CFATA-UNAM
Campus Juriquilla

19 de agosto de 2006

Seminario2006

¡Hola!

Tiene razón Luis al decir que el esquema que he trabajado corresponde a una binomial y también que la desviación estándard de una casilla con una lista de 750 nombres es menor que 2 %. De hecho, en tales condiciones, tal desviación está acotada por 1.83 %.

El principal error del esquema que yo he propuesto es precisamente utilizar los datos del IFE, pero eso no se puede remediar, ya que no hay otros. Sí podemos hacernos una idea de la confiabilidad de los números publicados por el IFE. Todos los que hemos seguido el proceso electoral fuimos testigos de la confrontación de dos cifras, hechas públicas poco antes del 2 de julio. Una era el padrón del IFE, con poco más de 71 millones, y la otra, el censo levantado por el INEGI, que el año pasado reportaba del orden de 61 millones de mayores de 18 años, misma que *proyectada* de acuerdo con las tazas de natalidad y mortandad, permitía pensar en una cifra cercana a los 63 millones. Estoy hablando de memoria: la última cifra pueden ser 62 millones aunque sí recuerdo que se afirmaba una diferencia del orden del 16 %. . A pesar de que sí leí esos detalles, nunca supe la contestación del IFE, si es que la dió.

Llamemos p la probabilidad de que alguien vote –sí, en el esquema binomial. En tales condiciones el número esperado de votantes en una casilla cuya lista nominal tiene n nombres es pn. ¿Esto es correcto?

La afirmación anterior es incorrecta, ya que se habla[1] de dos procesos realizados sobre los nombres de una casilla, uno de quitar, otro de añadir. Sea N el número de nombres en la *verdadera* lista nominal. Le añadimos a y le quitamos r, correspondiendo este último al proceso de *rasurado*. Podemos entonces afirmar que el número n antes introducido satisface:

$$n = N - r + a. \tag{1}$$

[1] entre los descontentos con la limpieza de los resultados reportados

De los ciudadanos normales espero una participación $(N-r)p$, ya que en esa *lista* sólo hay $N-r$ ciudadanos *normales* y al mismo tiempo una desviación estándard dada por $\sqrt{(N-r)p(1-p)}$ misma que se mide respecto de $(N-r)p$. Debemos, sin embargo, tomar en cuenta los *añnadidos*. Si todos estuvieran muertos, me espero que no haya ninguna contribución. Sin embargo, la participación ya no es p, como calcularíamos en condiciones normales, sino

$$\frac{(N-r)p}{N-r+a} = \frac{(N-r)p+ap-ap}{N-r+a} = p - \frac{ap}{N-r+a}. \tag{2}$$

Estos *añadidos* son inocuos ya que no votan. Si, por el contrario, estos añadidos tuvieran un papel más activo, el resultado puede cambiar. Por ejemplo, si estos añnadidos tuvieran una probabilidad de participar igual a uno, entonces la participación resulta tomar el valor

$$\frac{(N-r)p+a}{N-r+a} = \frac{(N-r)p+a[p+(1-p)]}{N-r+a} = p + \frac{a(1-p)}{N-r+a} \tag{3}$$

Podemos ver entonces que la participación oscila entre los valores dados por las ecuaciones 2 y 3, es decir, dentro de un intervalo cuya longitud esta dada por su diferencia,

$$\frac{a(1-p)}{N-r+a} + \frac{ap}{N-r+a} = \frac{a}{N-r+a} = \frac{a}{n}. \tag{4}$$

El *centro* respecto del que se debe calcular la desviación estándard tiene una *incertidumbre* del orden de a/n, y desplazar el punto respecto del cual se calcula el segundo momento da lugar a un cambio en la desviación estándar del orden de $\sqrt{a/n}$. Reconozcamos que quien sostenga que los datos del IFE son confiables sólo aceptará $a = 0$.

No conocemos a, pero la suma de tales *añadiduras* puede ser la explicación entre las dos cifras señaladas más arriba, colocándose de esta manera en un $16\,\%$.[2]

Si adoptamos $a/n \approx 0.16$, la desviación estándard cambia en $\sqrt{a/n} \approx 0.4$ (esto nada más es aproximado, ya que la dispersión está dada por la suma de dos cuadrados, es decir, que las desviaciones estándard corresponden con la *hipotenusa* y los *catetos* cuyos lados cuyas longitudes son las desviaciones estándard correspondientes).

[2]Es muy probable que esta cifra sea exagerada y sí me gustaría contar con una más precisa.

Por otro lado, la desviación estándard de la columna 13, correspondiente al número total de votos, calculada sobre las 130,777 casillas, es 102.06. Este número debe ser dividido por la *lista nominal*, cuyo promedio nacional de la lista nominal es 545.73, lo que da lugar a $102/546 \approx 0.19$. Este último resultado es la mitad del estimado más arriba, lo que sugiere que el número debe ser un poco más pequeño que $a = 0.16/4 = 0.04$.

Conviene completar el esquema pasando a las estimaciones numéricas: Se cuenta el número total de votos en la casilla, se encuentra n_t pero se escribe n_f. Estos números se espera que sean iguales, pero hay evidencia de que en muchas casillas no coinciden. Quizás convenga detallar: Se encuentra n_t, se escribe n_t, sin embargo *alguien* decide que habrá de ser n_f. Más tarde se rellenará el paquete, quedando finalmente en n_f, tanto en el paquete, como en el acta, como en el IFE, como en el archivo que yo bajo de su página. Y, con ese número n_f yo calculo la participación, mediante el cociente doblemente dudoso: n_f/n, mismo que termino identificando con el número p propuesto más arriba, ya que carezco de información sobre a y r.

Aprovecho para señalar que decidí hacer un análisis sobre el total de votos, ya que la localización de la casilla efectivamente es un factor determinante respecto del candidato: hay zonas que favorecen al PRI, otras que al CBT y otras al PAN, como confirma Luis:

> ... pues existen familias, grupos de vecinos, amigos, clubes, iglesias, etc., i.e., cada individuo forma parte de una red y el voto de uno esté fuertemente correlacionado con el voto de otros miembros de su red.

Y es precisamente este hecho lo que lo hace inaplicable al caso del PRI, que tiene zonas geográficas con muy diferentes niveles de apoyo, y estas variaciones se deben añadir a las que hemos propuesto más arriba, para producir la desviación observada. La *participación* tiene la ventaja de ser un poco más uniforme.

El uso de binomiales para describir problemas de votación puede verse en el libro 'Introduction to Statistical Thought and Practice' de George Hilton, capítulo IX. El uso de binomiales en las compañias que se dedican a hacer encuestas es típico. Cito la siguiente referencia
`www.realpoor.com/presidential_poll_statistics_t30909.html`,

> Some of you must be taking statistics. What's up with the way they quote errors on presidential polls? They seem wrong to me, but all the polls seem to do it consistently.
>
> Ignoring 3'rd party candidates and undecideds (both of which are small), a poll for Bush or Kerry is a pure binomial distribution. If the poll size is N, the probability of a vote for Bush is 'p' and Kerry is (1-p). The standard deviation, by definition for a binomial is sqrt(N*p*(1-p)). Since p and (1-p) are both basically 1/2 (electorate is evenly split), that makes the standard deviation sqrt(N*.5*.5), or sqrt(N)/2. As a fraction, that means the S.D. is sqrt(N)/2N, or 1/(2*sqrt(N)).
>
> But all the polls report it as if the standard deviation is 1/sqrt(N). For example, a poll of 400 voters is reported to be plus or minus 5 %, when I'd say it should be 1/(2*sqrt(400)), or 2.5 %. 717 voters is reported to be ± 3.5 %. Both those numbers seem to be twice what they should be. Does anyone understand why?
>
> Some examples at
> `http1://www.cnn.com/ELECTION/2004/special/president\`
> `/showdown/FL/polls.html`

De acuerdo con lo anterior, *allá* parece que utilizan el número $g = 2$ en lugar del $g = 23$ (quede claro que no dudo que Luis haya ajustado correctamente el número g).

Las dos situaciones confrontadas en la cita tienen una explicación: $1/(2 * \sqrt{(N)})$ es la desviación estándard de las observaciones de una binomial con $p = 1/2$, mientras que $1/\sqrt{(N)}$ es la anchura del intervalo de confianza ofrecido por la compañia encuestadora que *contiene* con la **confiabilidad anunciada por la compañia** el parámetro p (desconocido) que se desea estimar (con una binomial).

Usando Google con 'binomial vote' aparecen 213,000 resultados, algunos interesantes, la mayoría basura.

Aprovecho para mandar un saludo al colega de Luis, a Max Aldana, de quien

estoy convencido que hará un excelente trabajo.

Atentamente

Miguel de Icaza Herrera