Data manipulation in the Mexican Election?

by

Jorge A. López, Ph.D.

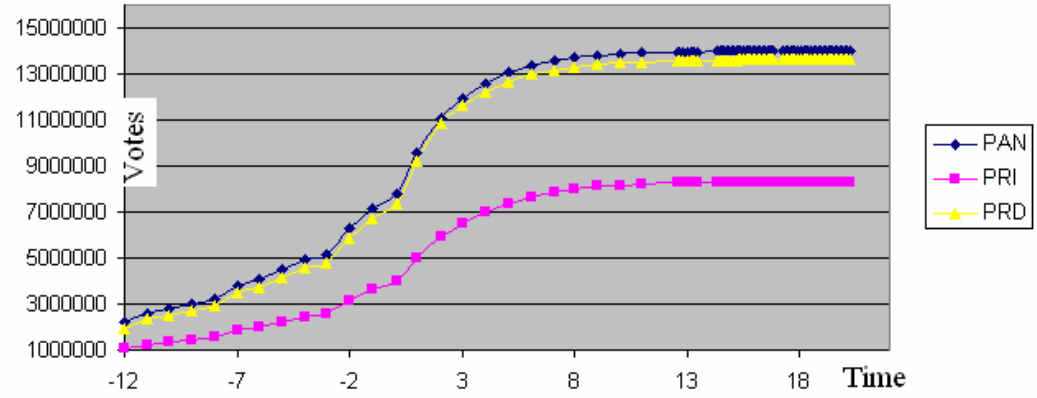
Many of us took advantage of the latest technology and followed last Sunday's elections in Mexico through a novel method: web postings of the votes through the Program of Preliminary Results, or PREP by its Spanish initials. What Mexico's Federal Electoral Institute (IFE) did not take into account is that the postings were not only informing, they were providing valuable data that can be –and was- examined to check its "health". The bottom line is that the data presented is ill, so ill that it appears to have been given artificial life by a computer algorithm.

What the web surfers saw is that after an initial strong showing, which began at Sunday noon with a Calderon advantage of more than 4% over López Obrador ("AMLO"), the lead began to decrease in percentages. The diminishing trend continued and, around midnight, many of us went to bed forecasting a tie by 3:00 AM Monday, and an AMLO advantage of about 1% by wake up time on Monday. The morning surprise was that the trend had changed overnight and Calderon appeared with a slim but *invariant* advantage of about 1%; this sent many of us to what we, physics professors, do for a living: data analysis.

By Monday afternoon the first sets of PREP data began to circulate on blogs and chat rooms, and the hints of manipulation began to take shape. Mexico's UNAM physicist Luis Mochan and countless anonymous contributors helped to put the picture together.

The Data

After digging data from several independent sources and confirming its reliability, the first sign of concern appeared when plotting the trends posted by the PREP as a function of time. The similarity between the curves of the votes belonging to different candidates was surprising: it presented a constant percentage-wise advantage of one candidate over the others as shown in the figure.



This mirroring effect is not to be expected as the votes being counted arrived from different parts of the country where the support of the different candidates varied by huge factors.

The "Scoop"

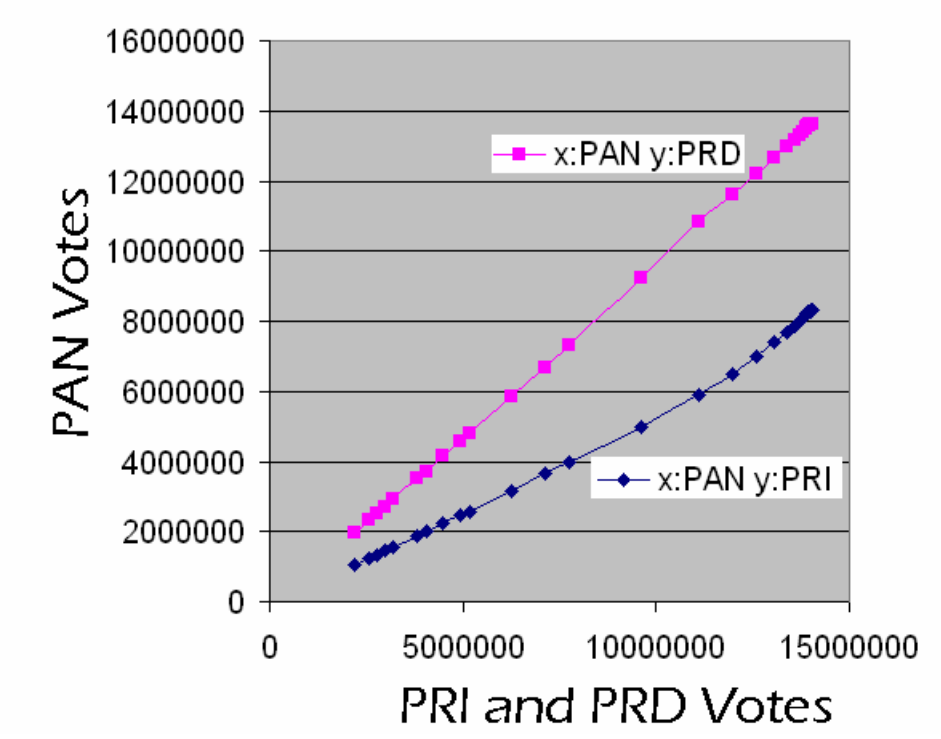
The immediate question was: how to quantify this abnormality? The obvious answer is by means of a test, for instance the Pearson's product moment correlation coefficient, which is used anywhere from social science to engineering. The coefficient is defined by

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

which, in plain English, determines if two variables vary together; a coefficient of zero means independence between the variables, a value of one means total dependence. Not surprisingly, the Pearson coefficient of the PAN-PRD voting trend was found to be 0.999974! [For comparison values of over 0.80 are generally viewed, by eg. NASA teams, as an indication of reliability.] Correlations of other curves were found to be 0.998205 (PAN-PRI) and 0.998196 (PRI-PRD); it was obvious that the control had been established over the PAN-PRD link, and – more important- it was now clear that the data was, if not fake, at least modified, *scooped* to put it in AMLO's southern Mexican jargon.

The Algorithm

Once a relationship had been uncovered, the next question was; what type of a relationship was imposed on the artificial votes? As the curves look extremely parallel one could expect a linear relationship between the voting trends. This is confirmed by the next graph.



The plot shows, in the y axis, the number of votes the PAN had at the time when the PRD and PRI had the votes in the x coordinate. The linear relationship is obvious. A quick fit to the PAN-PRD line produces the expression

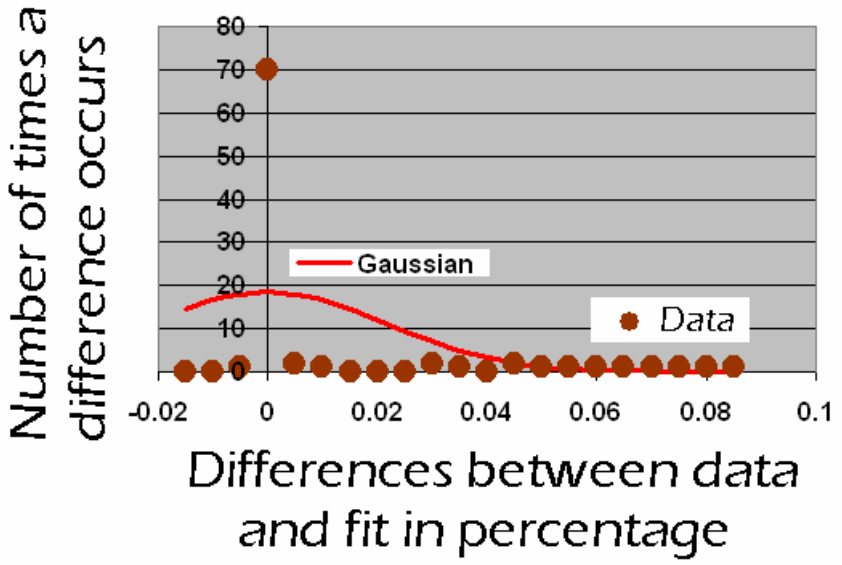
$$Y_{PAN}=279926.7904 + 1.008060294\ X_{PRD}.$$

Noteworthy is that the Calderon advantage reported by the IFE is of 257,532 votes, down from the 402,708 initially reported, and more in line with the intercept determined by the fit.

Goodness of fit

As any statistician would argue, linear trends are not proof of data manipulation. With this in mind the next question to answer is: can this fit be obtained from a sampling of numbers? The answer this time comes from a study of the deviations that the data has with respect to the mean behavior. Normal samplings always show small deviations from a trend, and these deviations tend to follow what is known as a "*Gaussian*" distribution, also called "Normal" for its repeatability in natural processes.

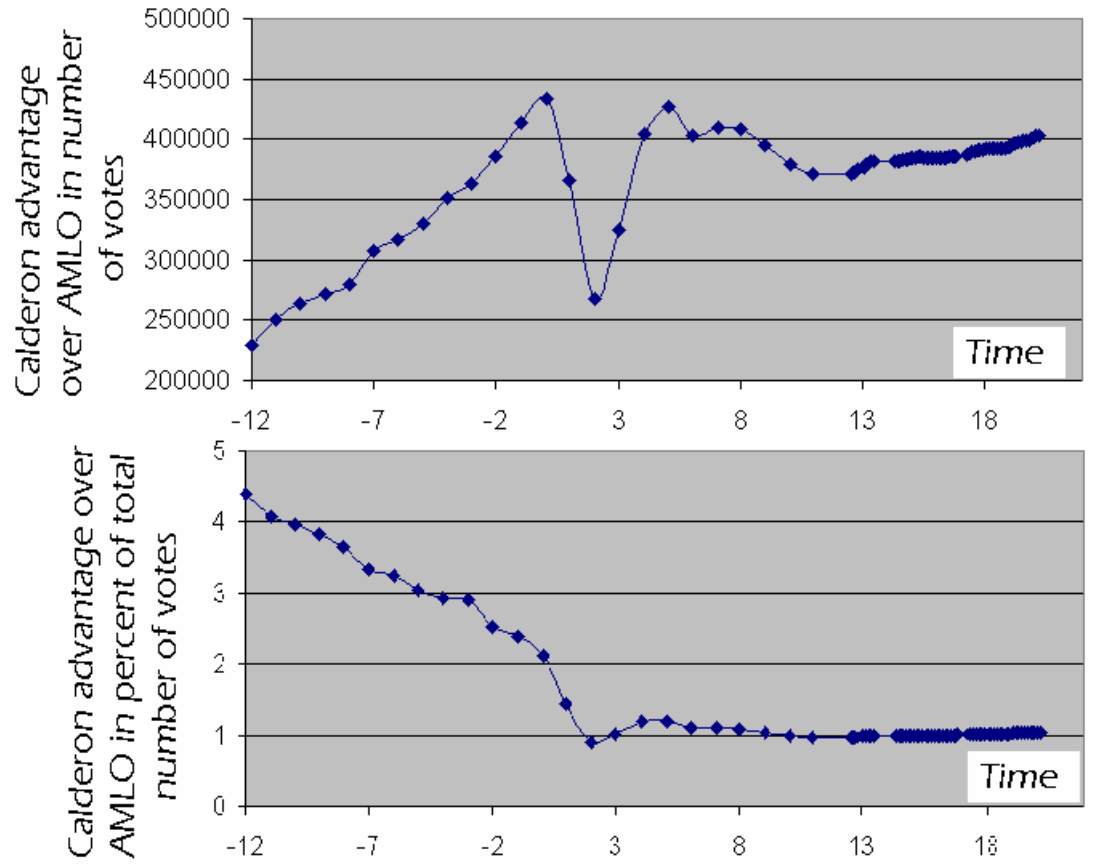
Looking at the differences between the data points of the PAN-PRD curve of the previous chart, and the analytic expression, one can obtain a distribution of these differences and plot them as a frequency chart as in the next graph.



With the dots representing the number of times a difference between the data and the fit occurs, and the red curve representing the expected normal distribution for such a sample, it is clear that the data does not follow a Gaussian distribution. The fact that a difference of zero percentage occurs many times more than other values, is a clear indication that the data was manufactured by an algorithm and does not stand a chance at passing as data originated at the actual voting.

The Scheme

What was then the scheme followed by the controllers of the PREP? This question can be answered by looking at the difference in votes between the two major candidates. The following two graphs shows such a difference both in real votes and in percentage (of the total number of votes) received as a function of time.



The plan is now easy to spot. Apparently the algorithm in operation in the PREP "count" was programmed to give Calderon an early large advantage to forge an illusion of invincibility, and press IFE into declaring Calderon as the winner at the Sunday 8:00 PM press conference. As the independent "*rapid count*" that IFE did under an independent group of five scientintists did not ratify PREP's fictitious advantage, the announcement of the winner was postponed to the 11:00 PM conference.

As the decreasing trend of Calderon's advantage (in percentage) continued, the announcement was again postponed, and the program apparently entered into a second mode of operation in which the fall of the advantage accelerated. [That's when many of us went to bed with a positive forecast for AMLO in mind.] But then, around 3:00 AM Monday, the code entered into a third mode of operation and constrained the Calderon-AMLO differences to about 1%. Incredible as it sounds, the relationship between the votes of the two top candidates kept on following the linear relationship imposed from the beginning.

What Lies Ahead?

As I finish writing this manuscript, a recount of the votes is taking place in all of Mexico. News from half an a hour ago (Wednesday 1:00 PM MST) show AMLO leading Calderon with an over 3% advantage with 37.32% of the ballots recounted. The moral of this exercise is twofold: 1) watch out for electronic methods of voting, and 2) never underestimate the power of statistics.