Análisis de datos de votación del referéndum reciente de Venezuela

Edward W. Felten
Departmento de Informática
Princeton University

Aviel D. Rubin
Departmento de Informática
Johns Hopkins University

Adam Stubblefield
Departmento de Informática
Johns Hopkins University

1 de Septiembre, 2004

Este articulo e informacion relacionada estan disponible en http://www.venezuela-referendum.com/.

1. Introducción

El 15 de agosto de 2004, los Venezolanos votaron en un referéndum nacional, para decidir si sacar a presidente Hugo Chavez de la presidencia. Luego, el CNE, la autoridad electoral venezolana (afectos a Chavez), anunció que aproximadamente 58% de los votantes habían votado en contra sacarlo, de modo que Chavez se quedaría como presidente. La elección fue realizada usando las máquinas de votación electrónicas de Smartmatic que marcan votos electrónicamente y producen un comprobante de papel comprobable de cada voto. Las figuras de la oposición han acusado al gobierno de fraude electoral.

Algunas de las alegaciones de fraude se pueden evaluar solamente a través una cuidadosa investigación y auditoria de los procedimientos y los documentos de la elección. No estamos en posición de realizar tal investigación, sino que esperamos que otros puedan hacerla. Otras alegaciones del fraude, sin embargo, si son verdaderas, crean anomalías estadísticas en los resultados de elección señalados. Este subconjunto de alegaciones de fraude se pueden evaluar por análisis estadístico.

Este trabajo describe nuestro análisis estadístico de los resultados de la elección señalados por el CNE. Los datos incluyen la cuenta de los votos de 19,055 máquinas de votación electrónicas. Recibimos estos datos de Sumate, un grupo de la oposición, que afirma que los datos corresponden con los resultados oficiales del CNE. Originalmente, el CNE había publicado datos por maquina en la Web, pero éstos fueron quitados, dejando únicamente los datos por mesa de votacion en el sitio web del CNE, alrededor del 18 de Agosto (véase http://www.cne.gov.ve/resultados/). Nuestro dataset está disponible en http://www.venezuela-referendum.com/.

Nuestro análisis no intentó producir ningún resultado determinado, sino que fue diseñado para analizar la cuestión técnica de si las anomalías afirmadas justifican una demanda del fraude. No tomamos ninguna posición con respecto a política venezolana en este documento, e impulsamos a los lectores validar o rechazar nuestro análisis estadístico por sus méritos, más que intentar predecir nuestros motivos.

Los miembros de la oposición han dicho que los resultados de elección señalados contienen anomalías estadísticas que no existirían en una elección hecha honestamente. Este trabajo considera si los modelos señalados por la oposición realmente indican fraude.

Acentuamos que hay tipos de fraude de elecciones que no crearían anomalías estadísticas y por lo tanto, no se podrían detectar por análisis estadísticos como los realizados por nosotros. Por ejemplo, si las máquinas

de votación electrónicas fueron programadas para cambiar cada voto de sí en un voto de no con una probabilidad de 10%, el análisis estadístico no podría detectar este tipo de fraude (sin embargo, probablemente podría ser detectado contando manualmente las comprobantes de votación). Nuestros resultados, en su mayoría, pueden mostrar si ocurrieron ciertos tipos de fraude; pero un análisis como el nuestro no puede eliminar la posibilidad de fraude.

2. Logística de la Elección

Según nuestros datos, la elección fue conducida en 4,582 lugares de votación. La mayoría de los lugares utilizaron las máquinas de votación electrónicas, pero algunos utilizaron votos de papel contados manualmente. Debido a que las demandas de fraude que estamos evaluando están relacionadas a las máquinas, no estamos considerando los lugares de votación que utilizaron el conteo manual, y consideramos solamente los 19,055 máquinas electrónicas de votación.

Cada lugar de la votación tenía una o más mesas, con una o más máquinas de votación en cada mesa. Los votantes fueron asignados a los lugares de la votación basados en donde viven, así que esperaríamos que los votantes en diversos lugares de votación se diferenciaran algo en factores demográficos y visiones políticas. Sin embargo, nos dicen que dentro de cada lugar de la votación, se asignaron a los votantes a una mesa y a una máquina al azar independientemente del sus puntos de vista políticas.

3. Alegaciones de fraude

Las figuras de la oposición afirman que los resultados señalados contienen varias anomalías estadísticas que no se habrían presentado al azar en una elección conducida honestamente. Tales anomalías, si existen, habrían podido presentarse debido a un fraude, o podrían existir debido a errores no intencionales en el manejo de las elecciones, o a fallos de funcionamiento en la programación de máquinas de votación electrónica.

Hemos recibido algunas alegaciones específicas del fraude. Entre los que circulaban en el Internet, hemos elegido mirar bien a dos posibilidades. La demanda más prominente es que el número de los votos del Sí fue predeterminado en las máquinas de votación, de modo que una vez que cierto número de los votos del Sí fuera marcado en una máquina de votación, la máquina contaría todos los votos subsecuentes como No, sin importar cómo los votantes eligieron. Si asumimos que cada máquina en un lugar de votación tenía el mismo patrón, tales patrones podrían ser detectados buscando los lugares de la votación que datos son consistentes con él. Particularmente, buscamos los lugares de votación que tenían dos máquinas con el mismo número de votos del Sí y con minoría de estos votos del Sí.

La otra alegación que se ha mencionado prominente es que dos o tres máquinas fuera de tres máquinas en la misma mesa tenían el mismo número de votos, Sí o No en una manera estadísticamente anómala. Examinamos las probabilidades de estas coincidencias del Sí o del No dentro de una mesa y los comparamos con los datos reales del CNE.

Nuestro objetivo fue determinar si las supuestas anomalías señaladas en los resultados oficiales son consistentes con la hipótesis que la elección fue conducida honestamente.

4. Fondo del análisis estadístico

Como algunos de los lectores quizás no saben mucho sobre análisis estadístico, vamos a repasar cómo tales análisis trabajan y qué conclusiones pueden sacarse de ellos.

Prueba De la Hipótesis

Los análisis estadísticos como los nuestros no pueden establecer la verdad de una hipótesis que es probada; pueden determinar solamente si los datos observados son consistentes o contrarios a la hipótesis. Si los

datos son contrarios a la hipótesis, ésta es evidencia en contra de la veracidad de la hipótesis. Pero si los datos son consistentes con la hipótesis, esto no establece que la hipótesis sea correcta. Si, por ejemplo, observamos que cada mañana canta un gallo y entonces sale el sol, esta observación es consistente con la hipótesis de que la canción del gallo causa la salida del sol, sin embargo no podemos deducir que la hipótesis es correcta.

Para ponerlo de otra manera, los análisis como el nuestro funcionan asumiendo que la hipótesis es correcta, y luego determinando las consecuencias que siguen de esa presunción. Si estas consecuencias son contrarias con los hechos observados, la inconsistencia es evidencia que la hipótesis (asumida) es incorrecta.

En nuestro caso, estuvimos probando la hipótesis de que la elección fue conducida honestamente. Si los datos señalados resultan ser contrarios con esa hipótesis, ésta será evidencia que algo incorrecto ocurrió en la elección.

Variación Al azar

Los procesos al azar exhiben una cierta variación por naturaleza. Por ejemplo, si lanzáramos 1000 monedas al aire, sabemos que en promedio, veríamos 500 caras. Sin embargo, las probabilidades son muy fuertes que no veríamos exactamente 500 caras, pero en lugar de otro considerarían que un cierto otro número cerca de 500. Nosotros podemos caracterizar esta clase de proceso al azar por dos números: la desviación "resultado malo" o medio, y "de estándar" que caracteriza la cantidad típica de variación al azar en el resultado. Por ejemplo, si en nuestro experimento de 1000 monedas el resultado promedio es 500 caras, y la desviación de estándar es cerca de 16.

Una regla general, que es verdad en todas las situaciones que encontramos en este papel, dice que cerca del 34% del tiempo el resultado de un experimento al azar será más de una desviación estándar lejano a la media , y cerca de 5% del tiempo más de dos desviaciones de estándar lejos del medio. Por ejemplo, en nuestro experimento de lanzar la moneda, el resultado estará fuera del rango 484-516 cerca de 34% del tiempo, y fuera del rango 468-532 cerca de 5% del tiempo. Si ensanchamos el intervalo, la ocasión que el resultado sea por fuera de nuestro intervalo ensanchado se hace más cercana a cero. Suponga que alguien nos dijo que hubieran lanzado 1000 monedas y hubieran conseguido 644 caras; este resultado es nueve veces la desviación estándar sobre el medio (es 644-500 = 144 sobre el medio, y 144 es 9x16). Un resultado desviaciones de este muchas estándar lejos del medio sucedería solamente 0.00000000000000002% del tiempo por al azar. Podríamos concluir con seguridad que mentía esta persona.

Evitar el error de la lotería

En el juego de Lotto de Nueva York, un jugador elige seis números entre 1 y 59. Entonces el estado elige aleatoriamente seis números de ese mismo rango. Un jugador gana si seis de los números que eligieron son los mismos escogidos por el estado. La probabilidad que un jugador gana es una en 45.057.474.

En el 28 de agosto de 2004, el estado trazó los números 6, 9, 34, 44, 49, y 58. El gráfico de estos números determinados era un acontecimiento muy inverosímil, que ocurriría solamente 0.000002% (1/45.057.474) del tiempo en una lotería justa (aleatoria). Con todo el hecho de que ocurrió este acontecimiento muy improbable, no es en sí mismo razón de sospechar un fraude. La razón de esto es que no hay nada especial sobre esta secuencia determinada de números - es solamente uno de los 45.057.474 acontecimientos igualmente improbable que habrían podido ocurrir. No importa que la secuencia saldría ese día, podriamos tener el mismo argumento.

Si discutiéramos (incorrectamente) que los resultados del 28 de agosto mostraron que la lotería del estado de Nueva York fue corrompida, estaríamos en error. Nuestro error se basaría en el hecho de que comenzamos con un conjunto muy grande de posibles patrones en la data; y luego seleccionar cuál patrón utilizar para explicar nuestros argumentos.

Para evitar el error de la lotería, necesitamos discutir si elegimos por adelantado buscar ese modelo como evidencia del fraude, o si hay alguna cierta razón para creer que ese patrón particular sería causado por algún probable mecanismo de fraude.

5. Lo que hicimos

Ya que probábamos la hipótesis que la elección fue conducida honestamente, comenzamos asumiendo que la elección fue conducida honestamente. (Acentuamos que este supuesto fue hecho hipotéticamente para el propósito del análisis.)

Bajo este supuesto, hicimos una serie de elecciones simuladas. En cada elección simulada, asumimos que los mismos votantes fueron a los mismos centros de votación y votaron de la misma manera que en los datos de la elección verdadera. Pero, dentro de cada centro de votación, reasignamos los votantes aleatoriamente a las máquinas de votación. En cada elección simulada, se colocó en cada máquina el mismo número de votos que fueron colocados en los datos de la elección real, pero asignando votantes diferentes (aleatoriamente elegidos) a cada máquina. Hicimos un total de 1.238 elecciones simuladas. Los datos de la elección verdadera, los resultados de estas elecciones simuladas, y el codigo del Python que usamos para generar y analizar los elecciones simuladas están todos disponibles en http://www.venezuela-referendum.com/.

Asumiendo que en la elección verdadera, los votantes de un centro de votación fueron asignados aleatoriamente a las máquinas de votación, o por lo menos independientemente de cómo votaron, esperaríamos que la elección verdadera fuera estadísticamente parecida a las elecciones simuladas.

6. Análisis

Para evaluar el reclamo de que habían límites pre-establecidos en el número de los votos de Si que se podrían contar en cada máquina, primero contamos el número de los lugares de la votación en los cuales dos o más máquinas tenían el mismo número de votos por el Sí y ninguna máquina con mayoría de votos por el Si. Esto corresponde a una situación en la cual el límite se ha alcanzado en por lo menos dos máquinas; observe que es difícil distinguir las máquinas con límites de las máquinas sin límites, a menos que un mínimo dos de las máquinas alcancen el límite. Llamaremos a esos lugares de votación "límite-constante."

En los datos de la elección verdadera, encontramos que 190 de los lugares de votación eran límite-constantes. En nuestras elecciones simuladas, había un promedio de 163 lugares límite-constantes con una desviación estándar de 12.33, un mínimo de 121, y un máximo de 204. Esto indica que los datos de la elección verdadera son 2.1 veces la desviación de estándar de la media, de modo que contáramos con tal resultado el cerca de 4% del tiempo. Es importante observar que ésta no es evidencia clara del fraude puesto que algunas de nuestras elecciones simuladas tenían aún más lugares límite-constantes, aunque no se limitó ninguna máquinas en nuestras simulaciones.

Para evaluar aun más la demanda de que existieron los límites, buscamos los lugares de votación en donde tres o más máquinas tenían el mismo número de los votos del Sí y ninguna máquina tenía mayoría de estos votos por el Si (es decir tres máquinas han alcanzado el límite). En los datos de la elección verdadera, cinco centros de votación cumplen con estos criterios; en las elecciones simuladas el promedio es 5.2 centros de votación. Este resultado es consistente con las máquinas que no son limitadas.

Para evaluar la demanda que había un número inusualmente alto de máquinas en el mismo centro con el mismo número de votos por el Si o por el No, contamos el número de tales "coincidencias" en ambos, los datos de la elección verdadera y los datos de nuestras elecciones simuladas. Esta estadística se muestra en la Tabla 1. Mientras que la desviación de las coincidencias del Si es mucho más alta que la desviación de las coincidencias de No, ésta no indica necesariamente el fraude por las razones explicadas abajo en la Sección 7.

	Si	No
Data de la elección verdadera	402	311
Promedio de elección simulada	360,90	317,35
Desviación estándar de elección simulada	17,75	16,99
Desviación estándar del promedio	2,3	0,37

Tabla 1: Comparación de las coincidencias entre la elección real y la simulada

7. Mecanismos del fraude posible

Como hemos dicho en la sección 3, examinamos las alegaciones de que había un límite en el número de los votos del Si en las máquinas dentro de un centro de votación, y que había un número inusualmente grande de coincidencias dentro de un centro, donde dos o tres máquinas tenían los mismos números votos por el Si o votos por el No. Claramente, un límite en el número de los votos por el Si habría podido predisponer con éxito la elección hacia el gobierno. Sin embargo, nuestro análisis no mostró ninguna evidencia que existió tal límite. El número de resultados repetidos quedó dentro del rango de los números predichos por nuestras simulaciones.

Nuestros experimentos si sugirieron que hubo una levemente alta incidencia de más alta de dos sí o de votos de no en cada mesa. La pregunta, entonces es si esto se relaciona o no con cualquier escenario realista del fraude. Nosotros no pudimos pensar en alguna manera con la cual alguien habría podido engañar y predisponer la elección en una dirección o la otra, que daría lugar a un número levemente más alto de estas coincidencias. Si el número de tales coincidencias fuera (por ejemplo) 100 desviaciones estándar de la media prevista, entonces tendríamos que concluir que hay algo inusual. Sin embargo, dado que el número levemente más alto de coincidencias es todavía dentro de los valores exhibidos en nuestros experimentos, nosotros atribuimos eso a la variación estadística más bien que a un fraude. Reevaluaríamos esa conclusión si alguien podría presentarnos una explicación creíble de cómo las coincidencias del Si y el No se habrían podido utilizar para favorecer un resultado determinado en la elección. Nosotros observamos que hay muchas cosas que pudieron haber sido medidas que resulten en estadísticas dentro de unas pocas desviaciones estándar de la media.. El hecho de que existe una variación en esta particular estadística, aparece ser más el producto de una búsqueda de anomalías estadísticas arbitrarias (es decir, una versión suave del error de la lotería) que evidencia de que hubo fraude en la elección.

8. Resumen

Después del referéndum el 15 de Agosto en Venezuela para que el presidente Chávez siga o no en el poder, grupos de la oposición examinaron los datos de la elección e hicieron acusaciones de fraude basándose en las anomalías estadísticas en los resultados de oficiales de la elección, demandando que esto no habría podido ocurrir si la elección fue ejecutada justamente. Sin embargo, nuestro análisis de los mismos datos, basado en simulaciones, no detectó ninguna anomalía estadística que indicara un evidente fraude en la elección.

Acentuamos que una falta de evidencia estadística no implica la ausencia del fraude. Sin embargo, elimina ciertas clases del fraude. En cualquier caso, el fraude que se alega que no es el tipo que esperaríamos que un gobierno que engañe emplearía. En detalle, creemos que las formas de fraude con mayor probabilidad de tener éxito como que las máquinas de votación cambien silenciosamente alguna fracción de los votos del Si a votos sopor el No dentro del computador, no produciría anomalías estadísticas observables.

La votación electrónica es más susceptible al fraude que mecanismos menos automatizados. El hecho de que la oposición sospeche altamente del resultado se debe, en parte, a que se decidió utilizar máquinas de votación en una elección simple de Si/No. Mientras que no encontramos ninguna evidencia estadística para las demandas de límites pre-establecidos en las máquinas u otras acusaciones específicas de fraude, nos preocupa que un amplio rango de fraude no detectable es mucho más fácil de generarse cuando se utilizan máquinas electrónicas de votación que cuando se utiliza, por ejemplo, un sistema basado en votos de papel.