903 Getafe (Madrid), Spain (e-mail: rauljose.jimenez@uc3m.es).

papers mentioned above cover the different claims of fraud investigated by the panel of experts convened by The Carter Center [(2005), Appendix 13]. These are:

(1) Discrepancy between official results and exit polls (Prado and Sansó, 2011) and unexpected correlations between computerized vote counting, the number of signatures for the recall petition and audit results (Delfino and Salas, 2011).

(2) Anomalous distributions of votes among voting notebooks (Febres and Marquez 2006; Taylor 2005), including high rates of ties (Taylor, 2005) and failure of fit to *Benford's Law* for significant digits (Pericchi and Torres 2011; Taylor 2005).

I am very skeptical about the use of data from other sources. To make a long story short, below I mention only key facts that can be extracted from the Comprehensive Report of The Carter Center:

The months previous to the referendum were highly polarized, with mass rallies for and against the government, with aggressive campaigns for attracting new voters and to intimidate and persecute both signers (people who signed for the recall petition) and supporters of President Chávez. Even the referendum day was hot. The electoral actors took ad hoc decisions that generated suspicions and lack of confidence in the whole process.

In this political atmosphere, we must assume that any unofficial information will be controversial. If there are many doubts about the official results, one cannot expect consensus with other data. Furthermore, one must be very careful with the statistical assumptions that one will use.

This article has two purposes: (1) to bring order to the ruckus caused by different statistical analyses, some of them carried out by non-experts, and (2) to examine, by a proper forensics analysis, the allegations of fraud. Section 2 reviews the referendum framework, introduces the main notation used throughout this paper and presents a critical revision of the five papers cited above. In Section 3 we propose a methodology, based exclusively on vote counting, to test the recall referendum of 2004. The presidential elections of 1998 and 2000 are also reviewed. Far from being a statistical headache, the referendum is an excellent dataset to exercise a wide variety of elementary but powerful statistical tools. Additionally, the case of study is also useful for illustrating some common mistakes in stochastic modeling. Section 4 summarizes the main findings and conclusions.

2. REFERENDUM FRAMEWORK AND CRITICAL REVIEW

The electoral process is fully described in the report of The Carter Center (2005). The crucial features for the present analysis are:

(i) *A voting center consists of one or more electoral tables and each table consists of one, two or three voting notebooks, which are the official data units with the lowest number of votes.*

(ii) *Within the time allowed, voters were registered to a center. Voters usually chose a center close to their residence or workplace, many of them long before the referendum. When the time was over, the referee decided the number of voting notebooks in a center according to the number of voters. In addition, notebooks are grouped in tables (no more than three per table), mainly for logistical reasons related to the voting process.*

(iii) *In each center, voters are randomly assigned to the notebooks.*[1]

(iv) *There were only two options to vote: YES or NO. Although there was a very small percentage of invalid votes (0.3%), there was a significant percentage of abstentions (30%).*

(v) *The voting notebooks were computerized (touchscreen voting machines which collected 86% of the valid votes) and manual (ballot boxes which represented 14% of the votes).*

Both (i)–(ii) and (iv)–(v) are simple true facts but (iii) is a *statistical hypothesis*. The secrecy of the ballot lies in the random assignment of voters to notebooks. For this reason, (iii) is essential for a fair election. Thus, we assume it is true throughout our analysis, with the exception of Sections 3.5–3.7, where we suppose there were irregularities in the allocation of voters to notebooks.

Next, let us introduce the basic notation used throughout this paper. To do so, I will use the term *polling unit*

---

[1]Every Venezuelan citizen is assigned an ID number. These numbers are assigned in sequential order by date of request. Usually, it is done when a Venezuelan girl or boy is ten years old. By this I mean that the number is independent of the entire electoral process. The ID number of the voters (older than 18 years old) has up to nine digits and, except for a case of extreme longevity, at least six digits. The mechanism to assign voters to notebooks can be described as follows: According to the last two digits, the voters were uniformly distributed to the notebooks. For example, in a center with four notebooks, if the last two digits ended between 00 and 24, then it was assigned to notebook 1. If the last two digits ended between 25 and 49, then it was assigned to notebook 2, and so on.

generically in the next three paragraphs to refer to a center or a table or a notebook.

– Let Y_i be the number of YES votes (those favoring recalling President Chávez) and N_i the number of NO votes in polling unit i.
– Let $T_i = Y_i + N_i$ be the total number of *valid votes* in polling unit i and τ_i the number of *voters* assigned to that polling unit (the *size* of the polling unit). Note the difference between *voters* and *valid votes*.
– Let $O_i = \tau_i - T_i$ be the number of invalid votes and abstentions in the polling unit i. For brevity, we refer to them as the OUT votes (out of the electoral consultation).

In the rest of this section, where we review different papers, the subscript can refer to centers, tables or notebooks. However, in Section 3 the subscripts are used only to identify voting notebooks.

2.1 Discrepancies Between Two Exit Polls and Official Results

Prado and Sansó (2011) addressed the controversial discrepancy between two independent exit polls and the official results. Roughly, the official result was 41% YES votes and 59% NO votes, while the exit poll results were 61% YES votes. The polls were collected by a political party (Primero Justicia) and a nongovernmental organization (Súmate), both opposition to president Chávez. The authors' main claims are:

C1: There was no selection bias in choosing the centers to be polled.
C2: The discrepancies per center cannot be explained by sampling errors.

C1 is settled by noting that the proportion of YES votes for the overall population matches the proportion of YES votes for the polled centers.

Claim C2 is addressed by assuming that the sampling distribution of the number of YES *answers* for a given polled center i, say y_i, is a Binomial(t_i, p_i) random variable. The parameters of this Binomial are: t_i the size of the sample collected at the center and p_i the proportion of YES votes, namely $p_i = Y_i/T_i$. Under this assumption, Prado and Sansó (2011) showed that there are significant differences between the official results and the exit polls in about 60% of the 497 polled centers. The authors also considered the pairwise comparison between the two exit polls among the common polled centers (27 in total). We remark that eight of them (30%) differ significantly.

It appears that Prado and Sansó had the following assumptions in mind to determine that y_i is Binomial with the parameters above:

A1: Given a polling center, the persons to interview were selected by *simple random sampling.*
A2: Each interviewed person responded to the question with the truth.

A careful reading of Section 2 of Prado and Sansó (2011) suggests that the sample at each center may correspond to a more complex model than simple random sampling. How could the used model affect the estimates and conclusions of their analysis? If, for example, and as seems to be, *stratified sampling* was used, it will depend on the *stratification* schema and the *allocation* criteria used by the pollsters (Lohr, 2004). In the absence of concrete information, the assumption of the binomial distribution is the most reasonable one. However, we cannot ignore the uncertainty about the model and, consequently, about the sampling errors computed under A1.

The authors discussed briefly the consequences of the non-veracity of A2. "It has been demonstrated repeatedly that non-response can have large effects on the results of a survey" (Lohr, 2004). It is quite possible that, in a highly polarized political climate, voters that supported Chávez were associated with non-response, since they could identify the pollsters as members of the opposition to Chávez. Unfortunately Prado and Sansó had no estimates of non-responses and so had to ignore their effects.

Other sources of voter selection bias and measurement error are discussed in this paper. Some of them could imply a systematic bias across the pollsters. Such is the case of the late closing of the voting centers:

The voting centers had to be open until 4:00 p.m. but the electoral umpire extended the closing time twice, first until 9:00 p.m. and finally until midnight. This was not foreseen by the pollsters and during the afternoon and evening, there was a fierce campaign to promote the attendance of the supporters of President Chávez to the voting centers (The Carter Center, 2005).

Prado and Sansó (2011) also studied this possibility, but the available data are very limited. Although the statistical procedure and motive are correct, missing data can produce results that have no validity at all (De Veaux and Hand, 2005).

It is hard to believe that the discrepancies between exit polls and official results are due to sampling and random non-sampling errors. Unfortunately the information about exit polls is limited and does not allow a more rigorous analysis.

2.2 YES Votes Versus Number of Signers in the Recall Petition

Delfino and Salas (2011) focused on the association between the YES votes and the number of signers in the recall petition.[2] In the first four sections of this paper the authors described the electoral process well. However, from the fifth section onward, I have major concerns.

Let S_i be the number of signers in voting center i. The authors considered the following two *relative numbers* of YES votes and signers:

$$(1) \qquad k_i = \frac{Y_i}{S_i} \quad \text{and} \quad s_i = \frac{S_i}{T_i}.$$

They conducted a bivariate data analysis with k as a response variable and s as an input variable. Since $k \leq 1/s$, they said: "In voting centers with a large value of s, we expect a value of k around $1 \ldots$ The situation is completely different in voting centers with a small value of s. The singularity can produce very high values of k in the neighborhood of $s = 0$. Hence the level of uncertainty in k becomes very large." Later on, they added: "The computerized centers are very far away from $1/s$, clearly contradicting the expected non-linear behavior with respect to s." Finally, they claimed fraud because the data contradict this behavior and even ventured to establish a hypothesis: "In computerized centers, official results were forced to follow a linear relationship with respect to the number of signatures."

What can justify the previous conjecture? All that we really know is that the range of k is larger when s decreases. How can we infer the *expected nonlinear behavior of k with respect to s* from this fact? As is shown in equation (4) of their paper,

$$k_i = \frac{p_i}{s_i},$$

$p_i = Y_i/T_i$ being the proportion of YES votes in center i. Then, k decreases as $1/s$, of course, but increases as p does and there is a strong relation between these two variables. In fact, as we will explain next, one expects the value of k to be constant with respect to s, not only showing that the conjecture of Delfino and Salas (2011) is false, but showing that the results observed are as expected.

Following their schema, we analyze the (full) computerized centers and (full) manual centers separately.

[2] For readers who do not know the intricacies of the referendum, the signatures were collected eight months before the referendum. Many signers were invalidated and some had to sign again in a second runoff (The Carter Center, 2005).

Manual centers are peculiar. They usually correspond to remote locations and they have a much smaller number of votes than the computerized ones (Prado and Sansó, 2011). For this reason many authors perform a separate analysis of these data. There was also a small number of mixed centers where there were both manual and computerized notebooks. These centers represent only 1.26% of the total YES votes, 1.3% of the valid votes, and are excluded in what follows.

Let $\gamma_1 = 389{,}862$ and $\gamma_2 = 3{,}548{,}811$, the total YES votes in manual and computerized centers, respectively. Consider also the total number of signers in manual and computerized centers, that we shall denote by $\theta_1 \gamma_1$ and $\theta_2 \gamma_2$, so that θ_1 and θ_2 are ratios between total signers and total YES votes. As mentioned before, I am skeptical about the use of data that are not official results of the referendum. So, we will assume θ_1 and θ_2 are unknown parameters and will only assign values to them for simulation purposes.

As The Carter Center (2005) remarked, the signers were the hard core of the YES votes. In fact, Delfino and Salas (2011) claimed that "each signature has a high probability of resulting a YES vote." Let us simplify the scenario and assume that each signature in a center was a YES vote in that center. Thus, the ratios θ_1 and θ_2 are less than 1. Under this assumption, the conditional distribution of S_i given Y_i can be fitted by a hypergeometric distribution with parameters γ_c (*the number of marbles* in the hypergeometric jargon), $\theta_c \gamma_c$ (*the number of white marbles*) and Y_i (*the number of draws*), c being equal to 1 or 2 according to whether i represents a manual or computerized center. The expected value and variance of the hypergeometric variable are

$$(2) \qquad \mathbb{E}[S_i | Y_i] = \theta_c Y_i \quad \text{and}$$
$$\text{Var}[S_i | Y_i] = Y_i \theta_c (1 - \theta_c) \frac{\gamma_c - Y_i}{\gamma_c - 1}.$$

Using the standard normal approximation one obtains

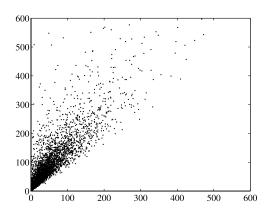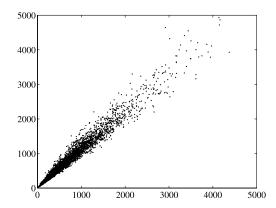$$(3) \qquad \frac{S_i - \theta_c Y_i}{\sqrt{Y_i \theta_c (1 - \theta_c)}} \approx \mathcal{N}(0, 1),$$

$\mathcal{N}(\mu, \sigma^2)$ a Normal random variable with mean μ and variance σ^2. Relation (3) leads us to consider the two heteroscedastic linear models

$$(4) \qquad S = \theta_c Y + \mathcal{N}(0, \theta_c (1 - \theta_c) Y),$$

for manual centers ($c = 1$) and computerized centers ($c = 2$).

For each center, we simulated the number of signatures given the number of YES votes at the cen-

FIG. 1. *YES votes versus simulated signatures according to the heteroscedastic linear model* (4). *The left panel corresponds to manual centers with* $\theta_1 = 1/1.81$. *The right panel corresponds to computerized centers with* $\theta_2 = 1/1.15$.

ter using (4). Typical outcomes of these simulations with $\theta_1 = 1/1.81$ and $\theta_2 = 1/1.15$ are shown in Figure 1. The values of θ_c were chosen with the intention of comparing our simulated clouds of points with those shown in Figure 6 of Delfino and Salas (2011). Note that the least squares regression lines of the latter ones have slopes 1.81 and 1.15 using a reverse relation between the variables, namely $Y = a_c S + b_c +$ error. Thus, we take $\theta_c = 1/a_c$. It is difficult to see how to reject the regression model (4) using statistical testing, even under the classical homoscedastic linear model. The differences between the clouds associated with manual and computerized centers are due to differences in scale and variances, included in the heteroscedastic linear model (4). There is nothing mysterious about this difference, as Delfino and Salas (2011) suggested. Reversing the relationship between Y and S in regression model (4) yields a heteroscedastic linear model

$$(5) \qquad Y = \beta_c S + \mathcal{N}(0, \sigma_c^2 S).$$

Dividing by S, the above equation becomes

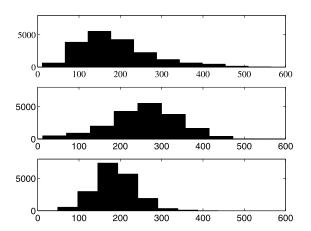$$(6) \qquad k = \beta_c + \frac{1}{\sqrt{S}} \mathcal{N}(0, \sigma_c^2),$$

which precisely describes the clouds of points shown in Figures 3 and 5 of Delfino and Salas (2011), with observations around a constant for any value of s, although the range of k is larger when s is smaller. In summary, it is expected that $\{k_i\}$ will be constant with a dispersion which decreases as $1/\sqrt{S}$ (note the difference between $S = sT$ and s). Note that, although s will be small, if T is large (like almost every computerized center), the variance can be small, explaining why computerized centers are more concentrated around the expected value of k.

There is an additional comment related to Figures 3, 4 and 5 of Delfino and Salas (2011) worth making. Note that all right panels have a gap for small values of the input variables (almost without observations). Compare the figures removing these gaps in both panels. For example, remove the windows with $s < 0.1$ in Figures 3 and 5 and the windows with less than 200 total votes in Figure 4. The behavior is very similar for manual and computerized voting centers. Their conclusion about the different behavior between manual and computerized centers seems inaccurate.

There are more intriguing statistical arguments in the paper of Delfino and Salas (2011). Although we have focused only on their main claim, I should add a comment related to the data. From the least squares regression lines shown in Figure 6 of Delfino and Salas (2011) one can estimate the total signatures in fully manual or computerized centers (excluding the mixed ones) on which the authors base their study. This total is 3,310,200, close to the 3,467,051 signatures submitted to the electoral umpire (Delfino and Salas, 2011). However, the total number of valid signers was 2,553,051 (The Carter Center, 2005). I leave the conclusion to the reader.

2.3 Anomaly Detection by Benford's Law

Pericchi and Torres (2011) compared empirical distributions with Benford's Law governing the frequency of the significant digits (Hill, 1995). Considering several electoral processes in three countries, the only case compellingly rejected by their test is the NO votes at computerized notebooks in the Venezuelan recall referendum. In addition, they made reference to recent contributions in which compliance or violation of the law in electoral processes has been studied. Some crit-
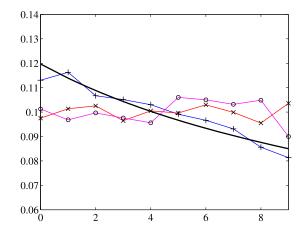
FIG. 2. *Left panel*: *Histogram of the YES votes (top), NO votes (middle) and OUT votes (bottom) per notebook. Right panel*: *Benford's Law for the second digit (solid line) versus relative frequencies of the second digit for YES votes = +, NO votes = × and OUT votes = ◦.*

icicisms related to the use of the law in electoral data (The Carter Center, 2005; Taylor, 2005) were also discussed. As theoretical contributions, the authors obtained a generalization of the law under restrictions of the maximum number of votes per polling station and discussed technical issues related to measuring the fit of the law.
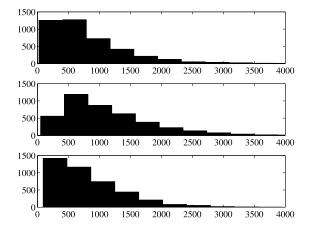
It is important to note that Pericchi and Torres (2011) did not analyze the OUT votes or abstentions. Figure 2 shows the marginal distributions of each option of vote per notebook (left panel) and compares the empirical distributions of the second digit with Benford's Law (right panel). Regarding Figure 2:

- As Pericchi and Torres showed, the YES votes conform to the law, while the NO votes do not. However, the strongest widespread departure from the law is

related to the OUT votes. The χ^2 test statistic for this option is the highest of the three.

- It is known that compliance with the law is more likely when the skewness is positive (Wallace, 2002), and the only distribution with positive skewness is related to the YES votes.

We should remark that violations of Benford's Law may be due to unbiased errors (Etteridge and Srivastava, 1999). Thus, deviations from the law can arise regardless of whether an election is fair or not (Deckert et al., 2010). On the other hand, there are many types of fraud that cannot be detected by Benford's analysis (Durtschi et al., 2004). So, electoral results that conform to the law are not neccessarily free of suspicion.

To illustrate the comments above let us consider results by centers rather than by notebooks. In Figure 3
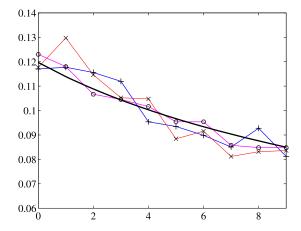

FIG. 3. *Left panel*: *Histogram of the YES votes (top), NO votes (middle) and OUT votes (bottom) aggregated by center. Right panel*: *Benford's Law for the second digit (solid line) versus relative frequencies of the second digit for YES votes = +, NO votes = × and OUT votes = ◦.*

(left panel) we show the distributions of the number of votes at this aggregation level. Note that now all distributions have positive skewness. In the same figure (right panel) we also show Benford's Law for the second digit and the related empirical distributions of vote per center. All voting options confirm the law. According to this analysis, there is no reason to doubt the official results by center, despite that the test suggests the contrary when we use the results by notebook. Is the former a false negative or the latter a false positive? Could unbiased errors in the vote counting by notebooks reproduce such a scenario? Or, conversely, could the results by centers be masking a fraud in notebooks? Benford's test does not address this controversy.

2.4 Irregularity in the YES Votes Distribution

Febres and Marquez (2006) tested the distribution of YES votes in the voting notebooks. In a first round, they applied a Z test to compare the proportion of YES votes in each notebook with the proportion from the center to which the notebook belongs. The number of *irregular notebooks* (notebooks with a proportion significantly different from the proportion of the center) resulting from this round is expected. Therefore, this analysis suggests no inconsistency. According to the territorial organization of Venezuela, the voting centers are grouped into parishes. The authors subdivided the parishes into clusters of centers, using a criterion that we discuss later. They then applied Pearson's χ^2 test to compare the distribution of YES votes among the notebooks at each cluster with the conditional expected distribution given the overall results by cluster and valid votes by notebook. In this second round, they reported a high percentage of *irregular clusters* (clusters with an outlier χ^2 statistic). Their main finding was that the irregular clusters favor the NO option. Moreover, they showed a monotone relationship between the proportion of YES votes by cluster and the p-value of the Pearson χ^2 test. Tuning the confidence level to block irregular clusters, they estimated the overall result and the winning option is YES.
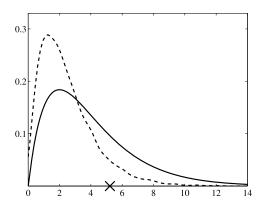
As mentioned earlier, voters within the same center were randomly assigned to the notebooks. Thus, each notebook is a random sample without replacement from the voting center population. The framework can be completely different when notebooks are grouped by clusters of centers. If the proportions of YES votes of two centers in the same cluster are not equal, no matter how similar they are, and if the total number of votes by notebooks is large enough, any

consistent test will detect significant discrepancy between the proportions in the notebooks and the proportion in the cluster. The authors took care of this fact. They made a trade-off between the homogeneity of the cluster (how similar the proportions of the centers within the cluster are) and the number of votes per notebook. Basically, the clusters were chosen such that the Z test does not detect a significant difference between the proportion of YES votes at the notebook with the greatest number of votes and the cluster proportion. In this way they ensured that each notebook is a representative sample of the cluster. The authors referred to this as the *minimum heterogeneity distance for clustering analysis* and made reference to the books of Sokal and Sneath (1973) and Press (1982).

I have two concerns about these results. The first deals with a general concern about the validity of ad hoc mechanisms to identify false positives (detecting fraud when none is present), which might be the case. The second is a technical issue that must be resolved before subscribing to the authors' conclusions.

In the referendum context, the standard cluster units of notebooks are the voting centers. I guess that the authors did not report results at this level of aggregation because they did not observe inconsistencies at this level of aggregation. In fact, if we apply Pearson's χ^2 test to detect *irregular centers*, in the same way that the authors applied this test to detect *irregular clusters*, we do not observe major inconsistencies. Therefore, their results depend on a particular way of clustering the notebooks. Why these clusters instead of ones more or less homogeneous? Why keep the hierarchical ordering by parishes instead of another more related to political preferences? With these questions I am only trying to illustrate natural doubts that can arise when we introduce ad hoc criteria for grouping notebooks. If the results were independent of the grouping level, then this would not matter, but this is not the case.

My second concern is the use of the usual asymptotic distribution of Pearson's χ^2 statistic to determine when an observed value of the test statistic is an outlier. This asymptotic does not hold in the framework that we are considering. In general, it is doubtful that this holds when the multinomial distribution, which is the standard underlying assumption when this test is performed, is replaced by a multivariate hypergeometric distribution (Zelterman, 2006), which is the reference model for the distribution of votes among notebooks. In particular, because *all* the votes of each cluster are distributed among the notebooks, the correlations are not negligible. Despite this, I do not deny
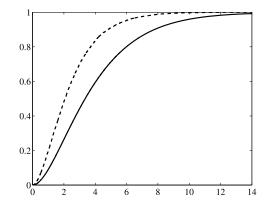
FIG. 4. *Left panel*: *Exact probability density function* (*dashed line*) *of the* χ^2 *test statistic related to the cluster with five notebooks described in Table* 9 *of Febres and Marquez* (2006). *Probability density function of the reference distribution used by the authors* (*solid line*). *The cross marks the observed value for the test statistic. Right panel*: *Exact cumulative distribution function* (*dashed line*) *and usual asymptotic approximation for the* χ^2 *test statistic* (*solid line*).

that there is a high percentage of irregular clusters. To illustrate the previous comment, we consider the cluster with five notebooks described in Table 9 of Febres and Marquez (2006). Following the standard asymptotics, the authors used the χ^2 distribution with four degrees of freedom to compute the p-value of the test statistic related with this cluster. We compute the exact distribution of this statistic to compare with the $\chi^2(4)$ distribution. How to compute the exact distribution is not relevant for now (it is a simple exercise following the discussion in Section 3). The important thing here is that an outlier for the $\chi^2(4)$ distribution is also an outlier for the exact distribution (see the left panel of Figure 4). In fact, as the right panel of Figure 4 shows, the test statistic for this cluster is *less* than $\chi^2(4)$ *in the usual stochastic order*. If we had a similar result for all clusters, then we could ensure that the percentage of irregular clusters is equal to or greater than the percentage reported in the paper. I believe that such a result could be obtained. An alternative would be to compute the exact distribution for each cluster to recompute the p-values and the percentage of outliers. This involves high computational costs but it would also allow us to test the authors' main claim about a monotone causal relationship between the proportion of YES votes and the p-value.

The conjectures of Febres and Marquez are interesting and point in a concrete direction, but require a further analysis before raising them to conclusions of fraud.

2.5 Too Many Ties?

Taylor (2005) considered the following six models of "fair elections":

T1. A model in which the YES/NO votes in computerized notebooks are independent and identically distributed Poisson random variables, with common expectation according to the results in the country.

T2. The same model as above but with a common distribution which is not necessarily Poisson.

T3: A model in which the YES/NO votes in the notebooks of each electoral table are independent and identically distributed Poisson random variables, with common expectation according to the results in the table.

T3.1. A model in which the distribution of YES/NO is multinomial, splitting up the YES/NO votes of each electoral table equally among the notebooks.

T4. A multivariate hypergeometric model, conditioned on the results per electoral table and valid votes per notebook.

T5. A parametric bootstrap where total votes of notebooks $\{T_i\}$ are generated according to the integer part of a multivariate Normal distribution. Then YES votes in notebook i are sampled according to a Binomial(T_i, p), p being the proportion of YES vote in the electoral table.

Although in Taylor's paper it is not always explicitly said, T3–T5 are conditioned on the official results by electoral table and T4 is additionally conditioned on the official number of valid votes by notebook.

From these models, the author analyzed different statistical anomalies related with claims of fraud. Two of them have been previously discussed in this section (Febres and Marquez 2006; Pericchi and Torres 2011).

The third is related to high rates of YES ties: A YES tie is a perfect match of YES votes between two notebooks. Accordingly, his analysis can be divided into three parts:

- Global test for goodness of fit for models T3 and T3.1.
- Comparative study between the distribution of the significant digits according to T3.1 (also to a slight improvement of T1), the observed distribution and Benford's Law.
- Computation of the expected number of electoral tables with one or more YES ties, for each model; and comparison with the observed number of ties.

His main results and conclusions can be summarized as follows:

R1. "The more powerful χ^2 test" strongly rejects the Poisson model T3. However, a *False Discovery Rates* analysis (Benjamini and Hochberg, 1995) shows "there is not evidence of widespread departures for the Poisson model." This result "shows no systematic fraud in the form of vote-capping."

R2. The distribution of the significant digits of the multinomial model T3.1 does not conform to Benford's Law and is virtually identical to the observed distribution. Thus, Benford's Law is of "little use in fraud detection in this instance."

R3. The Z scores used to compare the observed number of electoral tables with one or more YES ties with the expected number according to his models "are fairly high" (I will make an exception with the Z test related to T4, which is equal to 2.37). "This only means that we can reject the global null hypothesis" (i.e., the global models) "and not that there indeed was fraud."

First of all, the validity of a statistical model is not entirely justified by the fact that it fits the data, especially if one wants to test the quality of those data. The costs, here associated with a false negative (failing to identify a fraud condition when one exists), are too high. The model should at least not be at odds with our knowledge about the system that is being modeled.

According to Taylor's web page,[3] "the first two models" (T1 and T2) "are clearly unrealistic." The next two (T3 and T3.1) also are:

(a) As discussed in the previous subsection, the assumption of independence among the notebooks is meaningless. There are links on the sums of votes across an electoral table. All the votes of each table

are distributed among its notebooks, so the correlations are not negligible.

(b) The number of voters (note again the difference between voters and votes) by notebooks varies among the notebooks of the same electoral table, so it makes no sense to equally split the votes of a table among the notebooks.

The last two models (T4 and T5) take into account (a) and (b). In particular, I agree that the multivariate hypergeometric approach used in T4 is the right way to generate vote configurations. However, T5 resorts to assumptions that can be questionable, as to the use of the integer parts of multivariate Normal random variables to generate valid votes by notebooks. Given that the two models provide similar results according to his own analysis, we will apply the principle of Occam's razor[4] to reduce his list to just one *realistic* model.

What can we conclude when a questionable dataset does not show evidence of widespread departures for an unrealistic model? What if the distributions of the significant digits are similar between them but differ from Benford's Law? The conclusions in R1 and R2 are baseless. We cannot conclude anything useful from these analyses.

Let us move to R3, where he considers the multivariate hypergeometric model and simulations carried out by Felten et al. (2004) to analyze the YES ties phenomenon. These simulations show that the number of electoral tables with one or more ties is high, but not high enough to be considered a sign of fraud (around 1% of cases can have an equal or greater number of tables with YES ties, according to this model). This part of his analysis did not detect extreme statistical anomalies that would indicate obvious fraud in the referendum. Of course, as Felten et al. emphasized, this does not imply the absence of fraud, either.

3. REEXAMINING THE REFERENDUM

The purpose of this section is to reevaluate the claim of fraud. An electoral fraud occurs if the results are altered to favor one of the options. Having evidence that the changes are enough to overturn the winner, the outcomes of the referendum should not be recognized. Moreover, if the handling does not change the winner, but changes the proportions significantly, it must be considered a fraud. Electoral results can affect drastically future electoral processes. In particular,

---

[3] http://www-stat.stanford.edu/~jtaylo/venezuela.

[4] *Entia non sunt multiplicanda praeter necessitatem* (entities must not be multiplied beyond necessity).

this could have happened during the Venezuelan parliamentary elections, one year after the referendum, in which the political parties that supported the YES option withdrew, claiming the possibility of *new* fraud. Also, a tight result can have a different political meaning than an outcome with a winner by a wide margin, especially in a recall referendum. At the end of this section we evaluate the hypothesis of irregularities in the vote counting to favor significantly the NO option. We begin by describing the joint probability distribution of results per notebook, conditioned on the complete set of information of each center. This corresponds to a multivariate hypergeometric model, similar to that used in Felten et al. (2004) and Taylor's T5 model (the differences are explained below). This is a key tool in the hypothesis test methodology that we develop through this section.

3.1 Shuffling Voting Cards

Consider a center with m notebooks, labeled by $1, 2, \ldots, m$. Let $\nu = \sum_{i=1}^{m} \tau_i$ be the total voters in the center. Identify each voter by a number in $\{1, 2, \ldots, \nu\}$ such that the first τ_1 voters are in notebook 1, the following τ_2 voters are in notebook 2, and so on. In the vote counting, each voter is represented by a voting card according to her/his electoral option. It can be YES, NO or OUT. Let X_i be the voting card of voter i. Then, the vote configuration at the center can be represented by

$$(7) \quad \mathcal{X} = (\underbrace{X_1, \ldots, X_{\tau_1}}_{\text{notebook 1}}, \ldots, \underbrace{X_{\nu-\tau_m+1}, \ldots, X_\nu}_{\text{notebook } m}).$$

Let $y = \sum_{i=1}^{m} Y_i$ be the total YES votes in the center. Similarly, let $n = \sum_{i=1}^{m} N_i$ be the total NO votes. Then, $\mathcal{X}$ is an outcome of shuffling the voting cards of the center:

$$\mathcal{C} = (\underbrace{YES, \ldots, YES}_{y \text{ YES's}}, \underbrace{NO, \ldots, NO}_{n \text{ NO's}}, \underbrace{OUT, \ldots, OUT}_{\nu-y-n \text{ OUT's}})$$

(8)

That is to say that $\mathcal{X}$ is a permutation of $\mathcal{C}$.

According to the random mechanism used by the electoral umpire to assign voters to notebooks, given (y, n, ν), any permutation of voting cards has the same probability of occurring. This is the underlying statistical principle shared by Febres and Marquez (2006), Felten et al. (2004) and Taylor (2005) for testing the referendum data. However, these authors do not consider all possible permutations:

- The sampling distribution of the test used by Febres and Marquez (2006) in their first round, where they conditioned on results by centers and valid votes by notebooks, corresponds to sampling on the set of outcomes of shuffling YES cards and NO cards in centers, leaving fixed OUT cards in notebooks.
- The samples from the multivariate hypergeometric model considered by Felten et al. (2004) and Taylor (2005) belong to a set of permutations even smaller than the previous one. They conditioned additionally on the total of YES votes and NO votes by electoral tables. That is, they just considered shuffling YES cards and NO cards in tables, also leaving fixed OUT cards in notebooks.

Both approaches fail to consider a large number of equiprobable results that match the referendum results at the centers. In this paper, we compute sampling distributions of test statistics considering all possible permutations of the voting cards at each center. To simplify the writing, in what follows, we will refer to the result obtained by shuffling randomly the cards across all centers as a *random sample of the electoral process*.

3.2 Statistical Hypothesis of Fair Referenda

If we assume that the referendum was properly conducted, the results by notebook correspond to a random sample of the electoral process. Therefore, the hypothesis of a *properly conducted referendum* is

$\mathcal{H}_0$: *The votes per notebook correspond to a random sample of the electoral process.*

But, the rejection of $\mathcal{H}_0$ does not imply that the results per notebook were altered to favor one of the options. It only implies that there is a significant presence of outliers in the distribution of votes per notebook. Innocent irregularities, as the incorrect allocation of voters to notebooks, can generate such outliers. We consider the most innocent alternative to $\mathcal{H}_0$, assuming that: (1) there is a significant presence of outliers in the votes per notebook, (2) the outliers are the result of neutral irregularities, and (3) the irregularities affect a random set of notebooks, regardless of whether they belong to strongholds of the winning option or not. Therefore, we consider the hypothesis of an *atypical fair referendum*, namely,

$\mathcal{H}_1$: *There is a significant presence of outliers in the votes per notebook that is consequence of inno-*

cent irregularities that affect a random set of note-books.

If there is in fact a significant presence of outliers, we can reject $\mathcal{H}_1$ because: (1) the irregularities are not innocent, introducing a significant bias in the vote counting, or (2) they affect mostly notebooks in bastions of one of the options. Therefore, we have to consider the bizarre, but fair, scenario in which the irregularities that generate the outliers are neutral, no matter what, and, for some reason, they affect mostly a set of notebooks that are in strongholds of one of the options. Thus, we consider the hypothesis of a *bizarre but fair referendum*:

$\mathcal{H}_2$: *The significant presence of outliers in the votes per notebook is the result of innocent irregularities that affect mostly a set of notebooks from strongholds of one electoral option.*

The remaining alternative is a clear signal that the irregularities are not innocent.

Before testing the hypotheses, we describe the dataset.

3.3 Description of the Dataset

It is required to have at least two notebooks per center for shuffling voting cards, so we restrict our analysis to these centers. In addition, since all allegations of fraud are related to computerized notebooks, we only consider full computerized centers (centers where there are no manual votes). We also exclude a very small number of centers with empty notebooks (notebooks without valid votes). Empty notebooks could arise for technical problems, affecting the distribution of voters to notebooks in such centers. After this simple cleaning on full computerized centers with two or more notebooks, a consistent dataset is obtained with 4,162 centers, all of them with comparable notebooks. This means that the votes among the notebooks of a center are in the same order. These 4,162 centers represent 18,297 notebooks, more than 83% of the total voters, and here will be the base of the study. The mean and the standard deviation of the number of voters per unit polls are 634 and 73, respectively.

For the last two subsections of this section, we will also use the results of presidential elections of 1998, in which Chavéz was elected to his first term as President of Venezuela with 56% of valid votes against a coalition of roughly the same political parties that supported the recall. This election was carried out with an automated voting system, which featured a single integrated electronic network to transmit the results from the polling stations to central headquarters (McCoy,

1999). The legitimacy of the electoral process and the acceptance of the results by political parties and international observers is a guarantee of the reliability of the results. At that time, in each center, voters were also randomly assigned to polling units, according to the last number of their ID, similarly to the process described in Section 2. As we do with the referendum's dataset, we exclude centers with only one unit poll and those with empty unit polls. After the cleaning, a dataset is obtained with 3952 centers, 15,667 unit polls, that represent 85% of the total voters. The mean and the standard deviation of the number of voters per unit poll are 594 and 112. For all the above, both sets of data are comparable for the statistical purposes of Section 3.6.

A different scenario overshadowed the presidential elections of 2000, that we consider in Section 3.7, in which Chavéz was elected to his second term with 59% of valid votes. After two years of important political changes, including the enacting of a new constitution, the criticism of Chavez's government increased, polarizing the political climate. Many claims of fraud, including machines not properly functioning, people whose names did not appear on the electoral registry and pre-marked ballots, were made at that time. While The Carter Center does not believe that the election irregularities would have changed the presidential results, they consider those elections as flawed and not fully successful (Neuman and McCoy, 2001). The election was carried out, roughly, with the same voting system used in 1998. However, there was an important difference in the number of voters per unit poll, increasing significantly the number of centers with only one unit. As with the referendum and the presidential elections of 1998, we exclude centers with only one unit poll and those with empty unit polls. In addition, we exclude unit polls with more votes than voters. Thus, we obtain a dataset with 1,600 centers, only 3,730 unit polls, that represents 53% of the total voters.

The three datasets provide estimates of high precision for the resultant percentage of votes per electoral option. Table 1 summarizes their main statistics.

3.4 Testing the Hypothesis of a Properly Conducted Referendum

The way to test irregularity is to determine whether an observed value is an outlier or not. Let i be a focal notebook and c the center to which it belongs:

– Since voters of a center are randomly assigned to notebooks, Y_i is the total of YES cards in a simple random sample (without replacement) of size τ_i

TABLE 1
Statistical summary of the dataset

Year	Unit polls	% of total voters	Mean of voters per unit poll	Standard deviation
1998	15,667	85%	594.70	111.97
2000	3,730	53%	1662.50	361.74
2004	18,297	83%	634.60	73.86

TABLE 2
Results in centers with notebooks associated with minimum (*)
and maximum (+) Z value

Notebook	Center 7990 m			Center 1123 M		
Y	174	81*	235	191	60	233+	62
N	272	70	375	396	137	359	143
τ	607	600	610	588	583	594	567

from the voting cards of the center. In particular, $\mathbb{E}[Y_i|\mathcal{H}_0] = p_c\tau_i$, with $p_c = y_c/v_c$ and

$$\text{Var}[Y_i|\mathcal{H}_0] = \tau_i p_c (1 - p_c)\frac{v_c - \tau_i}{v_c - 1}.$$

– The minimum τ_i in the 18,297 notebooks involved is 347 (the mean value $\bar{\tau} = \sum \tau_i/18{,}297$ is equal to 634.60, and the maximum is 975).

Coupling these facts, a straightforward application of the Central Limit Theorem implies that, under $\mathcal{H}_0$, the score

$$(9) \qquad Z_i = \frac{(Y_i - p_c\tau_i)}{\sqrt{p_c(1 - p_c)\tau_i(v_c - \tau_i)/(v_c - 1)}}$$

is approximately $\mathcal{N}(0, 1)$, for any i. Therefore, a test of regularity for a single notebook is reduced to determining the significance of Z.

To get an overall qualitative idea of the joint behavior of the Z scores under $\mathcal{H}_0$, the normal probability plot of these statistics from a random sample of the electoral process is shown in Figure 5 (left panel). In the same figure (right panel) the normal probability plot of the scores based on the observed values is also shown. This plot highlights many official results far from what is expected. Let us peer into the most atypical cases. Table 2 shows the official results of centers 7990 and

1123,[5] where are the notebooks associated with minimum (-9.08) and maximum (10.54) Z value. Let us call these notebooks m and M, respectively. Under $\mathcal{H}_0$, the expected value of Y_m is 161.81, almost twice the observed value, which is 81, while the expected value of Y_M is 139.08, just over half of what is observed, which is 233.

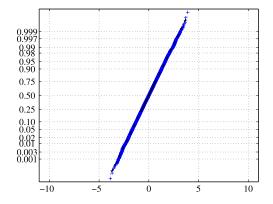An overall comparison is handled by summing squares of Z scores. Let

$$(10) \qquad S^2 = \sum_{i=1}^{18,297} Z_i^2.$$

A straightforward computation gives $\mathbb{E}[S^2|\mathcal{H}_0] = 18{,}297$. The variance can be estimated by Monte Carlo, shuffling the voting cards. We performed 1,000 random samples of the electoral process and obtained a standard deviation of 216. Next we show that the sampling distribution of the test statistic

$$(11) \qquad T^{\text{YES}} = \frac{S^2 - 18{,}297}{216}$$

can be approximated by a standard Normal distribution.

[5]We use the center encoding used for the referendum. Codes, as well as the list of centers, varies from election to election.
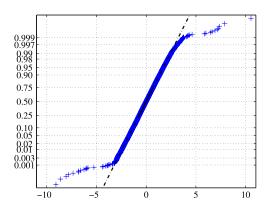

FIG. 5. *Normal probability plot of Z scores based on a random sample (left) and observed values (right).*

TABLE 3
Number of clean and fully computerized centers with m notebooks

m	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
C_m	1,044	820	665	496	380	300	208	110	54	41	19	12	4	4	2	2	1

The centers have between 2 and 18 notebooks. The distribution of the centers according to the number of notebooks is shown in Table 3.

The sum of squares can be decomposed as follows:

$$S^2 = \sum_{i=1}^{18,297} Z_i^2 = \chi^2_{nb(1)} + \chi^2_{nb(2)} + \cdots + \chi^2_{nb(18)},$$

$\chi^2_{nb(i)}$ being the sum along all the centers of the squares of the Z scores related to the ith notebook of a center. Although the results of notebooks belonging to the same center are correlated, given $\mathcal{H}_0$ they are independent of results in other centers. In turn, each $\chi^2_{nb(i)}$ is the sum of independent random variables and each one is approximately the square of a standard normal random variable. Then, $\chi^2_{nb(i)}$ is approximately χ^2 with $\sum_{m \geq i} C_m$ degrees of freedom, C_m being the number of centers with m notebooks. Table 4 lists the degrees of freedom related to $\{\chi^2_{nb(i)}, 1 \leq i \leq 18\}$.

In general, approximating the distribution of sums of correlated χ^2 can be difficult. Fortunately, this is not case here. Two remarks:

– For $i \leq 10$, the degrees of freedom are large enough to fit the distribution of $\chi^2_{nb(i)}$ by a Normal.
– $\chi^2_{nb(1)} + \cdots + \chi^2_{nb(10)}$ represents 99% of the Z^2 statistics in S^2.

Therefore, S^2 is approximately a sum of Normal random variables. Letting ς^2 be the sample variance obtained from k independent samples of S^2, under $\mathcal{H}_0$, the test statistic

$$(12) \qquad T^{\text{YES}} = \frac{S^2 - \mathbb{E}[S^2|\mathcal{H}_0]}{\varsigma}$$

is approximately $\mathcal{N}(0, 1)$, for any large k. As we said above, we simulated 1,000 random samples of the electoral process, obtaining $\zeta = 216$. We also used the

samples to confirm that, under $\mathcal{H}_0$, the distribution of T^{YES} is approximately $\mathcal{N}(0, 1)$. For that, we test normality with different methods, all of them with the same conclusive results. To illustrate, Figure 6 compares the kernel density estimator of the probability density function of T^{YES} with the probability density of a standard Normal.

The T^{YES} observed value, according to the official results, is $T^{\text{YES}}_{\text{obs}} \approx 13.12$, which establishes that the results of YES votes per notebook are not credible, given the results by centers. The p-value, less than the MatLab precision, is strong evidence against $\mathcal{H}_0$.

Following the same approach, we can test regularity on the distribution of NO votes and abstentions. For that, we define the Z statistics

$$
(13) \qquad
\begin{aligned}
Z_i^{\text{NO}} &= \frac{(N_i - q_c \tau_i)}{\sqrt{q_c(1 - q_c)\tau_i(\nu_c - \tau_i)/(\nu_c - \tau_i)}} \quad \text{and} \\
Z_i^{\text{OUT}} &= \frac{(O_i - r_c \tau_i)}{\sqrt{r_c(1 - r_c)\tau_i(\nu_c - \tau_i)/(\nu_c - \tau_i)}},
\end{aligned}
$$

$q_c = n_c/\nu_c$ and $r_c = (\nu_c - y_c - n_c)/\nu_c$ being the proportion of NO votes and OUT votes in the center c to which notebook i belongs. As an illustration, Figure 7 shows the normal probability plots of these Z statistics based on a random sample of a properly conducted referendum. The figure shows also the normal probability plots of the scores based on the official results. These plots show a widespread departure from the expected values, even stronger than for the YES case (be careful with the scales of these figures). In fact, if we define test statistics to the distribution of NO votes and OUT votes, T^{NO} and T^{OUT} respectively, similar to what we did for the YES votes, then we have

$$T^{\text{OUT}}_{\text{obs}} > T^{\text{NO}}_{\text{obs}} > T^{\text{YES}}_{\text{obs}}.$$

Clearly, $\mathcal{H}_0$ can be completely rejected.

TABLE 4
Degrees of freedom (df) related with $\chi^2_{nb(i)}$

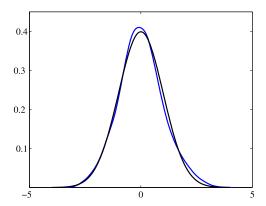| i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
df	4,162	4,162	3,118	2,298	1,633	1,137	757	457	249	139	85	44	25	13	9	5	3	1

FIG. 6. *Kernel estimator of the probability density function of T^{YES} versus a standard Normal probability density.*

3.5 Testing the Hypothesis of an Atypical Fair Referendum

As mentioned previously, the widespread departure of YESs, NOs and OUTs per notebook from their expected values could be the outcome of innocent irregularities in the conduct of the referendum. Incorrect allocation of voters to notebooks and the passing of votes from one notebook to another during the vote counting, by bugs in the programming, are examples of such ir-

regularities. These irregularities may generate, in particular, Z^{OUT} outliers but, by the secrecy of the ballot, they should not be associated with a trend in the vote counting. Next, we propose a testing methodology, based on a simple statistical control chart, for testing trend in the vote counting on potentially irregular notebooks. The methodology can be easily extended to other electoral audit frameworks. It relies on the assumption that unexpected irregularities can occur in any unit poll with the same probability.

Denote by R the ratio between NO votes and total valid votes in the target population, consisting of $K = 18{,}297$ notebooks, namely,

$$(14) \qquad R = \frac{\sum_{i=1}^{K} N_i}{\sum_{i=1}^{K} T_i}.$$

In sampling jargon, R is the *population ratio* and K is the *size of the population*. Let $\mathcal{S}_k$ be the sample consisting of the k notebooks with the most extreme Z^{OUT} values. This is the set of k notebooks with Z^{OUT} values furthest away from zero. Given a confidence level $1 - \alpha$, there is a $k := k(\alpha)$ such that $\mathcal{S}_k$ matches the set of notebooks with Z^{OUT} values that we consider that are outliers, that is, the set of notebooks with Z^{OUT}
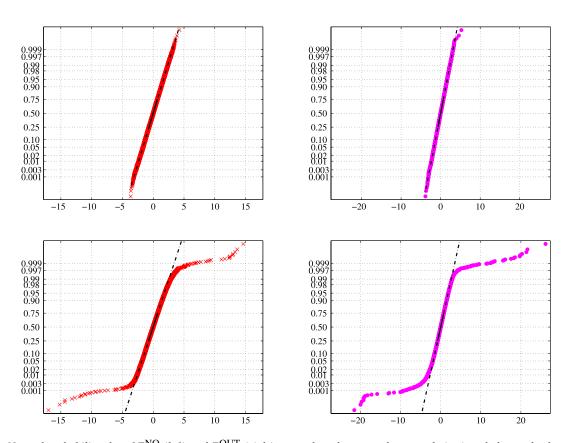

FIG. 7. *Normal probability plot of Z^{NO} (left) and Z^{OUT} (right) scores based on a random sample (top) and observed values (bottom).*

values out of the $(1 - \alpha) \times 100\%$ normal confidence interval. In our case study, if the confidence level is 99%, then $k = 706$. Roughly, 4% of the Z^{OUT} values are out of the 99% confidence interval. In what follows, k varies in a range such that $\mathcal{S}_k$ corresponds to the set of outliers, according to some reasonable confidence level.

Denote by r_k the *sample ratio* based on $\mathcal{S}_k$. That is,

$$(15) \qquad r_k = \frac{\sum_{i \in \mathcal{S}_k} N_i}{\sum_{i \in \mathcal{S}_k} T_i}.$$

Note that r_k is not the usual ratio estimator, since we are sampling notebooks with atypical Z^{OUT} values. Thus, we might expect that observations (N_i, T_i) in $\mathcal{S}_k$ are larger or smaller than those from a simple random sample (SRS). However, if the irregularities are innocent, if they do not introduce bias in the vote counting, r_k should be similar to the sample ratio based on a SRS. In particular, if k is large, the bias of the estimator will be small and the variance can be approximated by
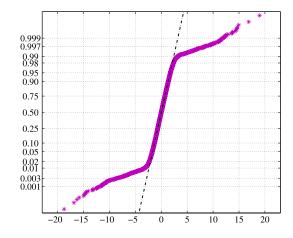
$$\mathrm{Var}(r_k) \approx S_k^2 := \left(1 - \frac{k}{K}\right) \frac{1}{\mu_T^2} \frac{s_r^2}{k}$$

(Lohr, 2004), with

$$\mu_T = \frac{1}{K} \sum_{i=1}^{K} T_i \quad \text{and} \quad s_r^2 = \frac{1}{k-1} \sum_{i \in S_k} (N_i - r_k T_i)^2.$$

Thus, under the hypothesis $\mathcal{H}_1$ defined in Section 3.2, if k and $K - k$ are large enough,

$$(16) \qquad \zeta_k = \frac{r_k - R}{S_k^2}$$

is distributed approximately as a standard normal variable. In what follows, we will only consider $100 \leq k \ll K$.

To illustrate the above, consider 1,000 independent copies of ζ_{500} from a random sample of atypical fair referenda. We simulate an atypical fair referendum by introducing 700 innocent irregularities on a random sample of the electoral process. Each irregularity consists in passing a random proportion of votes (10% on average) from a notebook to another located in the same center. This handling produces a significant number of Z^{OUT} outliers (outside the 99% normal confidence interval) to those already obtained before the manipulation. The normal probability plot of the Z^{OUT} scores of one of these atypical fair referenda is shown in Figure 8 (left panel) as an example. As we can see, the shape of the plot is similar to that observed for the referendum (Figure 7, right bottom panel). We test normality of ζ_{500} with different methods, all of them with the same conclusive positive results. To illustrate, the right panel of Figure 8 compares the kernel density estimator of the probability density function of ζ_{500} with the probability density of a standard Normal.

We can test the hypothesis of an irregular fair referendum using the ζ_k scores. High values of ζ_k imply that irregularities introduce a bias in favor of the NO option in the vote counting. Small values of this score imply a bias in favor of the YES option. Under $\mathcal{H}_1$, we expect ζ_k to be within a confidence interval.

The ζ_k scores corresponding to the official results are plotted in Figure 9 (top line), for k between 100 and 706. To illustrate the behavior that we expect in an atypical fair referendum, we plot, in the same figure, 100 simulated scores series of the atypical fair ref-
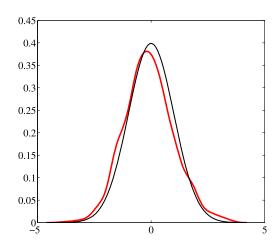

FIG. 8. *Left panel*: *Normal probability plot of Z^{OUT} scores from an atypical fair referendum. Right panel*: *Standard normal probability density versus kernel estimator of the probability density function of r_{500}, based on 1,000 independent copies of an atypical fair referendum.*
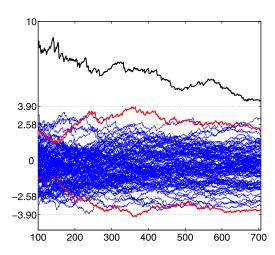
FIG. 9. ζ_k versus k for official results (top line) and simulated atypical fair referenda.

erendum discussed above. The ζ_k scores corresponding to the official results are well above the 99.99% confidence interval $(-3.9, 3.9)$ for $100 \leq k \leq 706$. Although a small number of simulated trajectories also reach high values, all of them are embedded in $(-3.9, 3.9)$ and most of them are in the 99% confidence interval $(-2.58, 2.58)$, as one expects. We observed similar behavior in 1,000 additional simulated trajectories (not plotted). The scores series of the referendum reaches values higher than any that we observed in simulations, being the only one always well above 3.9, for $100 \leq k \leq 706$. This provides strong evidence against $\mathcal{H}_1$ than a fairly small p-value of a ζ_k score, for some k. We are seeing a significant bias in the vote counting on notebooks associated with irregularities, which is almost impossible to observe under $\mathcal{H}_1$. All the above is strong evidence for rejecting it.

3.6 Testing the Hypothesis of Bizarre but Fair Referendum

Most political scientists expect more innocent administrative errors in areas with more poor voters (M. Lindeman, personal communication, July 2010). In addition, "the conventional wisdom about contemporary Venezuelan politics is that class voting has become commonplace, with the poor doggedly supporting Hugo Chávez while the rich oppose him" (Lupu, 2010). If both beliefs are true, we expect more innocent irregularities in strongholds of the NO option, which would explain the atypical result observed in the above section. That is what $\mathcal{H}_2$ describes, a general scenario in which there are more innocent irregularities in centers that support the winning option. To illustrate this possibility, we show in Table 5 the results in Center

TABLE 5
Results in Center 1123 (*C. M. A. Dr. Angel Vicente Ochoa, in Santa Rosalía, Caracas*)

Notebook	1	2	3	4
Y	191	60	233	62
N	396	137	359	143
τ	588	583	594	567

1123 (C. M. A. Dr. Angel Vicente Ochoa, in Santa Rosalía, Caracas), one of the most extreme results. All its notebooks are associated with very extreme Z^{OUT} values (greater than 18.53!). But the overall NO proportion (65%) is even less than that observed in the presidential elections of 1998 (67%). This center appears to be a bona fide Chávez stronghold. However, we have to remark that, in this election, the Z^{OUT} values of that center are quite normal, all of them between -0.4 and 0.40. The results, only three unit polls in that election, are shown in Table 6.

A naive procedure to see if the irregularities affected mostly notebooks in Chávez's strongholds would be repeating the previous analysis on all centers with outliers. We consider an alternative analysis for an important reason: If indeed tampering occurred, it is possible that, in order to mask the stuff, the irregularities were committed precisely in Chávez's bastions. In addition, we have to admit that we do not have access to the recoding of centers to automatize the procedure.
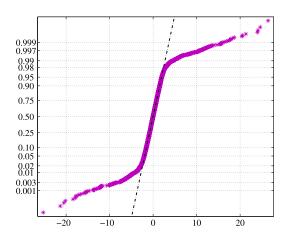
Lupu (2010) provided evidence that the presidential election of 1998 was more monotonic in class voting than the referendum. This means, the poor were more likely to vote for Chávez in 1998 than in 2004. Thus, we expect more innocent irregularities in Chávez's strongholds in 1998 than in 2004. In addition, there is not doubt about the legitimacy of this election (Neuman and McCoy, 2001). For these reasons, the election of 1998 is very appropriate to test if irregularities affect mostly notebooks in centers that support Chávez, that is, to test $\mathcal{H}_2$, defined in Section 3.2. The testing

TABLE 6
Presidential elections of 1998, *results in C. M. A. Dr. Angel Vicente Ochoa Center*

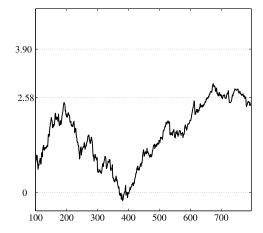| Unit poll	1	2	3
Y	182	115	123
N	357	265	247
τ	899	645	624

FIG. 10. *Left panel*: *Normal probability plot of* Z^{OUT} *scores of the presidential elections of* 1998. *Right panel*: ζ_k *versus k for presidential elections of* 1998.

schema we use is to reject $\mathcal{H}_2$ if we fail to reject $\mathcal{H}_1$ for the elections of 1998. We begin verifying that there is a significant presence of Z^{OUT} outliers in 1998: 5% of Z^{OUT} values (797 of 15,667) are out of the 99% normal confidence interval. The evidence against $\mathcal{H}_0$ is of the same order as in 2004. Furthermore, the most extreme Z^{OUT} values of 1998 are higher than those observed in 2004. We omit the details and summarize results by showing the normal probability plot of the Z^{OUT} scores in Figure 10 (left panel). It seems possible that, in complex elections, ad hoc decisions are made to resolve problems that arise on the fly. As we have discussed previously, this can produce large outliers in the vote distribution. However, the test discussed in the previous section strongly supports $\mathcal{H}_1$ for the presidential elections of 1998. The corresponding scores series $\{\zeta_k, 100 \leq k \leq 797\}$ is almost embedded in the 99% confidence interval; see right panel of Figure 10. Therefore, we see that there is little reason to think that the significant presence of Z^{OUT} outliers, that are the result of innocent irregularities, affect mostly a set of notebooks from Chávez's strongholds. Irregularities seem to occur randomly, regardless of whether the notebook belongs to a Chávez bastion or not, and thus, we reject $\mathcal{H}_2$.

3.7 Estimating the Effect of the Irregularities

We have provided statistical evidence that there was a significant presence of irregularities that favored the winning option in the vote counting of 2004. But, how much could the irregularities affect the overall results? Suppose that the Z^{OUT} outliers are the *tip of the iceberg* and there is bias in the vote counting of a high proportion of notebooks, not just in the notebooks with

extreme Z^{OUT} values. To evaluate this assumption, we analyze the behavior of the sample ratio r_k in (15) for higher values of k than those that we have already considered. Figure 11 shows that proportion for k varying from 100 to 3000 (top line). The shape shows a strong correlation between the trend in the vote counting and the discrepancy between valid votes and its expectation, not only in notebooks with Z^{OUT} outliers. We remark that for $k = 100$ we are considering 41,533 valid votes, a very large sample size for estimating proportions (an accepted standard for pollsters is above 1,000). What we expect when we increase the sample size is exactly what we have for the presidential elections of 1998 (line from the middle to the bottom in Figure 11): The proportion is a function of the sample size that slightly varies around the population proportion, and that quickly stabilizes around this
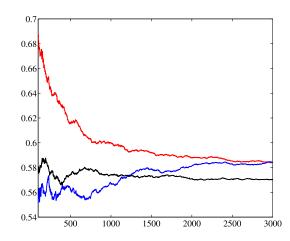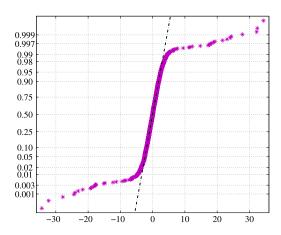

FIG. 11. *Proportion of Chávez's votes in* $\mathcal{S}_k$ *versus k for the referendum* (*top line*) *and presidential elections of* 1998 (*line from middle to the bottom*) *and* 2000 (*line from the bottom to the middle*).
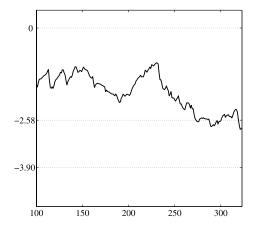
FIG. 12. *Left panel*: *Normal probability plot of* Z^{OUT} *scores of the presidential elections of* 2000. *Right panel*: ζ_k *versus k for presidential elections of* 2000.

value. So, our assumption of irregularities that affect the vote counting across all the notebooks is quite possibly true.[6] Let us measure how much it could affect the totals.

Let R be the population ratio defined in (14). Bounds for the relative error, introduced in the vote counting by the irregularities, can be obtained maximizing and minimizing the relative error $(r_k - R)/R$. Thus, we can provide a prediction interval for the *corrected proportion* of votes in favor of Chávez, namely,

$$
(17) \quad \left[\max_{100 \leq k \leq K/2} \left(1 - \frac{r_k - R}{R} \right) R, \right.
$$
$$
\left. \min_{100 \leq k \leq K/2} \left(1 - \frac{r_k - R}{R} \right) R \right],
$$

K being the total number of notebooks. Note that we are considering up to 50% of the notebooks ($k \ll K$), those with the highest $|Z^{OUT}|$ values.

For example, the prediction interval (17) for the presidential election of 1998, in which Chávez won with 56% of valid votes, is [55%, 57%]. We remark that this is an example of an atypical but fair election, where the results were well accepted by political parties and international observers.

Let us consider next the presidential elections of 2000. As mentioned above, The Carter Center considers this election as flawed and not fully successful. However, they also emphasize that the irregularities did

not change the presidential results. Our methodology confirms their conclusion. Figure 12 summarizes our testing analysis. We observe the highest presence of Z^{OUT} outliers in 2000: 9% of Z^{OUT} values (327 of 3730) are outside the 99% normal confidence interval. Also, the most extreme Z^{OUT} scores of 2000 are higher than the observed in 1998 and 2004. But, there is not evidence to reject $\mathcal{H}_1$ for this election. The ζ scores series is always in the 99% normal confidence interval, except for a short excursion. Moreover, the prediction interval (17) for this election is [59%, 62%], and Chávez was elected with 59% of the valid votes.

We do observe a controversial result in the referendum, managed by a different electoral umpire from those that managed the elections of 1998 and 2000: The prediction interval is [47%, 57%]. The official result (59%) is out of range, while results that overturn the winner are within. We remark that while this is not proof that irregularities changed the overall results, it does illustrate that such a scenario is plausible. Certainly, the result should be, at least, more in line with the prediction interval.

4. CONCLUSIONS

The main tool for conciliating political actors in an election under suspicion of fraud is a full audit. When this is not possible, statistical methods for detecting numerical anomalies and diagnosing irregularities can be useful for evaluating the likelihood of the allegations of fraud. This is the aim of *election forensics* (Mebane, 2008), an exciting area of applied statistics. Election forensics has been applied for several recent controversial elections, including 2004 USA, 2006 Mexico, 2008 Russia and 2009 Iran (Mebane, 2011); see the

---

[6]Martín (2011), which unfortunately was not available for my review, studies the volume of traffic in incoming and outgoing data between notebooks and totalizing servers. It provides evidence that the vote counting of a high percent of notebooks could be affected from the totalizing servers.

personal web page of Walter Mebane.[7] The Venezuelan recall referendum is a case study that shows a wide pallet of the commonly used statistical tools and problems that can arise in this type of analysis, as shown by our review in Section 2. In particular, we have highlighted problems related to exit polls, causal relationship between number of votes and dependent variables, Benford's Law, different levels of data aggregation, goodness of fit, and election modeling. Beyond the statistical learning, the hard criticism of some of the papers reviewed relates to a deep concern about the future of this emerging area. I am convinced that the diffusion of inaccurate analyses only causes founded allegations of fraud to be undervalued. At least, this was the case of the Venezuelan referendum.

We propose a forensic election methodology, based only on vote counting, to analyze the referendum. Also the Venezuelan presidential elections of 1998 and 2000 are reviewed. Unlike previous work, we used the full information of the official dataset. This consists not only of the number of votes for and against revoking the mandate of President Chavez, but also the number of abstentions and invalid votes at the official data unit with the lowest number of votes. The main conclusion of the present paper is that there were a significant number of irregularities in the vote counting that introduced a bias in favor of the winning option. We provide prediction intervals for the bias, showing that the scenario in which the bias could overturn the results is plausible. This places solid evidence in the arena, substantiating the allegations of fraud made at the time.

ACKNOWLEDGMENTS

REFERENCES

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

THE CARTER CENTER (2005). Observing the Venezuela presidential recall referendum. Comprehensive report, The Carter Center. Available at http://www.cartercenter.org/documents/2020.pdf.

DE VEAUX, R. D. and HAND, D. J. (2005). How to lie with bad data. *Statist. Sci.* **20** 231–238. MR2189000

DECKERT, J., MYAGKOV, M. and ORDESHOOK, P. C. (2010). The irrelevance of Benford's Law for detecting fraud in elections. Working Paper No. 9, Caltech/MIT Voting Technology Project. Available at http://www.vote.caltech.edu/drupal/files/rpeavt_paper/benford_pdf_4b97cc5b5b.pdf.

DELFINO, G. and SALAS, G. (2011). Analysis of the 2004 Venezuela referendum: The official results versus the petition signatures. *Statist. Sci.* **26** 502–512.

DURSTCHI, C., HILLISON, W. and PACINI, C. (2004). The effective use of Benford's Law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting* **5** 17–34.

ETTERIDGE, M. L., HILLISON, W. and SRIVASTAVA, R. P. (1999). Using digital analysis to enhance data integrity. *Issues in Accounting Education* **4** 675–690.

FEBRES, M. M. and MARQUEZ, B. (2006). A statistical approach to asses referendum resuslts: The Venezuelan recall referendum 2004. *International Statistical Review* **774** 379–389.

FELTEN, E. W., RUBIN, A. D. and STUBBLEFIELD, A. (2004). Analysis of the voting data from the recent Venezuela referendum. Available at http://venezuela-referendum.com.

HAUSMANN, R. and RIGOBON, R. (2004). In search of the black swan: Analysis of the statistical evidence of fraud in Venezuela. Working paper, J. F. Kennedy School of Government, Harvard Univ. Available at http://www.hks.harvard.edu/fs/rhausma/new/blackswan03.pdf.

HAUSMANN, R. and RIGOBON, R. (2011). In search of the black swan: Analysis of the statistical evidence of fraud in Venezuela. *Statist. Sci.* **26** 543–563.

HILL, T. P. (1995). The significant-digit phenomenon. *Amer. Math. Monthly* **102** 322–327. MR1328015

LOHR, S. (2004). *Sampling: Design and Analysis*, 2nd ed. Brooks/Cole, Boston, MA.

LUHNOW, D. and DE CORDOBA, J. (2004). Academics' study backs fraud claim in Chavez election. *The Wall Street Journal*, September 7, 2004.

LUPU, N. (2010). Who Votes for chavismo? Class Voting in Hugo Chávez's Venezuela. *Latin American Research Review* **45** 7–32.

MARTÍN, I. (2011). 2004 Venezuelan presidential recall referendum (2004 PRR): A statistical analysis from the point of view of data transmission by electronic voting machines. *Statist. Sci.* **26** 528–542.

McCOY, J. (1999). Chávez and the end of "Partyarchy" in Venezuela. *Journal of Democracy* **10** 64–77.

MEBANE, W. (2008). Election forensics: The Second-digit Benford's Law Test and recent American presidecial elections. In *Election Fraud*: *Detecting and Deterring Electoral Manipulation*. (R. M. Alvarez, T. E. Hall and S. D. Hyde, eds.) 162–181. Brooking Press, Washington, DC.

MEBANE, W. (2011). Fraud in the 2009 Presidential Election in Iran? *Chance* **23** 6–15.

NEUMAN, L. and McCOY, J. (2001). Observing political change in Venezuela: The Bolivarianan constitution and 2000 elec-

---

[7]www-personal.umich.edu/~wmebane.

tions. Final report, The Carter Center. Available at http://www.cartercenter.org/documents/297.pdf.

PERICCHI, L. and TORRES, D. (2011). Quick anomaly detection by the Newcomb–Benford Law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. *Statist. Sci.* **26** 513–527.

PRADO, R. and SANSÓ, B. (2011). The 2004 Venezuelan presidential recall referendum: Discrepancies between two exit polls and official results. *Statist. Sci.* **26** 502–512.

PRESS, S. J. (1982). *Applied Multivariate Analysis*: *Using Bayesian and Frequentist Methods of Inference*, 2nd ed. Dover, Mineola, NY.

SNEATH, P. H. A. and SOKAL, R. R. (1973). *Numerical Taxonomy*: *The Principles and Practice of Numerical Classification*. Freeman, San Francisco, CA. MR0456594

TAYLOR, J. (2005). Too many ties? An empirical analysis of the Venezuelan recall referendum. Available at http://esdata.info/pdf/Taylor-Ties.pdf.

WALLACE, W. A. (2002). Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management* **51** 21–23.

WEISBROT, M., ROSNICK, D. and TUCKER, T. (2004). Black swans, conspiracy theories, and the Quixotic search for fraud: A look at Hausmann and Rigobon's analysis of Venezuela's referendum vote. Briefing paper, Center for Economic and Policy Research. Available at http://www.cepr.net/documents/publications/venezuela_2004_09.pdf.

ZELTERMAN, D. (2006). *Models for Discrete Data*, Revised ed. Oxford Univ. Press, Oxford. MR2218182