Todas las cursivas corresponden a textos de Luis Mochán, o en su defecto, a lo publicado en su página de Internet (http://em.fis.unam.mx/public/mochan/elecciones/). No se realizaron modificaciones a los textos originales en ningún caso. Todos los textos en itálicas corresponden a las respuestas.

Pregunta (i) de Luis Mochán: *Las inconsistencias en la base de datos del PREP, sobre todo los puntos 3-6 de la lista http://em.fis.unam.mx/public/mochan/elecciones/#dificultadesprep y sobre todo, el que el número de votantes supere al número de boletas depositadas en al menos 220,000, i.e., hay 200K más votantes que boletas. Las bases de datos del Cómputo distrital no mencionan al número de boletas depositadas, por lo cual es imposible verificar si el problema se resolvió. Otra preocupación es el enorme número de actas incompletas y/o inconsistentes, que suman casi el 50%.*

Respuestas (i): De la lista a la que se hace referencia arriba:

1. *El número de registros que contiene (*la base de datos del PREP*) es 117,287. Como no he tenido tiempo de seguir las noticias no estoy seguro en cual de las cuentas entrarían los 13,200 registros faltantes necesarios para completar las 130,488 reportado en las páginas del PREP durante el conteo.*

El total de actas que suman las 130,788 está desglosado de la siguiente manera, e incluso puede descargarse cada uno de estos archivos desde la página en Internet:

http://www.ife.org.mx/prep2006/bd_prep2006/bd_prep2006.htm

En esta misma página puede descargarse un documento que explica las estructuras de los archivos, así como el contenido de cada uno de ellos:

http://www.ife.org.mx/prep2006/bd_prep2006/PREP2006_presidente_descripcion.pdf

Tabla 1. Resumen de Actas de Presidente. PREP 2006.

Tipo	Archivo	Actas
Actas Contabilizadas	*PREP2006-Presidente.txt*	117,287
Actas Inconsistentes	*PREP2006-Presidente-AI.txt*	11,184
Sin Acta	*PREP2006-Presidente-SA.txt*	1,637
Actas NO Recibidas	PREP2006-*Presidente-NR.txt*	380
Total Nacional		**130,488**
VMRE	PREP2006-*Presidente-VMRE.txt*	300
Total Nacional + VMRE		**130,788**

Actas Fuera de Catálogo	*PREP2006-Presidente-FC.txt*	**611**

2. *Ya conseguí también las bases de datos de senadores y diputados. Contienen 120,032 y 120,091 registros respectivamente. ¿Por qué difieren en alrededor de 2700 registros de la base para presidente?*

La diferencia de actas esperadas entre Senadores y Diputados contra Presidente es de **822** actas, las cuales son las actas para casillas especiales de representación proporcional que no se contemplan en la elección de presidente.

Para PRESIDENTE al sumar las 117,287 actas contabilizadas de PREP más las 11,184 actas con inconsistencias se tienen **128,471** actas procesadas sin contemplar las 300 actas de voto en el extranjero, ya que senadores y diputados no las contabilizan.

Para SENADOR al sumar las 120,030 actas contabilizadas de PREP, más las 9,150 actas con inconsistencias se tienen **129,180** actas procesadas.

Para DIPUTADO al sumar las 120,089 actas contabilizadas de PREP, más las 9,015 actas con inconsistencias se tienen **129,104** actas procesadas.

Por lo tanto, en realidad la diferencia de actas entre senadores y presidente es **709** actas (129,180 actas de senador – 128,471 actas de presidente).

Con respecto a la diferencia de actas entre senadores y presidente es **633** actas (129,104 actas de senador – 128,471 actas de presidente).

La diferencia de actas mencionada en los dos párrafos anteriores son las actas de representación proporcional para senadores y diputados de las casillas especiales que fueron procesadas.

A continuación se presentan las tablas que muestran el resumen de actas que pueden descargarse desde la página en Internet, mencionada en el punto 1.

Tabla 2. Resumen de Actas de Senadores. PREP 2006.

Tipo	Archivo	Actas
Actas Contabilizadas	*PREP2006-Senadores.txt*	120,030
Actas Inconsistentes	*PREP2006-Senadores-AI.txt*	9,150
Sin Acta	*PREP2006-Senadores-SA.txt*	1,714
Actas NO Recibidas	PREP2006-*Senadores-NR.txt*	416
Total Nacional		**131,310**

Actas Fuera de Catálogo	*PREP2006-Senadores-FC.txt*	372

Tabla 3. Resumen de Actas de Diputados. PREP 2006.

Tipo	Archivo	Actas
Actas Contabilizadas	*PREP2006-Diputados.txt*	120,089
Actas Inconsistentes	*PREP2006-Diputados-AI.txt*	9,015
Sin Acta	*PREP2006-Diputados-SA.txt*	1,791
Actas NO Recibidas	PREP2006-*Diputados-NR.txt*	415
Total Nacional		**131,310**

Actas Fuera de Catálogo	*PREP2006-Diputados-FC.txt*	315

3. *Además de los registros faltantes, hay otros 22,538 que tienen un asterisco ('*') en alguno de los campos numéricos. El problema me saltó a la vista al tratar de checar la consistencia de los datos numéricos. Aquí guardé la base de datos correspondiente a estos registros incompletos.*

Existen 24,148 registros en la base de datos de registros que se sumaron a los resultados del PREP, aun cuando estas actas contenían uno o más campos ilegibles o vacíos. Debido a que estos errores en el llenado del acta no se encontraron en alguno de los campos de votación por partido o coalición, no se restringió la contabilización del acta en los resultados sumados por el PREP[1].

4. *Eliminando los registros con asteriscos, hay 27,073 registros que considero inconsistentes, pues la suma de los campos PAN, ALIANZA_POR_MEXICO, POR_EL_BIEN_DE_TODOS, NUEVA_ALIANZA, ALTERNATIVA_SOCIAL_DEMOCRATA, NO_REGISTRADOS y NULOS no es igual al número de BOLETAS_DEPOSITADAS. Aquí guardé la base de datos correspondiente.*

La cantidad de actas con una diferencia en esta suma, contabilizando las 128,471[2] actas para presidente procesadas en el PREP, es de:

Tabla 4. Actas con diferencia en la suma de votación emitida. Elección presidencial. PREP 2006.

	TOTAL EN PREP
TIPO DE ERROR	Número de Casillas
(Votación Total) MAYOR QUE (Boletas Depositadas)	**31,630**
(Votación Total) MENOR QUE (Boletas Depositadas)	**15,539**
Totales	**47,169**

5. *El NUMERO_VOTANTES siempre es consistente con la suma de PAN+ALIANZA_... Habiendo tantos errores en otros campos es sorprendente que en este campo no haya un solo error en más de 117,000 registros. El NUMERO_VOTANTES ¿fue uno de los campos que llenaron los funcionarios de casilla al llenar las actas? De ser así, la ausencia de errores no es sorprendente sino imposible. Tal vez este campo no se tomó de las actas, sino que fue calculado por las computadores del IFE, definiéndolo como la suma de votos por partidos mas no registrados mas nulos. Verifiqué que el NUMERO_VOTANTES se conserva consistente aún si reemplazo todos los asteriscos por ceros en lugar de eliminarlos. Por lo tanto, en los análisis subsiguientes realizo dicha modificación.*

El NUMERO_VOTANTES es el total de votos, y equivale a la suma de votos para todos los partidos políticos o coaliciones más los votos para candidatos no registrados y votos nulos. La etiqueta no es clara en este sentido. Sin embargo, los archivos pueden descargarse desde la página de Internet:

http://www.ife.org.mx/prep2006/bd_prep2006/bd_prep2006.htm

Estos archivos contienen un campo llamado TOTAL_CIUDADANOS_VOTARON, el cual es el dato capturado por los funcionarios de casilla, y el campo TOTAL_VOTOS que refleja la suma total de votos para los partidos políticos, coaliciones, candidatos no registrados y votos nulos.

En realidad la cantidad de actas en donde difiere la cantidad de ciudadanos que votaron contra el total de votos, incluyendo las actas inconsistentes es de **64,123** de **128,471** actas procesadas.

6. *Reemplazando los asteriscos por ceros, obtengo que la suma de las BOLETAS_DEPOSITADAS es 35,876,783 y la de los NUMERO_VOTANTES es 38,516,730, por lo cual parece haber 2,639,947 más votos que boletas depositadas en las urnas. Por otro lado, si elimino los registros con asteriscos, obtengo 35,876,783*

[1] Acuerdo del IFE con los representes de los partidos políticos y coaliciones el 10 de febrero de 2006.
[2] No se incluyen las 300 actas de VME.

boletas depositadas y 36,100,471 votantes, 223,688 más votantes que boletas depositadas.

El dato referente a que la suma de las BOLETAS_DEPOSITADAS es 35,876,783 y la de los NUMERO_VOTANTES es 38,516,730, por lo cual parece haber 2,639,947 más votos que boletas depositadas es correcto. El primero de estos dos campos es un dato que se captura directamente del acta, por lo que los errores de cálculo provienen de un mal llenado de las actas. Es importante mencionar que en un número importante de actas, dichos campos (ciudadanos que votaron, boletas depositadas en la Urna, etc) no son siempre llenados correctamente o simplemente son dejados en blanco, por lo que estas cifras nunca cuadran con el total de votos.

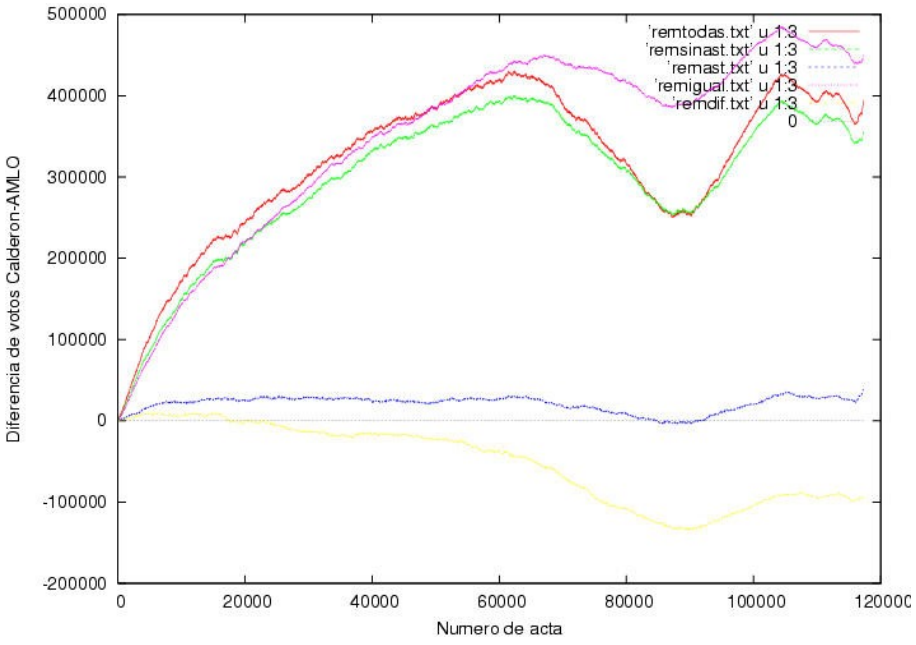
Pregunta (ii) de Luis Mochán: *Me preocupan también la lista de comentarios que precede a mi figura 23.1 http://em.fis.unam.mx/public/mochan/elecciones/#fig23.1 referente al PREP. En particular, me sorprende que haya una correlación entre los errores del PREP y la preferencia electoral. Por ejemplo, casi el 20% de las actas que sí entraron al PREP están incompletas[3]. Yo hubiera esperado que su contribución a la ventaja de Calderón sobre AMLO fuera aproximadamente del 20%, pero resulta ser mucho menor al 20% del total. Otro 20% de las actas tienen inconsistencias. Su contribución a la ventaja de Calderón sobre AMLO no sólo no es la fracción correspondiente del total sino que ¡tiene el signo contrario!*

De la liga a la cuál se hace referencia arriba:

La figura 23.1 es equivalente a la figura 5, pero elaborada con todos los datos de la bases de datos del PREP, casilla por casilla. Se muestran varias curvas correspondientes a:

a. *¿Por qué se correlacionan los errores y las inconsistencias con la preferencia hacia AMLO?*

Figura 23.1



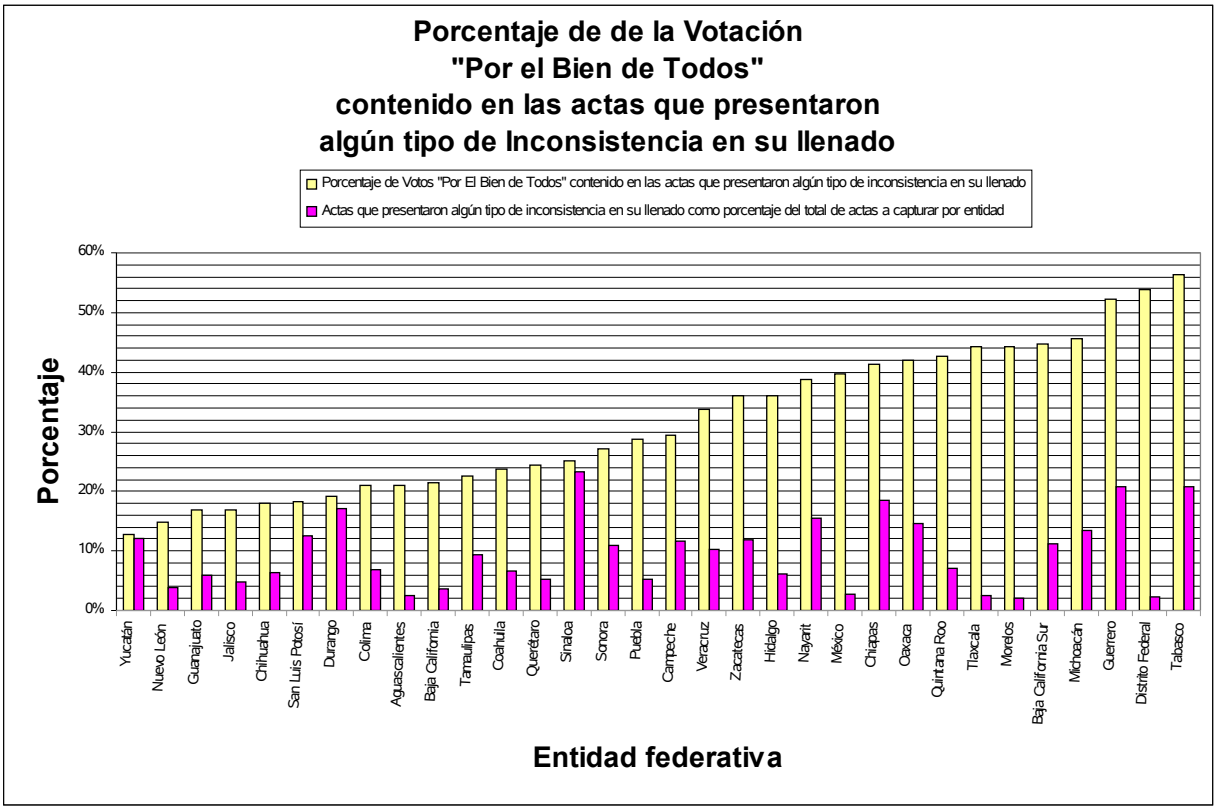
[3] A este respecto, favor de referirse al punto anterior.

Respuesta (ii): Las actas que presentan inconsistencias no se relacionan directamente con la votación a favor de la Coalición por el Bien de Todos (elección presidencial).

Las actas de la elección para presidente que presentaron algún tipo de inconsistencia en su llenado fueron capturadas y contabilizadas en una "base de Inconsistencias", durante el periodo de operación del PREP (http://www.ife.org.mx/prep2006/bd_prep2006/bd_prep2006.htm). Estos registros no fueron sumados a los resultados del PREP, pues tenían un carácter de no publicable[4], de acuerdo con los criterios establecidos con los representes de los partidos políticos en febrero del 2006.

El número de actas que presentaron inconsistencias no publicables es de 11,184 (8.55% del total de actas esperadas). No hay posibilidad de que exista una correlación inducida entre las actas con inconsistencias y el voto por la Coalición por el Bien de Todos (CBT), ya que las actas que no se publicaron provinieron de todas las entidades federativas. El gráfico 1 muestra el porcentaje de votación a favor de la CBT contenido en las actas que presentaron inconsistencias por entidad, y lo compara con el porcentaje el porcentaje de actas con inconsistencias.

Gráfico 1. Votación acumulada a favor de la CPBT, contenida en actas que presentaron inconsistencias. Nivel entidad, elección presidencial.



Al realizar una prueba de correlación de Pearson[5] a los datos a la votación porcentual a favor del candidato de la CBT contra el porcentaje de actas que presentan inconsistencias por entidad, se obtuvo una coeficiente de 0.231. Lo anterior indica que la relación entre ambas es muy pequeña.

[4] Los criterios bajo los cuales un acta que presentara errores en su llenado era integrada a la base de datos de actas con inconsistencias, y no a la suma de resultados difundidos por el PREP, se establecieron por acuerdo con los representes de los partidos políticos y coaliciones el 10 de febrero del 2006.

[5] A las variables involucradas se les aplicó una prueba de normalidad, misma que indicó en ambos casos, que no se puede rechazar la hipótesis de que se distribuyan normalmente con un 95% de confianza.

Claramente podemos observar que no hay correlación evidente entre las inconsistencias con la preferencia por el candidato de la CBT.

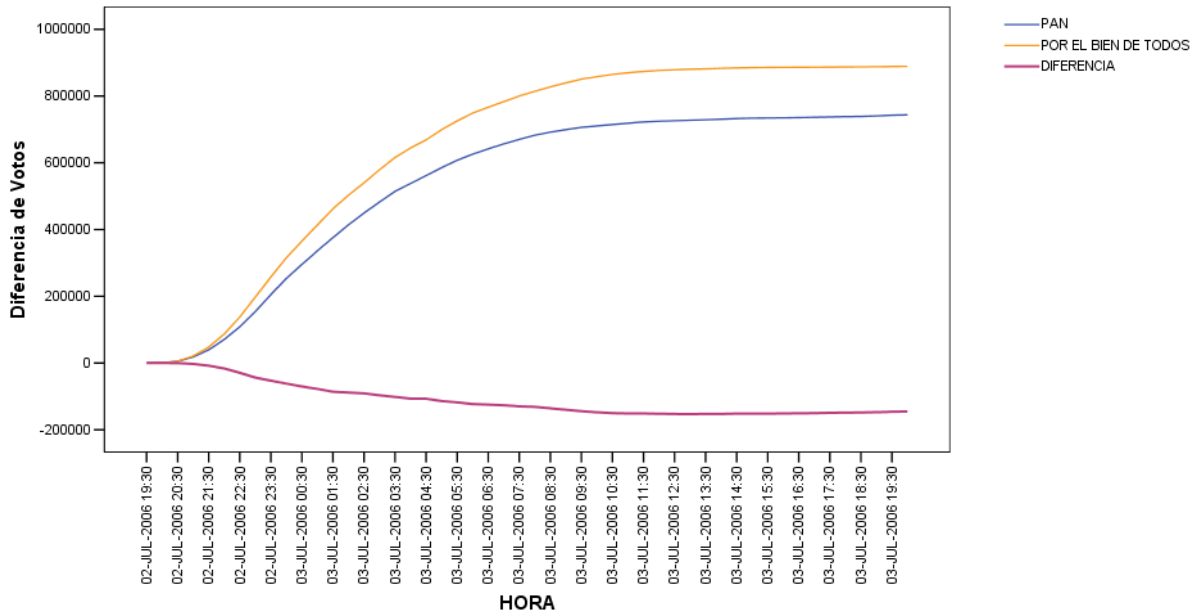
Efectivamente, la mayor cantidad de votos en las actas con inconsistencias fueron para la CBT, sin que por este motivo haya una relación causal en ello. A continuación se muestra una tabla con la cantidad de votos por partido contenidos en la base de inconsistencias.

Tabla 5. Votación por partido o coalición. Actas que presentaron inconsistencias en su llenado; 11,184 registros. Elección presidencial. PREP 2006.

PAN	ALIANZA POR MÉXICO	ALIANZA POR EL BIEN DE TODOS	NUEVA ALIANAZA	ALTERNATIVA SOCIALDEMÓCRATA	CANDIDATOS NO REGISTRADOS	NULOS	TOTAL
743,795	**809,003**	**888,971**	**13,946**	**28,040**	**15,019**	**82,452**	**2,581,226**

El siguiente gráfico muestra la votación acumulada para el PAN y la CBT, así como la diferencia entre dicha votación acumulada (PAN y CPBT, línea roja), tal como se calcula con base en lo asentado en las actas que presentan inconsistencias. Dicho diferencial muestra una trayectoria consistente con los resultados del Programa de Resultados Electorales Preeliminares. Es decir, la contribución de votos contenidos en las actas con inconsistencias no muestra “baches” ni estancamientos en su trayectoria, y por lo tanto no hay ningún indicio de una supuesta manipulación de las actas para perjudicar a algún partido o coalición en particular.

Gráfico 2. Actas con diferencia en la suma de votación emitida. Elección presidencial. PREP 2006.



Dado las consideraciones anteriores, no se puede aseverar que existe una correlación entre la votación que recibe la CBT y las actas que presentan inconsistencias.

Pregunta (iii) de Luis Mochán: *Los aspectos que me llamaron la atención de los Cómputos Distritales están en:* ***http://em.fis.unam.mx/public/mochan/elecciones/#analisisconteo*** *Muchos de ellos ya me los aclararon (aunque no he tenido tiempo de explicarlos en mi página), pero me sorprende mucho el punto 14; si usáramos las 13.5K actas que no llegaron al PREP para hacer una estimación del resultado de la elección cometeríamos un error grave, pues obtendríamos que AMLO hubiera ganado por más de 4.5% y que Madrazo casi hubiera empatado con Calderón. Es decir, este 10% de las casillas resulta no ser un buen muestreo del total. ¿Qué podría provocar una correlación entre los problemas que evitaran que estas actas llegaran al PREP y la preferencia electoral? Mi figura 40 muestra que la correlación está bien escondida pues las actas referidas contienen resultados que cubren todo el posible espectro de preferencias electorales. Un análisis de un colega (referido en mi página) muestra que en algunos estados las casillas correspondientes a estas dificultades son un excelente muestreo, como yo hubiera esperado, y en otros es un pésimo muestreo.*

Respuesta (iii):

El número de actas que no fueron sumadas a los resultados sumados del Programa de Resultados Electorales Preeliminares (PREP) fueron 11,184 para la elección de Presidente.

De estas 11,184 casillas, 3,846 son urbanas y 7,338 no urbanas. Si consideramos que el PAN recibió mayor votación en las zonas urbanas, mientras que la CBT obtuvo mayor número de votos en las zonas no urbanas, es posible comprender por qué las actas con inconsistencias concentran mayor votación para la CBT que para el PAN.

Bajo ninguna circunstancia podría considerarse que las 11,184 actas que presentan inconsistencias constituyen una muestra representativa.

De las 130,488 casillas aprobadas por los consejos a nivel nacional 90,286 (69.19%) son casillas urbanas; mientras que 40,202 (30.81%) son casillas no urbanas. En contraste, en el universo de actas con inconsistencias, 34.4% son urbanas y 64.6% son no urbanas. Así como

las actas que presentan inconsistencias no representan una muestra que permitiera predecir tendencias, tampoco es ése el objetivo de los sistemas PREP y Cómputos Distritales, como sí lo es para el Conteo Rápido.

Pregunta (iv) de Luis Mochán: *Otro aspecto de los Cómputos que me llama la atención está en el punto 26, i.e., en aquellas casillas donde hubo cambios respecto al PREP. Los cambios entre PREP->Cómputos a veces favorecen y a veces perjudican a los partidos. Para Madrazo y para Calderón fueron favorecidos/perjudicados 50%+-5% de las veces. AMLO fue favorecido en el cambio el 63% de las veces. Pero además, mientras que Calderón y Madrazo ganaron y perdieron un número similar de votos, AMLO ganó dos votos por cada uno de los que perdió en el cambio. ¿Es razonable? ¿No sería razonable esperar que cada candidato sea favorecido y perjudicado en el cambio con la misma probabilidad?*

Respuesta (iv):

Respecto al análisis que presenta las diferencias entre el PREP y el Sistema de Cómputos Distritales (CD), con relación a la diferencia de votos capturados para cada partido político y coalición (elección presidencial), la información que ofrecen dichos sistemas no coincide con lo que argumenta Luis Mochán. El análisis arroja los siguientes resultados:

Dado que el PREP pretende procesar el mayor volumen de información en el menor tiempo posible, dichas actas son registradas en el sistema tal cual, sin detenerse a definir si hubiese o no algún error o faltante de información. Por otro lado el Cómputo Distrital se realiza en presencia del Consejo Distrital, quien se encarga de cotejar el contenido de cada una de las actas para finalmente acordar, con el consentimiento de los representantes de los partido políticos, los resultados de la votación de cada una de ellas.

Cada acta representada en la base de datos es clasificada de acuerdo al tipo de inconsistencias que pudiera o no presentar y clasificada de acuerdo a su estado con respecto a su registro en el sistema.

De acuerdo con las inconsistencias reportadas para cada acta y al criterio acordado con partidos políticos, éstas se pueden clasificar como publicables o no publicables. Las actas no publicables son registradas en el sistema pero no son sumadas en los totales de los resultados preliminares debido a las inconsistencias que las clasifican como tales.

Tomando en cuenta los puntos anteriores, para realizar una comparación entre los dos sistemas en cuestión se deben de considerar los siguientes criterios:

1. Solamente se deben comparar aquellas actas que se encuentren registradas en el sistema PREP al cierre de su funcionamiento, ya que todas aquellas actas que no fueron reportadas dentro del sistema no pueden ser comparadas contra sus homólogas de CD.

2. Al comparar acta por acta, es necesario tener en cuenta que aquellas actas que presentan alguna inconsistencia serán diferentes en algunos campos con respecto a las actas registradas por cómputos, consecuencia de la misma definición de acta inconsistente.

A partir de los criterios recién citados, presentamos las siguientes comparaciones de resultados PREP – CD para la elección de presidente a nivel nacional: Totales de la votación considerando todas las actas registradas en el sistema PREP incluyendo actas con inconsistencias publicables y no publicables.

Tabla 6: Comparación nacional PREP - CD computando todas las actas registradas en el sistema PREP2006 (incluyendo actas con inconsistencias).

Partido Político o Coalición	Votación Nacional *PREP*	Votación Nacional *Cómputos Distritales*	Diferencia en votos (CD – PREP)	Diferencia porcentual en votos (CD – PREP)/PREP
PAN	14,751,993	14,799,267	47,274	0.32%
APM	9,126,529	9,153,187	26,658	0.29%
PBT	14,502,387	14,558,330	55,943	0.39%
NA	398,135	394,384	-3,751	-0.94%
ASDC	1,113,119	1,114,540	1,421	0.13%
No Registrados	296,135	294,562	-1,573	-0.53%
Nulos	909,658	889,918	-19,740	-2.17%

Nota: Ningún cálculo de la tabla incluye las actas de voto de mexicanos residentes en el extranjero.

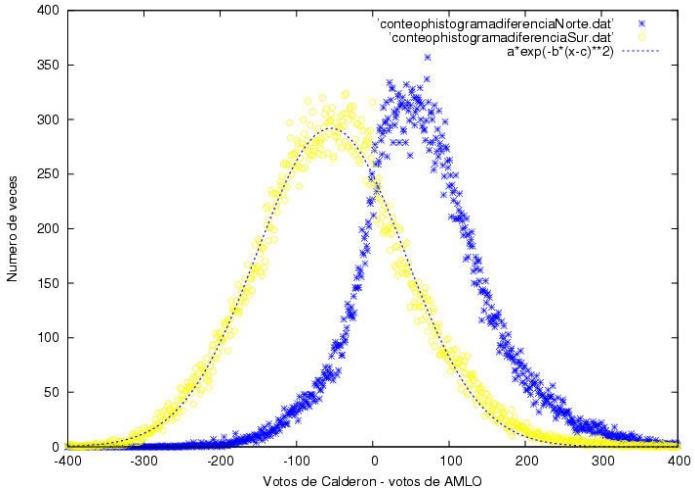
Como se puede leer de la tabla 6, tanto el PAN como la APM y la CBT obtienen más votos al completarse los cómputos distritales. La mayor variación porcentual se observa para la votación a favor de la CBT, seguido por el PAN y la APM.

Pregunta (v) de Luis Mochán: *Sobre las estadísticas de la elección, me sorprenden las figuras 34-38. Los histogramas que muestran la diferencia de votos entre Calderón y AMLO son una bonita Gaussiana en el sur y una curva extraña en el norte, la cual se aproxima a una Gaussiana al eliminar las contribuciones de ciertos estados. Las contribuciones de dichos estados sugieren (desde luego no demuestran) una manipulación de los resultados.*

De la página de Luis Mochán, gráficos referidos en la pregunta.

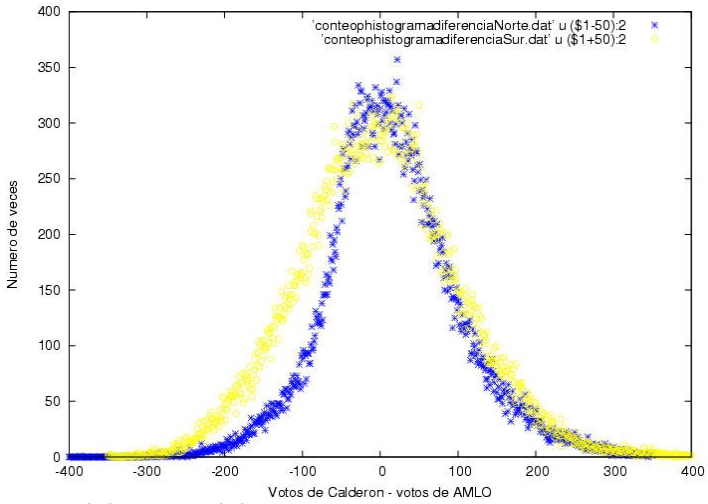
a. *La figura 34 muestra el mismo histograma que la figura 33, pero separado en contribuciones provenientes de los estados del norte y del sur, como en las figuras 29.1 y 29.2. La figura 34 muestra que mi interpretación original de las figuras 21.5 y 33 es errónea, como me había advertido la Dra. Gloria Koenigsberger. No se trata de una curva normal cuyo pico se ve desplazado hacia la derecha, sino de la suma de dos curvas, una centrada alrededor de -50 (más o menos) correspondiente al Sur, en la que AMLO domina las preferencias, y otra centrada alrededor de 50 correpondiente al Norte, y en la cual es Calderón quien domina las preferencias. La curva correspondiente al Sur se puede ajustar relativamente bien por una Gaussiana (a propuesta de Jaime Ruiz) de la forma a*exp(-b(x-c)^2), donde x es la diferencia de votos y a=292.1+/-1.0, b=5.28e-05+/-4e-07 y c=-54.8+/-0.4 son los parámetros de ajuste (línea punteada). Por otro lado, la curva correspondiente al norte no se parece a una Gaussiana ni a una Lorentziana. La curvatura en las colas donde número de veces es menor a 100 y la de la cima cima no es consistente con la subida donde el número de veces pasa de 100 a 275. Además, la curva es bastante asimétrica. Las diferencias entre la forma de las dos distribuciones se vuelven evidentes si las desplazamos horizontalmente para que se superpongan. En la figura 35 muestro las curva para el Norte desplazada 50 votos hacia la izquierda y la curva para el Sur deplazadas 50 votos hacia la derecha. Los astrónomos reconocerán en la curva del Norte el llamado Perfil P Cisne (según Gloria Koenigsberger), correspondiente al espectro que describe el color de la luz proveniente de ciertas estrellas cuya radiación es selectivamente absorbida por el viento estelar.*

   Figura 34



Datos del norte y del sur.

Figura 35

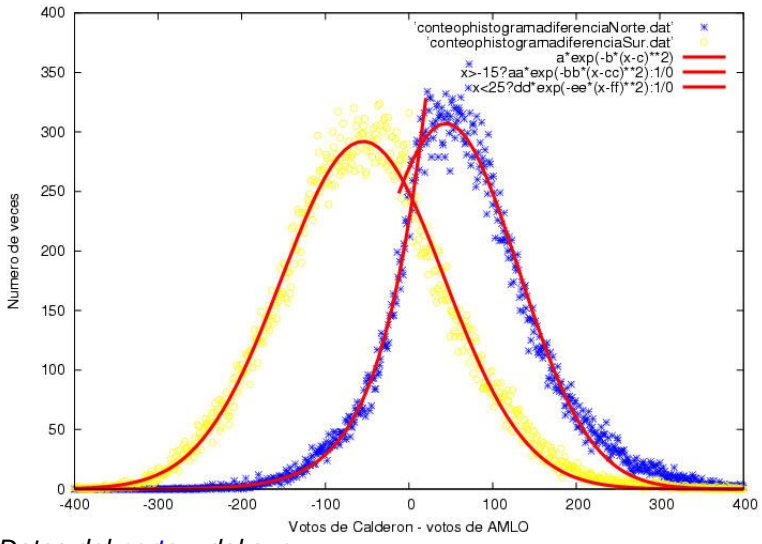


Datos del norte y del sur.

a. *Gerardo Horvilleur hizo la observación de que el lado derecho de la curva azul en la figura 35 no es demasiado distinta al lado derecho de la curva amarilla, descrita por una curva normal, mientras que el lado izquierdo difiere notablemente. Además, observó que el cambio de comportamiento coincide con la región donde las curvas azul y amarilla se intersectan en la figura 34, es decir, en aquella zona de la gráfica donde AMLO le lleva una ventaja ligera a Calderón en la región norte. Jaime Ruiz estudió estas curvas y obtuvo que se pueden describir como dos lorentzianas distintas. La figura 36 ilustra la misma idea pero empleando ajustes gaussianos. Como habíamos visto en la figura figura 34, la distribución del sur puede ajustarse bien a una curva normal. En cambio, es necesario dividir la distribución del norte en dos intervalos, cada uno descrito por una gaussiana de la forma a*exp(-b(x-c)^2) pero con parámetros a,b,c muy distintos. Una describe la región en que Calderón le gana a AMLO. Los parámetros correspondientes (a=307.1+/-1.5, b=6.71e-05+/-1.6e-06 y c=44.0+/-1.2) son similares a los de la gaussiana que describe la votación en el sur (a=292.1+/-1.0, b=5.28e-05+/-4e-07) excepto por la posición (c=-54.8+/-0.4) del máximo. Otra describe la región donde AMLO le gana a Calderón. Sus altura (a=73000) es ridículamente alta, lo cual indica que dicha región es muy anómala. Ambas gaussianas se cruzan donde la diferencia de votos es casi nula y a partir de ese punto se alejan muy rápidamente*

entre sí y de los datos subsiguientes. ¿Por qué la estadística en el sur, mayoritariamente perredista, es normal, mientras que la estadística en el norte, mayoritariamente panista, muestra una fuerte anomalía, pero sólo en el intervalo donde AMLO tiene más votos que Calderón?

Figura 36

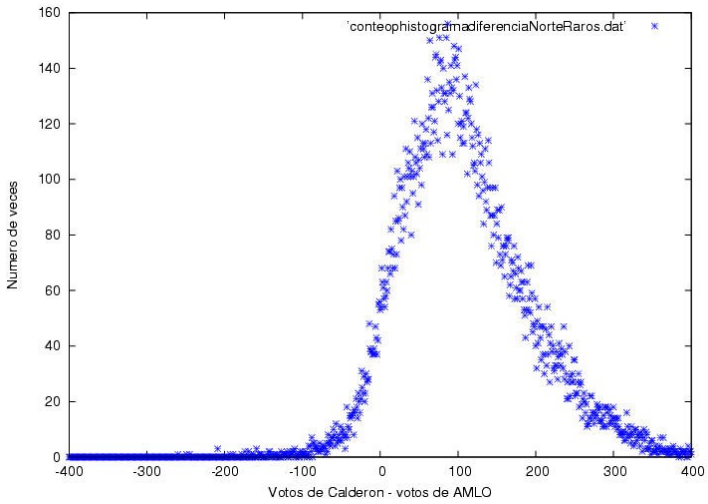


Datos del norte y del sur.

Indice

*Siguiendo sugerencias de Gerardo Horvilleur y de Jaime Ruiz, hice una búsqueda de uno en uno de aquellos estados que pudieran haber dado origen al comportamiento singular de las funciones de distribución de votos. En la figura 37 muestro el histograma de diferencia de votos correspondiente a los estados de Chihuahua, Guanajuato, Jalisco y Nuevo León. Se ve completamente anómalo, es muy asimétrico y tiene una enorme dispersión cerca del máximo. Aunque más importante es que al excluir dichos estados de la lista previa de estados del norte, el histograma correspondiente a todos los estados del norte restantes, mostrado en la figura 38 parece ser moderadamente normal, mucho más que el histograma mostrado en la figura 36. Sin embargo, el ajuste gaussiano a*exp(-b*(x-c)**2) con a=206+/-1, b= 0.000122+/-1.5-06, c=37.1+/-0.4) deje mucho que desear aún y hay permaece una dispersión extraña cerca del máximo.*

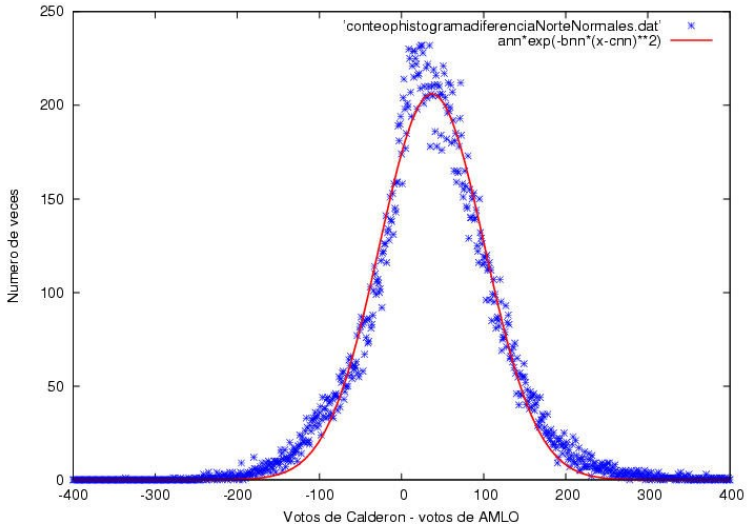
Figura 37



Datos.

Indice

Figura 38



Respuesta (v):

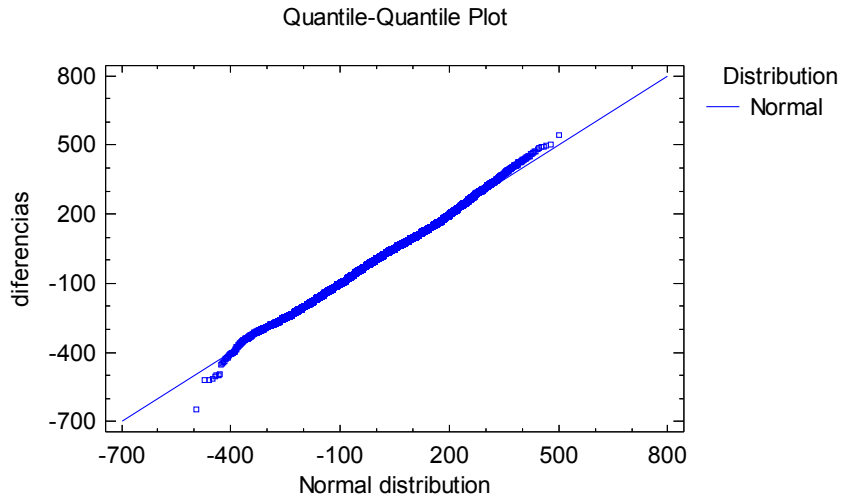
No hay evidencia de la supuesta manipulación de los datos. El histograma de diferencias entre la votación para el PAN y la CBT mostrado en la figura 33 está realizado con 130,673 observaciones, lo que no cuadra con ninguna combinación de las bases de datos del PREP (elección presidencial), que se encuentran disponibles en la página del IFE, y el rango de las diferencias abarca desde -400 a 400, lo que excluye las colas más pesadas en ambos sentidos.

Considerando la base de datos con los registros de la información sumada a los resultados del PREP – la cual contiene 117,287 observaciones – se tiene que el rango para la variable de diferencias en las votaciones para el PAN y la CBT va desde -648 hasta +543. Los histogramas

considerados en la página de Luis Mochán presentan un rango menor, lo que se puede reflejar en sesgos en la distribución de las variables.

Al aplicar pruebas estadísticas de ajuste a una distribución (Kolmogorov-Smirnov) para la variable de diferencias, encontramos que se rechaza la hipótesis de que los datos provengan de una distribución Normal, y observando el comportamiento del ajuste de la variable a dicha distribución, observamos que el rechazo de tal hipótesis proviene del efecto de las colas, que aparentemente se han excluido en el histograma general. El gráfico Q-Q siguiente muestra el efecto de las colas a partir de los rangos observados.

Gráfico 3. Q-Q en las colas de la distribución en el diferencial de la votación.



Al realizar el análisis de las variables de cantidad de votos para el PAN y la CBT, observamos que ninguna de ambas variables cumple con una distribución normal. Esto lo indican las medidas tales como el sesgo y la kurtosis, así como los gráficos Q-Q que se presentan a continuación:

Gráfico 4. Q-Q en las colas de la distribución en la votación para el PAN.

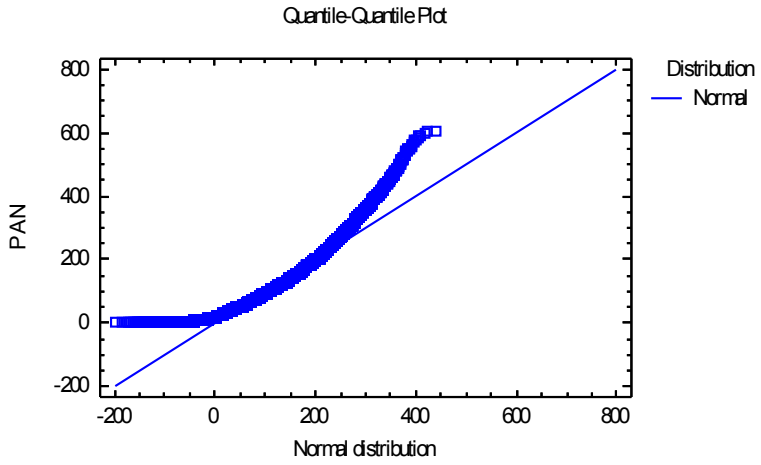
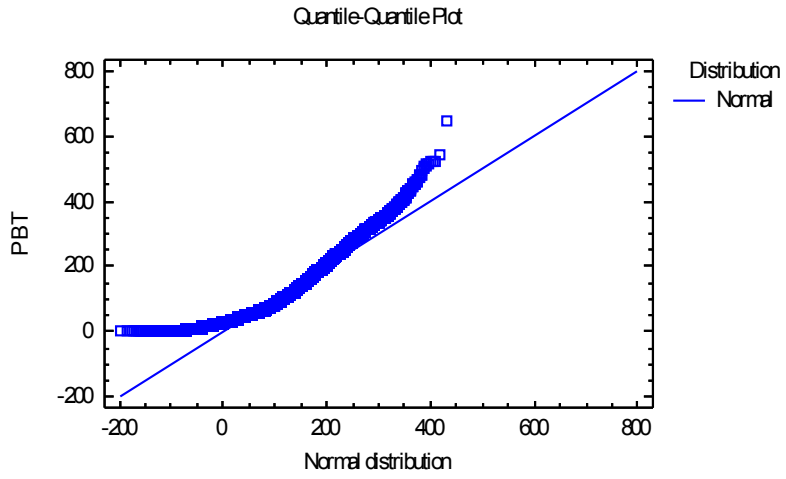


Gráfico 5. Q-Q en las colas de la distribución en la votación para la CPBT.



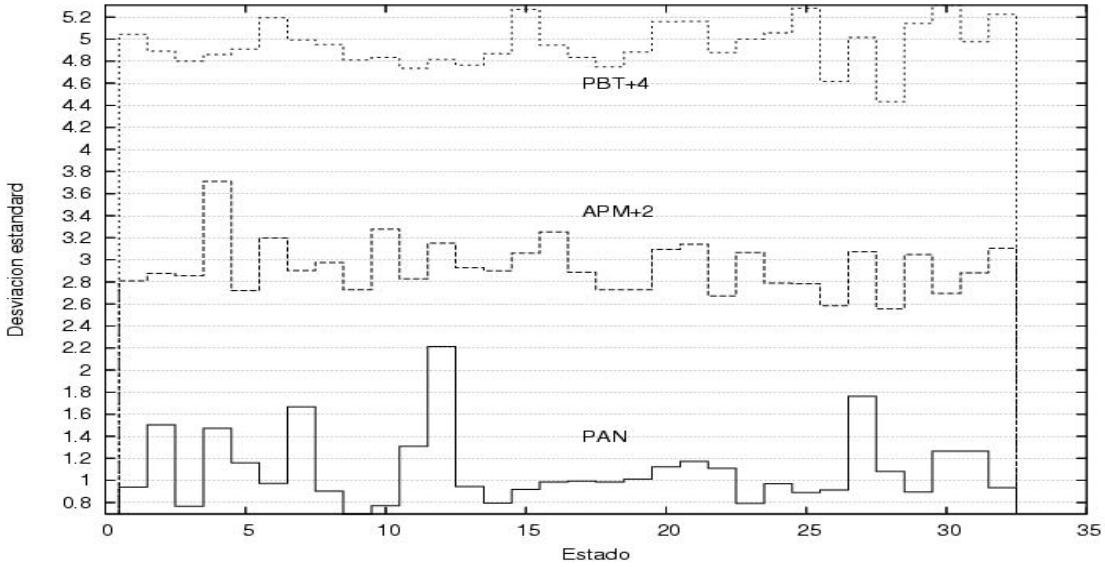
Para estar en condiciones de replicar el desglose por zona geográfica Norte-Sur, tal como lo hace Luis Mochán, se requiere contar con información sobre qué estados entran en cada subconjunto. No hay referencia alguna a esto en la página citada.

Pregunta (vi) de Luis Mochán: *Todavía no se si mi estadística de los dígitos menos significativos (figura 60) sea correcta. No arroja un resultado tan tajante como el que obtuvo el Dr. Barberán en su análisis de la elección del 88, pero muestra aspectos que vuelven a esta elección muy improbable, i.e., en menos de tres de cada 100,000 elecciones podría uno esperar una dispersión tan grande como la que corresponde al PAN.*

*La figura 60 muestra para cada estado de la la dispersión obtenida en la estadística del dígito menos significativo, i.e., el que va hasta el extremo derecho. Como se discutió al presentar las figuras 24-26, se espera que cada dígito sea equiprobable y que aparezca alrededor de 0.1 N veces en un estado, con fluctuaciones caracterisadas por una desviación estandard dada por la raiz cuadrada de 0.1*0.9*N, donde N es el número de actas contabilizadas. Calculé la varianza empleando para ello los diez dígitos 0,1,...9 en cada uno de los 32 estados, los cuales aparecen numerados en orden alfabético en el eje horizontal de la figura. En el eje vertical puse el valor de la desviación estandard de la muestra normalizada al valor de la desviación estandard esperada. Para que la gráfica no quedara amontonada, desplacé los resultados corresondientes a APM y a PBT una distancia de 2 y 4 respectivamente en la dirección vertical. De manera análoga a lo observado con los datos del conteo, observamos que la varianza toma valores más grandes en general para el PAN que para la Alianza por México, para la cual es más grande aún que para la Coalición por el Bien de Todos. Arriegándome a un primer ejercicio*

de principiante en cuantificación estadística, evalué crudamente la probabilidad de estos resultados empleando la distribución chi cuadrada con 9 grados de libertad (10 dígitos - 1) (Esto no es estrictamente correcto pues los resultados para los 10 dígitos no son estríctamente independientes). En Guerrero el PAN muestra una desviación estandard mayor a 2.2. La probabilidad de que esto hubiese ocurrido en un estado dado de acuerdo a la distribución chi cuadrada es menor a una parte en 100,000. La probabilidad de que hubiese sucedido en alguno de los 32 estados es entonces menor a 3 partes en 10,000. Del mismo modo, la probabilidad de obtener una variancia mayor a 1.6 es menor a 6 partes en 1000. Sin embargo, para el PAN dicha variancia se excede en tres estados. La probabilidad de dicho evento es menor a una parte en mil. Se puede concluir entonces que la probabilidad de una distribución de dígitos como la mostrada en la figura 60 ¡es sumamente improbable! *(Nota: Fernando Rodriguez ha hecho una crítica a la hipótesis de equiprobabilidad.)*

Figura 60



Lista de estados: 1 Aguascalientes, 2 Baja California, 3 Baja California Sur, 4 Campeche, 5 Coahuila, 6 Colima, 7 Chiapas, 8 Chihuahua, 9 Distrito Federal, 10 Durango, 11 Guanajuato, 12 Guerrero, 13 Hidalgo, 14 Jalisco, 15 Mexico, 16 Michoacán, 17 Morelos, 18 Nayarit, 19 Nuevo León, 20 Oaxaca, 21 Puebla, 22 Querétaro, 23 Quintana Roo, 24 San Luis, 25 Sinaloa, 26 Sonora, 27 Tabasco, 28 Tamaulipas, 29 Tlaxcala, 30 Veracruz, 31 Yucatán, 32 Zacatecas.

Respuesta (vi):

Al pensar en un análisis del dígito que se encuentra en la posición de las unidades en los registros de la cantidad de votos por partido o coalición, quizás una aproximación inicial que resultaría adecuada y lo suficientemente justa sería el echar un vistazo a los rangos de variación de los montos de votos por cada partido o coalición. Si el rango para cada una de las entidades se mantuviera relativamente constante, e incluso, si los rangos para cada candidato fuesen semejantes, quizás el supuesto de equiprobabilidad cobraría fuerza y validez en el agregado nacional, pero no a nivel entidad federativa, ya que el rango observado por entidad para cada partido o coalición es de carácter variado. Es decir, si un determinado partido presenta un rango observado de votos relativamente estrecho, parecería que los valores más frecuentes para ese rango aparecieran con mayor frecuencia que el resto. Por otro lado, si un determinado candidato presenta un rango mayor, parecería que ciertos dígitos (los de las unidades) presentarían casi la misma probabilidad. Así pues, los casos en los que los rangos de variación son menores, la varianza aparecería mayor que en los casos de mayor amplitud de rango, casos en los que la varianza parecería aminorar.

Otra alternativa que se entiende justa, es probar el ajuste a una determinada distribución con base en las frecuencias del último dígito, lo que daría una primera noción de qué esperar a nivel probabilístico en la aparición de los dígitos. Esto es, si el rango para determinado candidato termina en un “x” dígito, la probabilidad de ocurrencia de los dígitos subsiguientes al

dígito “x” se verían invariablemente reducidas, y el enfoque frecuentista igualmente podría arrojar nociones de las probabilidades de incidencia de los dígitos. En este sentido, quizás resultara más concreto el pensar al nivel de una distribución de carácter multinomial, ya que sería posible pensar el ejercicio como que cada cantidad de votos registrada (y fijándonos en el último dígito) es un ensayo el cual puede resultar en una de una cantidad fija y finita de k posibles resultados (en este caso, k=10), con probabilidades p_1, …, p_k, en N ensayos independientes. Así pues, se podría pensar entonces en una variable aleatoria Xi que indicase el número de veces que el número i se observara en los N ensayos (donde N bien podría ser la cantidad de actas contabilizadas en cada entidad federativa). Luego entonces, la distribución multinomial podría definirse como la distribución del vector (X_1,… X_k). Cada una de las k componentes podría verse como un ensayo binomial (si ocurre o no ocurre el dígito k). Bajo esta premisa, se podría calcular la correlación entre las variables aleatorias X_i y X_j y colocarlas a modo matricial para observar que las entradas que se encuentran fuera de la diagonal se encuentran negativamente correlacionadas, porque para una cantidad fija N (el número de actas), un incremento en una componente de un vector multinomial, necesariamente requiere un decremento en otra de sus componentes, por lo que quizás el enfoque equiprobablilístico no sea del todo acertado.

Pregunta (vii) de Luis Mochán: *Mi análisis aplicando la ley de Benford (fig. 27) es errónea. Sin embargo, un análisis del Dr. Walter R. Mebane de Cornell empleando la ley de Benford para el segundo dígito (presentado en el 2006 Summer Meeting of the Political Methodology Society en U. Calif. (Davis) el pasado 20 de julio) sugiere que nuestra elección es 'dudosa' y sugiere pruebas adicionales (que requirían recuentos parciales) para confirmarlo, y de ser así, recomienda un recuento total.*

De la página de Luis Mochán, texto y gráficos referidos en la pregunta.

a. *(vii) Existe otra prueba estadística sobre la probabilidad de aparición de dígitos en colecciones de números. Esta es la prueba de Benford. Yo no sabía de ella hasta hoy (11/vii/06) en que leí el artículo que escribió al respecto R. Mansilla. Resulta que desde 1881 se conoce la ley de probabilidad, conocida ahora como Ley de Benford, que describe el histograma de aparición del dígito más significativo de una colección de números aleatorios. Está demostrado que esta distribución se debe cumplir en una gran variedad de bases de datos donde hay algún elemento de azar tan diversas como áreas de ríos, pesos atómicos de los elementos químicos, números de las casa en una ciudad, etc. La aplicación actual más importante de la ley de Benford es la detección de fraudes fiscales.*

   ¿Qué es la ley de Benford (LB)? El dígito más significativo de una colección grande de números se distribuye de la siguiente manera: la probabilidad de hallar el digito D es log(1+1/d)/log(10). Por ejemplo, el dígito D=1 debería aparecer en la primera posición con una probabilidad de log(2)/log(10)=0.301, i.e., aproximádamente el 30% de las veces, mientras que el dígito D=6 debería aparecer con la probabilidad log(1+1/6)/log(10)=0.067, i.e., abajo de 7% de las veces. En la figura 27 muestro la probabilidad de obtener cada uno de los dígitos 1..9 en la posición más significativa, expresada como un porcentaje. Como referencia, marqué también el valor predicho por la LB (línea continua). Curiosamente ¡ninguno de los resultados del PREP es consistente con la LB.

   1. *Los datos de Calderón (+) parten de 45% en lugar de 30% y bajan rápidamente mostrando un mínimo para el dígito 4, subiendo posteriormente hasta aproximarse a la ley de Benford para digitos mayores.*
   2. *Los datos de Madrazo (X) empiezan por debajo de la ley de Benford, tienen un mínimo en 2 y un máximo en 5, y sólo se aproximan a la ley de Benford en 9.*
   3. *Los datos de AMLO (asteriscos) empiezan arriba de la ley de Benford, tienen un mínimo en 3 y siguen la ley de Benford aproximadamente a partir del 5-6.*
   4. *Los datos de Campa empiezan poco abajo de la LB y terminan un poco arriba. Decaen de manera monótona. Sin embargo su decaimiento inicial es muy lento comparado con el predicho por la LB.*
   5. *El comportamiento de Patricia Mercado sigue muy de cerca al de Calderón.*

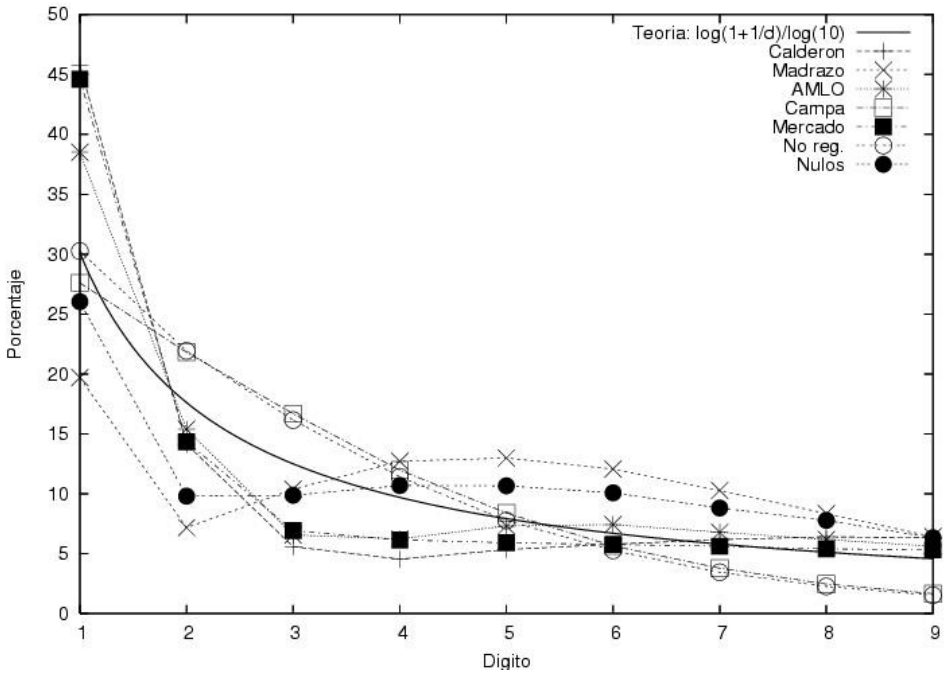
6. *Los no registrados empiezan sobre la LB pero siguen muy de cerca los resultados de Campa.*
7. *Los votos nulos siguen cualitativamente el comportamiento de AMLO, aunque con variaciones más pequeñas.*

¿Será posible que las violaciones a la LB se deban a que los números de nuestra muestra son muy chicos, todos ellos de 3 o menos dígitos? ¿Habrá efectos de tamaño finito? De ser esta la explicación de las discrepancias, yo esperaría que candidatos con números totales de votos similares siguieran curvas similares. Este no es el caso. Los datos de AMLO y los de Calderón difieren notablemente, a pesar de haber obtenido votaciones muy cercanas. Los datos de Calderón y de Mercado se parecen, a pesar de haber obtenido votacioes muy distintas.

De manera que ningún candidato cumple con la ley de Benford. Sin embargo, si vuelvo a hacer el cálculo sin distinguir los datos correspondientes a un candidato de los de los otros candidatos, es decir, si hago el histograma correspondiente a todos los votos recibidos por todos los candidatos en todas las casillas, incluyendo candidatos no registrados y votos nulos, ¡el resultado se vuelve consistente con la ley de Benford! (figura 28) Esta casualidad... parece milagrosa, aunque... ¡hay otra explicación! (sugerida por Hernán Larralde) Es posible que la ley de Benford no se aplique a nuestras distribuciones, las cuales no son invariantes de escala. Como las distribuciones tienen un máximo (por ejemplo, 53 en el caso de Madrazo), es factible que el dígito más significativo del mismo (5 en el caso de Madrazo) aparezca con una frecuencia mayor que el dígito anterior o que el posterior (4 o 6 para Madrazo). Al agregar todos los datos en un mismo histograma, sumamos candidatos con distintos números esperados de votos y creamos una distribución más parecida a una distribución invariante de escala, con lo cual mejoramos el ajuste a la ley de Benford.

Ayer (26/VII/06) me enteré de que, aunque no tenemos por qué esperar que la ley de Benford mencionada arriba se cumpla para el dígito más significativo de los datos de las elecciones, hay una generalización de la ley de Benford que incluye a los dígitos subsiguientes (segundo, tercero,...) y que su violación es una indicación seria de anomalías y posibles fraudes. En una nota electrónica, Luis Horacio Gutierrez me ha facilitado un artículo sobre la teoría matemática que sustenta a la Ley de Benford y a su amplia aplicabilidad y otro artículo escrito por el profesor W. R. Mebane de la U. de Cornell en que aplica dicha ley para estudiar sistemáticamente los resultados de nuestra reciente elección y la elección de Florida en 2004. Finalmente, aquí hay una presentación con un estudio detallado de nuestra eleeción empleando la ley de Benford.

Figura 27



Respuesta (vii):

La Ley de Benford y el PREP

La Ley de Benford (BL) indica una mayor incidencia de números pequeños con relación a números mayores como primer dígito[6]. El sustento empírico de la BL son bases de datos de toda índole (direcciones de calles, inundaciones, bacterias, etc).

En algunos documentos se hecho referencia a una supuesta suspicacia en la base de datos del PREP, misma que radica en que el número de votos, tanto para el candidato de la CBT, como para el candidato del Partido Acción Nacional (PAN) no se ajustan al modelo teórico. En otras palabras, que el número de votos asentados en un acta, a favor de cada uno de estos dos contendientes, comienza mayormente con un 1, luego con menor frecuencia con un 2, y así sucesivamente. Según las gráficas que se muestran en la página de Luis Mochán, es cierto que así sucede. Efectivamente hay mayor incidencia de actas donde el primer dígito de los votos a favor de CBT y de PAN es 1 ó 2. La inconformidad relativa a estos resultados radica en que a pesar de que así lo indica la BL, esto no se ajusta a los valores del modelo teórico.

Por qué la suma de votación a nivel casilla no se ajusta a la Ley de Benford aplicada al primer dígito (1BL)

El número máximo de boletas que tenía cada casilla era de 750 más 10, es decir, por diseño se disminuye la probabilidad de que el primer dígito de cada cantidad de votos sea 7, 8 ó 9. El universo de números para el primer dígito no incluye el 7, 8 ó 9 como se incluye en cualquier otro tipo de experimento. En el caso del número máximo de boletas por casilla la escala va del 1 al 9 para las unidades y para el primer bloque de decenas, pero del 1 al 7 para las centenas. Por lo tanto, esto representa una primera dificultad para que las sumas de votos por partido o coalición se ajusten a la 1BL bajo cualquier circunstancia, pues se parte de un supuesto erróneo (distribución específica de números para el primer dígito).

[6] También se puede aplicar al segundo, tercer y demás dígitos, haciendo las modificaciones pertinentes en la fórmula.

Jugando con la idea anterior, podemos hacer el siguiente ejercicio: si asumimos como dada la tasa de participación que el PREP registró (58.9%), el número promedio de votos por casilla para la elección presidencial es 448. Eso reduce el universo de primeros dígitos posibles para los votos en cada caso, hace más probable la incidencia de que los números del 1 al 3 aparezcan que los que van del 4 al 9.

Siguiendo con el ejercicio, si consideramos que la competencia se daba a nivel casilla con ciento setenta y tres votos para el candidato de la CBT en las casillas donde ganaba dicho candidato, y algo muy similar sucedía para el candidato del PAN en las casillas donde ese candidato ganaba, y en ambos casos el candidato contrincante en segundo lugar perdía con al rededor de 74 votos, entonces es correcto esperar una mayor incidencia de números menores en el primer dígito de la suma de votos por casilla, específicamente del 1.

Finalmente, si extrapolamos el resultado final de la contienda a la cantidad de votos que en promedio se emitió, estamos hablando de que ciento cincuenta y tantos votos son para cada uno de los candidatos punteros, por lo que la mayor incidencia del dígito 1 es correcta – incluso rebasando la predicción del modelo teórico benfordiano. Y evidentemente esto no representa anormalidad alguna.

Además, si se observan las curvas que se muestran en la página de Internet de Luis Mochán, los valores entre 3 y 6 son menos frecuentes en la suma de votos para todos los partidos y coaliciones (salvo en el caso de Nueva Alianza) de lo que predice la BL. Lo anterior es de esperarse, pues sucedió poco que la suma de votos por partido o coalición (a nivel casilla) estuviera entre 300 y 699, o entre 30 y 69.

La recomendación del profesor Mebane consiste en la realización de estudios exhaustivos antes de caer en la tentación de hacer un recuento de los votos

El profesor Walter R. Mebane, Jr. escribió un documento llamado Election Forensics: Vote Counts and Benford's Law, en el que sostiene que la aplicación de la Ley de Benford al segundo dígito del conteo de votos por partido o coalición puede ser conveniente en ocasiones determinadas. El estudio del profesor Mebane se puede aplicar a cualquier sistema electoral, en tanto la única finalidad de aplicar la Ley de Benford al segundo dígito de la suma de votos (el autor abrevia el término con la notación *2BL*) es detectar comportamientos que diverjan fuertemente de la distribución probabilística que tiene dicha ley. Sin embargo, a lo largo de todo su estudio, el profesor afirma que dicha divergencia entre resultados empíricos y teóricos no es un indicador robusto, debido a que la distribución de datos tal como los registra el sistema de conteo local puede tomar cualquier tipo de distribución[7]. Específicamente en el caso de México, el profesor recomienda que antes de sacar conclusiones sobre la necesidad de un recuento de votos, se lleven a cabo estudios exhaustivos con base en la 2BL a nivel distrito, así como con muestreos estratificados de los conteos parciales. Lo anterior con la finalidad de tener una mejor visión del panorama electoral antes de incurrir en costos de recuentos mayores de votos.

Las dos principales razones para recomendar estudios que incluyan distintos niveles de agregación del conteo de votos son la operatividad de las elecciones en México, y los criterios de asignación de votantes a casillas. La operatividad del conteo de votos mexicano se basa en los criterios poblacionales de la distritación. Sin embargo, el profesor Mebane, se limita a hacer cálculos a nivel casilla y a nivel sección. Con ello se altera la proporcionalidad del conteo, y no hay distribuciones de votación comparables.

En cuanto a la asignación de votantes a las casillas, el profesor señala que el problema para la aplicación de la 2BL es que la lista nominal en un buen número de casos es mucho menor que el promedio. Esto ocasiona que no exista un buen ajuste de la distribución a patrones de Ji cuadrada o de normalidad. Por lo tanto, es necesario observar niveles más agregados de conteo de votos.

[7] " Even if 2BL typically describes vote count data, it does not follow that deviations form 2BL indicate election fraud." Página 2, cuarto párrafo. "Because significant perturbations (on the vote count distribution) may occur in the absence of fraud, such a result can do no more than suggest the possibility of fraud" Página 14, segundo párrafo.

Por ende, tras una minuciosa lectura del estudio del profesor Mebane, es posible concluir que el autor no recomienda un recuento de los votos, sino la realización de una serie de estudios que eviten sacar conclusiones precipitadas y onerosas.

Por qué la votación emitida por todos los partidos o coaliciones puede tener un mejor ajuste a la Ley de Benford aplicada al primer dígito (1BL)

Las elecciones son fenómenos cuyos resultados están condicionados al comportamiento interactivo de los votantes y los contendientes. La decisión de votar por un candidato, partido o coalición determinado no es un evento aleatorio, y en una gran mayoría de casos responde, principalmente, a dos estímulos combinados: la afinidad y la animadversión hacia algún contendiente, aun que también es cierto que algunas veces la decisión de votar en una dirección se toma unos instantes antes de emitir el sufragio[8].

Debido a que el resultado de la elección afecta tanto a los que votaron a favor de un candidato, como a los que no votaron por él o ella, las preferencias electorales de distintos grupos no son independientes entre sí. Es muy probable, especialmente en elecciones competidas, que exista algún tipo de relación entre la votación a favor de alguien y la falta de votación a favor de los oponentes. Por lo tanto, al observar la votación a favor de un candidato determinado no se obtienen resultados de una muestra representativa del comportamiento electoral de la población de una región. Muy por el contrario, al estar correlacionada entre sí, la votación a favor de los diferentes candidatos, al observar únicamente un subconjunto de la población, cualquier estimación estadística va a contener necesariamente un sesgo. Con base en esto, se concluye que al incluir a todas las sumas de votos (por cada uno de los partidos o coaliciones, así como por los candidatos no registrados y los votos nulos) en cierto nivel de agregación poblacional (que no sea la casilla), se obtengan resultados que se ajusten a la 1BL.

Consideraciones finales

En la práctica se cuenta con un número reducido de elecciones documentadas, por lo tanto, es difícil identificar una distribución sobre las cuales hacer un modelo ajustado, que cumpla con reglas similares a la BL en cualquiera de sus etapas y acepciones.

Además, no es trivial recalcar que ninguna elección es igual a otra. Dado que una elección es un suceso poco frecuente, es extremadamente complicado equiparar dos o más coyunturas políticas e imponer reglas estadísticas rígidas para determinar si hay o no manipulación de la información.

[8] El profesor Mebane hace referencia a esto en la página 8 de su documento.