La estadística detrás de los conteos rápidos de 2018

Manuel Mendoza Ramírez
Profesor del Departamento Académico de Estadística del ITAM

Gerardo Orantes Jordan
Exalumno de Actuaría y Matemáticas Aplicadas del ITAM

Gian Carlo Diluvi
Exalumno de Matemáticas Aplicadas del ITAM

2018 fue testigo del proceso electoral más grande de la historia de México: más de 3,400 cargos a nivel tanto federal como local fueron decididos en las urnas por los más de 55 millones de mexicanas y mexicanos que salieron a votar el domingo 1° de julio.

El organismo encargado de la organización de las elecciones en México es el Instituto Nacional Electoral (INE)[1] y sus funciones incluyen, entre otras, supervisar las campañas de los partidos políticos, la planeación de la logística para el día de la jornada, la selección y capacitación de los ciudadanos que fungen como funcionarios de casilla, y el conteo de los votos y anuncio de los resultados.[2]

Debido a la relevancia política y social de la contienda electoral, es de interés conocer el resultado de las elecciones el mismo día en que se llevan a cabo. Sin embargo, los resultados oficiales solo se producen al término de los *cómputos distritales*, habitualmente una semana después de la jornada electoral. En esa circunstancia es necesario recurrir a procedimientos estadísticos de estimación. Para este propósito el ejercicio que otorga mayor certeza es un conteo rápido, que utiliza una muestra de casillas y requiere que los funcionarios encargados de ellas reporten los votos cuando terminen de contarlos, para estimar así los resultados de la votación en la totalidad de casillas.

Para el proceso electoral de este año el INE organizó 10 conteos rápidos diferentes: uno para la elección presidencial, uno para la elección de la jefatura de gobierno de la Ciudad de México, y uno para cada una de las demás entidades en las que se celebraron elecciones de gubernatura. Para llevar a cabo esta tarea el INE designó un comité técnico a cargo de la organización e implementación de tales conteos.

En este artículo se discute uno de los modelos empleados en este sistema de conteos rápidos, particularmente para la elección presidencial y las gubernaturas de Chiapas y Guanajuato. Primero se ahonda un poco más en la organización de los conteos rápidos y en el diseño muestral; luego se plantea el modelo original y se mencionan un par de modificaciones realizadas para esta jornada electoral; finalmente se discuten los resultados para la elección presidencial.

[1] El INE es responsable único solo de la organización de las elecciones federales; las elecciones locales las organiza en conjunto con los organismos electorales locales.

[2] Aunque el anuncio legal de los ganadores de la contienda lo realiza el Tribunal Electoral unos meses después de la jornada.

Los conteos rápidos del INE

Si bien el INE ha organizado conteos rápidos para distintas elecciones y en varias ocasiones, nunca se habían implementado tantos en una misma jornada electoral. Por ello el comité formado por el INE estuvo conformado por nueve expertos en Estadística y Matemáticas. Para distribuir la carga de trabajo el comité decidió seleccionar un responsable por cada elección local (jefatura de gobierno y gubernaturas), y crear tres equipos (A, B, y C) para la estimación federal.

Para la elección presidencial cada uno de los tres equipos era responsable de proponer una estimación, y al final éstas se consolidaron para así obtener la estimación final. En el caso de las elecciones locales, cada equipo determinó producir dos o tres estimaciones por entidad para luego consolidarlas. De esta forma cada miembro del comité fue responsable de producir una estimación para la elección presidencial, una para la entidad local que le correspondiera, y una o dos adicionales para las entidades de los miembros de su equipo.[3]

En el caso del equipo A, cada uno de sus integrantes utilizó un modelo Bayesiano diferente para sus estimaciones tanto locales como federales. En este reporte se discute uno de ellos.

Diseño muestral

Una de las funciones del comité es determinar el diseño muestral de los conteos rápidos. En esta ocasión el diseño para la elección presidencial se decidió en conjunto por todo el comité, mientras que cada miembro fue responsable de determinar el diseño para la entidad a su cargo. En todos los casos la unidad observacional fue la casilla, de las cuales hubieron 156,899 en el marco muestral federal. También en todos los casos el diseño fue estratificado.

En el caso federal, la estratificación estuvo definida por los distritos electorales federales en aquellas entidades sin elección local de gubernatura, y por la estratificación local establecida por cada miembro responsable en el caso de aquellas entidades que sí contaran con elección local. Dicha estratificación dio lugar a 350 estratos a lo largo de todo el país.

El comité determinó que el número de casillas en muestra necesario para obtener precisiones de 0.25 en la estimación de los porcentajes de votación, con un 95 % de confianza, era de 7,500. Tomando en cuenta que dos entidades tienen husos horarios con dos horas de diferencia, lo cual impacta la velocidad de captura y envío de la información, se decidió incluir 287 casillas de sobremuestra, para un tamaño final de muestra de 7,787 casillas (casi el 5 % del total).

[3] El número de estimaciones adicionales, así como el método para consolidarlas, se determinó por equipo.

El modelo

El modelo que se planteará está basado en el desarrollado por Mendoza y Nieto (2016), mismo que los autores emplearon en los conteos rápidos de las elecciones presidenciales de 2006 y 2012. Primero se presentará brevemente el modelo original y después se discutirán dos modificaciones que se incorporaron para los conteos rápidos de esta jornada. Se recomienda ampliamente revisar el artículo original y la bibliografía ahí citada si se desea ahondar más en el tema, particularmente en el desarrollo de los conteos rápidos de aquellas elecciones.

Comenzaremos introduciendo la misma notación empleada por Mendoza y Nieto. Como se ha indicado el diseño muestral es estratificado y la unidad observacional es la casilla. Hay que mencionar que en cada casilla pueden votar, como máximo, 750 personas. Además cada casilla cuenta con una lista de las personas que están registradas para votar ahí; se llama *lista nominal* y, en promedio, está conformada por alrededor de 550 ciudadanos.

Utilizaremos la letra i para indexar estratos, la letra j para indexar candidatos, y la letra k para indexar casillas dentro de cada estrato. Supongamos que la población se divide en N estratos, y que el estrato i cuenta con K_i casillas. Sea n_i^k el tamaño del listado nominal de la casilla k del estrato i, para $k = 1, ..., K_i$ e $i = 1, ..., N$. Definimos X_{ij}^k como el número de votos en la casilla k del estrato i en favor del candidato o categoría j, $j = 1, ..., J$. En general habrá $J' = J - 2$ candidatos que contienden en la elección y, además, dos categorías adicionales: la categoría de *Otros*, que se refiere a aquellos votos que fueron anulados o que se emitieron por candidatos no registrados; y la categoría de los votos que no fueron emitidos (personas que no acudieron a votar). En las elecciones presidenciales de este año $J = 6$, puesto que había $J' = 4$ candidatos contendientes (Ricardo Anaya Cortés, José Antonio Meade Kuribreña, Andrés Manuel López Obrador y Jaime Heliodoro Rodríguez Calderón). En este reporte se utilizará el término *candidato* para referirse tanto a los J' candidatos contendientes como a las dos categorías adicionales, a menos que se especifique lo contrario.

Con la información de las casillas en la muestra, el modelo produce estimaciones por estrato. Por ello se definen, para cada estrato i y candidato j,

$$n_i = \sum_{k=1}^{K_i} n_i^k, \quad X_{ij} = \sum_{k=1}^{K_i} X_{ij}^k$$

como el tamaño de la lista nominal del estrato i y el número de votos en el estrato i para el candidato j, respectivamente.

Ahora bien, lo que se desea estimar no es el número total de votos sino la *proporción* de votos en favor de cada candidato. En el estrato i esta proporción se define como $\theta_{ij} = X_{ij}/n_i$. Ya que se desea reportar el valor no por estrato sino de toda la población, se define la proporción de votos en favor del candidato j,

$$\theta_j = \sum_{i=1}^{N} \frac{n_i}{n} \theta_{ij}. \tag{1}$$

Aquí $n = \sum_i n_i$ es el tamaño de la lista nominal de todo el país (¡casi 90 millones de personas en estas elecciones!). Note que se están contabilizando también los votos no emitidos, por lo que en realidad θ_j mide la proporción de votos por candidato del total de la lista nominal. Sin embargo, los resultados generalmente se miden con respecto al total de votos emitidos, por lo que se define

$$\lambda_j = \frac{\theta_j}{\sum_l \theta_l}, \tag{2}$$

la proporción de votos *emitidos* en favor del candidato j. Es crucial mencionar que la suma del denominador se realiza sobre todas las θ_j *excepto la referente a los votos no emitidos* ($j = 1, 2, 3, 4, 5$ en la elección presidencial de este año, por ejemplo).

Ahora bien, en el estrato i se seleccionan aleatoriamente c_i casillas, para un tamaño de muestra total de $c = \sum_i c_i$ casillas. Entonces Mendoza y Nieto proponen, para cada estrato i, candidato j, y casilla en muestra k,

$$X_{ij}^k \,|\, \theta_{ij}, \tau_{ij} \sim \mathcal{N}\left(n_i^k \theta_{ij}, \frac{\tau_{ij}}{n_i^k}\right), \tag{3}$$

donde τ_{ij} es un parámetro de precisión independiente de θ_{ij} y desconocido y constante en cada estrato.

En cuanto a la distribución inicial, hay que considerar que su papel es reflejar adecuadamente los conocimientos iniciales sobre las proporciones de votos. Los tiempos de procesos electorales son interesantes porque en los meses y días previos a la jornada se realizan una variedad de estudios, encuestas, y sondeos con el propósito de estimar las tendencias de la votación. Por ello la cantidad de información sobre las votaciones es considerable. Si una casa comercial se diera a la tarea de implementar un modelo Bayesiano con el propósito de pronosticar el resultado de la contienda, sería sensato incorporar información de algunas de estas encuestas en la distribución inicial.

Sin embargo, el INE debe mantener una postura absolutamente imparcial. Es inconcebible que, a priori, la autoridad electoral suponga que algún candidato tiene una mayor proporción de votos que el resto de los candidatos. Por lo mismo, Mendoza y Nieto proponen una distribución mínimo informativa para (θ_{ij}, τ_{ij}):

$$p(\theta_{ij}, \tau_{ij}) \propto \tau_{ij}^{-1} \mathbb{1}(\tau_{ij} > 0)\, \mathbb{1}(0 < \theta_{ij} < 1). \tag{4}$$

Note que, a priori, θ_{ij} y τ_{ij} son independientes y θ_{ij} sigue marginalmente una distribución uniforme en (0,1).

Considerando (3) y (4) se obtiene que la distribución final para θ_{ij} y τ_{ij} está dada por

$$p(\theta_{ij}, \tau_{ij} \mid X_{ij}) \propto \mathcal{N}\left(\theta_{ij} \;\middle|\; \frac{\sum_{k=1}^{c_i} x_{ij}^k}{\sum_{k=1}^{c_i} n_i^k},\ \tau_{ij} \sum_{k=1}^{c_i} n_i^k\right) \mathbb{1}(0 < \theta_{ij} < 1) \times \text{Gamma}\left(\tau_{ij} \;\middle|\; \frac{c_i - 1}{2},\ \frac{1}{2}\left\{\sum_{k=1}^{c_i} \frac{(x_{ij}^k)^2}{n_i^k} - \frac{\left(\sum_{k=1}^{c_i} x_{ij}^k\right)^2}{\sum_{k=1}^{c_i} n_i^k}\right\}\right), \tag{5}$$

que describe una distribución Normal truncada condicional en τ_{ij} para θ_{ij}, multiplicada por una distribución Gamma para τ_{ij}. Particularmente es necesario que $c_i \geq 2$ para que esta última esté bien definida, i.e., cada estrato debe contener al menos dos casillas en muestra.

Mendoza y Nieto discuten con mayor detalle la razón de utilizar (3) como modelo de muestreo, así como algunas propiedades de la distribución inicial (4).

Implementación del modelo

Es posible hacer inferencias con base en la distribución posterior vía simulación. Simplemente es necesario, dada una muestra, generar observaciones de τ_{ij} vía la distribución Gamma de la ecuación (5) y utilizarlas para generar observaciones de θ_{ij} vía la distribución Normal truncada. Note que es necesario asegurarse de truncar la distribución Normal para que θ_{ij} esté contenido en (0,1). Para cada vector simulado $(\theta_{1j}, \theta_{2j}, ..., \theta_{Nj})$ se obtiene un valor de θ_j con la ecuación (1), para cada candidato j. Así, con los valores obtenidos de $\theta_1, ..., \theta_J$, se calcula un valor de interés λ_j vía la ecuación (2). El procedimiento se repite M veces para producir una muestra (de tamaño M) de la distribución posterior del vector $(\lambda_1, ..., \lambda_{J-1})$.

En la práctica, el personal de campo acude a cada casilla en cuestión en la muestra y, al finalizar el conteo de los votos en ella, informa al INE (generalmente vía telefónica) de los resultados de la votación. Este proceso implica que no toda la información llega al mismo tiempo, sino que lo hace conforme los votos de las diferentes casillas se van terminando de contar. Por ello, el INE no espera a contar con la información de toda la muestra, sino que cada 5 minutos elabora un compilado de las casillas cuya información ha recibido hasta el momento y se lo envía automáticamente al comité mediante una red interna.

Así pues, cuando la información llega generalmente está incompleta. Particularmente puede ocurrir que no haya información de uno o más estratos, ocasionando que no se puedan simular observaciones de (5). En esos casos el conocimiento sobre los parámetros θ_{ij} está descrito por la distribución inicial (4), y es de esta distribución de la que se simula.

Modificaciones al modelo original

Si bien el modelo discutido fue exitosamente empleado en las jornadas de 2006 y 2012, hay dos detalles que fueron revisados en esta jornada electoral. La primera y más relevante tiene

que ver con la distribución inicial.

En 2018 el INE solicitó al comité que, además de las estimaciones de las proporciones de votos de cada candidato contendiente j, realizara una estimación de la *participación ciudadana*, denominada ρ. Ésta se refiere al porcentaje de personas de la lista nominal que fueron a votar, y se calcula como el total de votos emitidos dividido por n, el tamaño de la lista nominal del país. En el contexto del modelo es sencillo estimarla, puesto que

$$\rho = \sum_{l} \theta_l \tag{6}$$

donde, al igual que en (2), la suma se hace sobre todas las θ_l exceptuando la referente a los votos no emitidos. En la práctica es posible realizar inferencias sobre la participación sumando los valores simulados de θ_j y obteniendo así una muestra posterior de ρ.

Sin embargo a priori, para cada j, $\mathbb{E}[\theta_j] = 1/2$ puesto que es un promedio ponderado de variables uniformes en (0,1). Si en un estrato no se tiene información entonces se simularán observaciones de la distribución inicial (4), y éstas tendrán un impacto en el promedio ponderado (1), ocasionando que las inferencias sobre la participación sobreestimen a ρ.

Para corregir esta anomalía, se introdujo una modificación a la distribución inicial (4) como sigue. Se sabe que la participación histórica es de 66.56 % y que la proporción de votos anulados o por candidatos no registrados histórica es de alrededor de 3 %. Así pues, esta proporción se traduce en $0.6656 \times 0.03 \approx 2\,\%$ de la lista nominal. Luego, $0.6656 - 0.02 = 64.56\,\%$ es la proporción que, con respecto al listado nominal, representan los votos emitidos por todos los candidatos contendientes. Una forma de reflejar la imparcialidad inicial frente a las proporciones de votos obtenidas por los candidatos es suponer que la participación ciudadana se distribuye a partes iguales, a priori, entre ellos. Si se recuerda que el caso presidencial cuenta con $J' = 4$ candidatos, se puede pensar en una distribución para cada θ_{ij} con media $\mathbb{E}[\theta_{ij}] = 0.6456/4$. En esta jornada se utilizó una distribución $p(\theta_{ij}) = \text{Beta}(\alpha, \beta)$, donde los hiperparámetros α y β se escogieron de forma que la distribución estuviera centrada en el valor requerido. Utilizando la expresión de la media de esta distribución se obtiene la relación

$$\beta = \alpha \left(\frac{J'}{0.6456} - 1 \right).$$

Adicionalmente, y puesto que sujeta a esta condición la varianza inicial aumenta en la medida en que α tiende a cero, este hiperparámetro se fijó en un valor cercano a cero; concretamente en $\alpha = 0.1$, de forma que la distribución inicial modificada resulta

$$p(\theta_{ij}, \tau_{ij}) \propto \tau_{ij}^{-1}\, \mathbb{1}(\tau_{ij} > 0)\, \text{Beta}(\theta_{ij} \,|\, 0.1,\, 0.1((J'/0.6456) - 1)) \tag{7}$$

para el estrato i y candidato j. Para el caso de la proporción de votos para *Otros* se empleó el mismo principio pero centrando la Beta en 0.02.

Es crucial insistir en que esta modificación *no* altera el estado de conocimientos vago en relación con la proporción inicial de votos en favor de cada candidato. Si bien se está incorporando información histórica acerca de la participación ciudadana y de los votos nulos y por candidatos no registrados, el supuesto sobre los candidatos contendientes es que, a priori, *deben* estar empatados. Por lo mismo esta distribución inicial conserva el estado de imparcialidad del INE sobre los candidatos contendientes.

El segundo cambio está relacionado con las llamadas casillas especiales, aquellas dispuestas para personas en tránsito. Éstas solo tienen 750 boletas y únicamente se puede votar para elecciones federales (dependiendo del distrito, sección, y entidad). Si bien estas casillas cuentan con 750 boletas, su lista nominal es oficialmente cero, lo que ocasiona problemas con la expresión de la distribución final (5).

Una primera solución consiste en suponer artificialmente que su lista nominal es 750 y estimar θ_{ij} de esta forma. El problema con esta solución es que ocasiona que se subestime la participación, ya que las personas que votan en casillas especiales *sí* están contempladas en la lista nominal: la de su casilla. Esta solución las cuenta dos veces, ocasionando errores en la estimación.

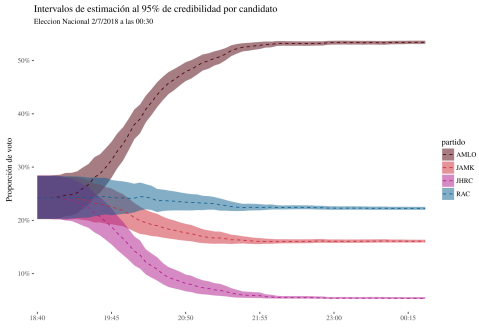
La solución que se implementó es la siguiente. Supongamos que hay s casillas especiales en el marco muestral y m casillas no especiales. En esta elección $s = 1,054$ y $m = 156,899 - 1,054 = 155,845$. Entonces a las s casillas especiales se les asignó artificialmente una lista nominal de 750, ocasionando un incremento de $750s$ elementos en la lista nominal. Para compensar este efecto, estas $750s$ personas se restaron equitativamente a las restantes m casillas, es decir, la lista nominal de cada una de dichas m casillas fue decrementada en $750s/m$, de forma que el listado nominal total quedó del mismo tamaño.

Los resultados

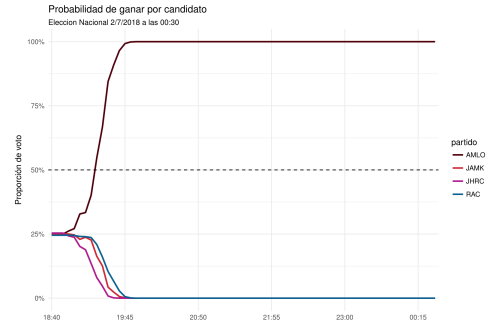
El modelo y las modificaciones discutidos fueron implementados en los conteos rápidos del 2018. Particularmente se estimaron las tendencias de la votación de la elección presidencial y, con las adecuaciones pertinentes por el número de candidatos contendientes, de las elecciones por las gubernaturas de Chiapas y Guanajuato. A continuación se discutirán con detalle los resultados para la elección presidencial.

El código desarrollado para el día de la elección buscaba información de forma automática en el servidor dispuesto para tal fin. En caso de existir información nueva se generaban 10,000 observaciones de la distribución posterior (5) para cada estrato $i = 1, 2, ..., 350$ y candidato $j = 1, 2, ..., 6$ (o de la inicial modificada (7) para el caso de estratos con menos de dos casillas). Posteriormente se obtenían las estimaciones a nivel federal θ_j vía (1) y éstas se transformaban mediante (2) a observaciones del vector $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. Además se estimaba la participación ρ utilizando (6), obteniendo así 10,000 observaciones de su distribución posterior.

Con esas simulaciones se obtuvieron intervalos del 95 % de credibilidad para cada candidato y para la participación, así como estimaciones puntuales utilizando la media de cada muestra. Además se estimó la probabilidad de ganar de cada candidato. Como complemento se generaban una serie de gráficas de verificación y control, así como un pequeño reporte de las estimaciones en un formato previamente establecido por el INE.



(a) Evolución de las estimaciones por candidato y remesa.



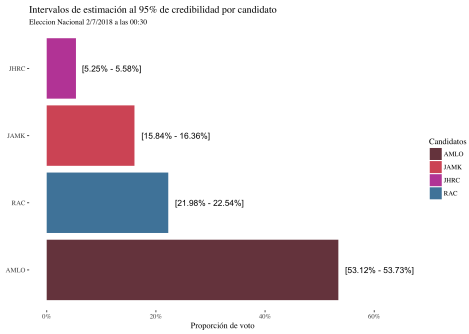
(b) Evolución de la probabilidad de ganar de cada candidato por remesa.

Figura 1: Evoluciones de las estimaciones de los candidatos (a) y sus probabilidades de ganar (b) con corte a las 00:30 horas del 2 de julio de 2018.

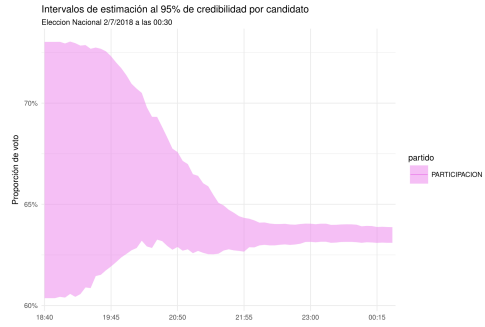
Las Figuras 1 y 2 muestra algunas de las gráficas generadas con corte a la remesa de las 00:30 horas del 2 de julio. Las Figuras 1a y 1b hacen evidente que la distribución inicial no aporta información al análisis, en tanto que los candidatos comienzan con la misma proporción estimada de votos en favor y la misma probabilidad de ganar (como es de esperar cuando hay pocas casillas en la muestra y, por lo tanto, muchos estratos sin información).

Por su parte la Figura 2a muestra las estimaciones producidas con la última remesa analizada, la correspondiente a las 00:30 horas del 2 de julio. La evolución de la estimación de la participación ρ se observa en la Figura 2b.

El comité decidió informar al INE de los resultados de la elección con la información de las 22:30 horas. Para ese momento solo se había recibido 67.5 % de la muestra originalmente planeada (5,254 de las 7,787 casillas). Sin embargo, se contaba con información proveniente de todas las entidades federativas y, más aún, de todos los estratos en el diseño. Con esa evidencia se procedió a reportar los resultados de la estimación que identificaron correctamente al ganador y, en ese sentido, contribuyeron, como era su objetivo, a dar certeza al proceso electoral, en el que por primera vez un presidente de izquierda obtuvo la presidencia.



(a) Tendencias finales de la votación con corte en la última remesa analizada.



(b) Evolución de las estimaciones para la participación ciudadana por remesa.

Figura 2: Tendencias finales de la votación (a) y evoluciones de las estimaciones de la participación ciudadana (b) con corte a las 00:30 horas del 2 de julio de 2018.

Es interesante observar que, si bien continuaron llegando datos al servidor interno del INE, las estimaciones realizadas con ellos no difirieron de los reportados a las autoridades. Lo mismo ocurrió con los conteos rápidos locales, ya que en todos los casos se logró determinar acertadamente a las y los ganadores de las diferentes contiendas electorales.

El código utilizado por los autores el día de la jornada electoral se desarrolló en R y está disponible en `https://github.com/GiankDiluvi/Mexico2018Elections_QuickCounts`. Se invita al público interesado a replicar los resultados aquí mostrados.

Referencias

[1] Mendoza, M. y Nieto Barajas, L. E. (2016). *Quick counts in the Mexican presidential elections: A Bayesian approach.* Electoral Studies. 43, 124-132.