IEEM
Instituto Electoral del Estado de México

4 de Junio 2017

COTECORA
Comité Técnico Asesor de los Conteos Rápidos

# Comité Técnico Asesor para el Conteo Rápido de la Elección para Gobernador del Estado de México 2017

# Criterios Científicos, Protocolo para Selección y Resguardo de la Muestra para la realización del Conteo Rápido

Mayo 3, 2017

# Contenido

# I. Antecedentes

1. El siete de septiembre de dos mil dieciséis, el Consejo General del Instituto Electoral del Estado de México, celebró sesión solemne por la que se dio inicio formalmente al Proceso Electoral Ordinario 2016-2017, para elegir al Gobernador Constitucional del Estado de México, para el periodo comprendido del 16 de septiembre de 2017 al 15 de septiembre de 2023.

2. En sesión extraordinaria del siete de septiembre de dos mil dieciséis, el Consejo General del Instituto Nacional Electoral aprobó el Acuerdo INE/CG661/2016, denominado: "Acuerdo del Consejo General del Instituto Nacional Electoral por el que se aprueba el Reglamento de Elecciones del Instituto Nacional Electoral", mismo que fue publicado en el Diario Oficial de la Federación, el trece de septiembre del mismo año.

3. El doce de septiembre de dos mil dieciséis, la H. "LIX" Legislatura Local, expidió el Decreto número 124, publicado en el Periódico Oficial del Gobierno del Estado Libre y Soberano de México, "Gaceta del Gobierno", en la misma fecha, en donde se convoca a los ciudadanos del Estado de México y a los partidos políticos con derecho a participar en la elección ordinaria para elegir al Gobernador Constitucional del Estado de México, para el periodo comprendido del 16 de septiembre de 2017 al 15 de septiembre de 2023.

4. En sesión ordinaria del veintidós de septiembre del dos mil dieciséis, el Órgano Superior de Dirección de este Instituto, aprobó el Acuerdo IEEM/CG/83/2016, por el que creó la Comisión Especial para la atención del Programa de Resultados Electorales Preliminares y Conteos Rápidos; la cual quedó integrada de la siguiente forma: Presidenta, Consejera Electoral Mtra. Palmira Tapia Palacios; como integrantes, los Consejeros Electorales, Mtro. Miguel Ángel García Hernández y Mtro. Saúl Mandujano Rubio; como Secretario Técnico el Secretario Ejecutivo y con un representante de cada partido político con registro ante este Instituto.

   En el Considerando XXVII, se estableció entre otros aspectos, como motivo de creación y dos de sus objetivos, los siguientes:

   "***Motivo de Creación***

   *Contar con una Comisión que auxilie al Consejo General en la coordinación de las actividades relativas a la implementación y operación del Programa de*

*Resultados Electorales Preliminares (PREP), así como para la realización de los conteos rápidos, en la elección de Titular del Poder Ejecutivo del Estado de México.*

***Objetivos***

- *Proponer al Consejo General la designación de los órganos internos responsables de la coordinación de las actividades del Programa de Resultados Electorales Preliminares (PREP) y de los Conteos Rápidos.*
- *…*
- *Supervisar la correcta integración y funcionamiento del Comité Técnico Asesor del Programa de Resultados Electorales Preliminares (COTAPREP) y Conteos Rápidos (COTECORA).*
- *…"*

**5.** En sesión extraordinaria celebrada el ocho de diciembre de dos mil dieciséis, la Comisión Especial para la atención del Programa de Resultados Electorales Preliminares y Conteos Rápidos, aprobó el Acuerdo IEEM/CEPREP y CR/03/2016 por el que propone al Consejo General, la integración del Comité Técnico Asesor de los Conteos Rápidos (COTECORA); y

**6.** El 14 de diciembre de 2016 fue aprobado por el Consejo General del IEEM el acuerdo IEEM/CG/115/2016 por el que se integra el Comité Técnico Asesor de los Conteos Rápidos para el Proceso Electoral 2016-2017, integrado por la Dra. Karla Beatriz Valenzuela Ocaña, el Dr. Alberto Alonso Y Coria, el Dr. Francisco Javier Aparicio Castillo, el Dr. Arturo Erdely Ruiz y el Dr. Carlos Erwin Rodríguez Hernández-Vela.

En apego a las disposiciones normativas en la materia, el COTECORA desarrolló los criterios científicos, así como el protocolo para la selección de la muestra, además conoció el proceso logístico operativo para el acopio de los datos de las Actas de Escrutinio y Cómputo de Casilla, y los consideró pertinentes para los fines del Conteo Rápido del Estado de México 2017. Los Criterios Científicos, el Protocolo para Selección de la Muestra y el Proceso Logístico Operativo propuesto se describen en este documento.

# II. Criterios científicos

Los criterios científicos del Conteo Rápido de la Elección del Estado de México son todos los procedimientos que, con base en la probabilidad y en la estadística, se usan para estimar el porcentaje de votos a favor de los candidatos para gobernador del estado.

## 1.1 Objetivos

El objetivo general de los conteos rápidos es crear certidumbre, confianza y transparencia en el proceso electoral, ofreciendo una estimación de las tendencias de la votación de alta calidad estadística en la noche del mismo día de la elección.

Los objetivos particulares del Conteo Rápido para la Elección de Gobernador en el Estado de México serán:

- Estimar el porcentaje de votos en favor de cada aspirante a gobernador por el Estado de México.
- Estimar el porcentaje de participación en la elección.

Todas las estimaciones del COTECORA estarán dadas en forma de intervalos de confianza del 95%.

## 1.2 Esquema de muestreo

### Diseño de muestreo

Las estrategias de estimación que implementarán los miembros del COTECORA estarán basadas en la misma información muestral. El proceso de selección será un muestreo aleatorio estratificado, en donde al interior de cada estrato se seleccionarán casillas mediante un muestreo aleatorio simple sin reemplazo.

Para definir los estratos, se usaron los 45 Distritos Locales del Estado de México, (ver Tabla 1).La muestra total será de alrededor de 1,800 casillas, esperando recibir al menos 1,200 casillas para realizar la estimación final. Vale la pena resaltar que el tamaño de muestra definitivo se calculará con base en la lista final de casillas a instalarse el 4 de junio y considerando la tasa de no respuesta de cada distrito local observada en elecciones pasadas.

La distribución de casillas en cada estrato se realizará vía afijación proporcional. Suponiendo afijación proporcional y utilizando como referencia la distribución de casillas instaladas para la elección federal del 7 de junio de 2015, la muestra quedaría distribuida como se sugiere en la Tabla 1.

**Tabla 1**. Distribución preliminar de la muestra para el Conteo Rápido en el Estado de México 2017 (con base en el listado de casillas al 5 de mayo de 2017).

| Distrito Local | Lista nominal | Casillas a instalarse ($N_h$) | Muestra ($n_h$) según lista nominal | Muestra ($n_h$) según número de casillas |
|---|---|---|---|---|
| 1 | 245,295 | 370 | 39 | 36 |
| 2 | 228,852 | 358 | 37 | 35 |
| 3 | 241,066 | 365 | 39 | 36 |
| 4 | 246,484 | 393 | 40 | 38 |
| 5 | 249,813 | 389 | 40 | 38 |
| 6 | 213,716 | 341 | 34 | 33 |
| 7 | 248,510 | 418 | 40 | 41 |
| 8 | 225,127 | 382 | 36 | 37 |
| 9 | 261,131 | 521 | 42 | 51 |
| 10 | 238,123 | 433 | 38 | 42 |
| 11 | 236,537 | 382 | 38 | 37 |
| 12 | 245,386 | 386 | 39 | 38 |
| 13 | 248,809 | 439 | 40 | 43 |
| 14 | 241,414 | 446 | 39 | 44 |
| 15 | 238,904 | 397 | 38 | 39 |
| 16 | 251,088 | 391 | 40 | 38 |
| 17 | 237,368 | 378 | 38 | 37 |
| 18 | 285,900 | 496 | 46 | 48 |
| 19 | 254,349 | 386 | 41 | 38 |
| 20 | 291,938 | 462 | 47 | 45 |
| 21 | 236,380 | 387 | 38 | 38 |
| 22 | 220,256 | 375 | 35 | 37 |
| 23 | 271,806 | 424 | 44 | 41 |
| 24 | 271,108 | 470 | 44 | 46 |
| 25 | 289,109 | 515 | 47 | 50 |
| 26 | 256,333 | 429 | 41 | 42 |
| 27 | 251,454 | 394 | 40 | 39 |
| 28 | 242,018 | 398 | 39 | 39 |
| 29 | 270,936 | 461 | 44 | 45 |
| 30 | 257,598 | 444 | 41 | 43 |
| 31 | 246,782 | 392 | 40 | 38 |
| 32 | 263,363 | 418 | 42 | 41 |
| 33 | 304,726 | 474 | 49 | 46 |
| 34 | 262,679 | 433 | 42 | 42 |
| 35 | 249,302 | 393 | 40 | 38 |
| 36 | 239,715 | 364 | 39 | 36 |
| 37 | 244,139 | 379 | 39 | 37 |
| 38 | 260,304 | 415 | 42 | 41 |
| 39 | 258,304 | 427 | 42 | 42 |
| 40 | 246,248 | 411 | 40 | 40 |
| 41 | 282,493 | 500 | 45 | 49 |
| 42 | 211,926 | 370 | 34 | 36 |
| 43 | 252,537 | 419 | 41 | 41 |
| 44 | 269,486 | 416 | 43 | 41 |
| 45 | 224,105 | 364 | 36 | 36 |
| **TOTAL** | **11,312,917** | **18,605** | **1,818** | **1,818** |

## Consideraciones para definir el tamaño de muestra y diseño de muestreo

*Ideas básicas del muestreo probabilístico para el Conteo Rápido*

Como se describió anteriormente, el objetivo particular del conteo rápido es estimar el porcentaje,$\boldsymbol{p}$,de votos enfavor de un candidato a gobernador[1]. Esta cantidad es el valor real y se obtiene una vez finalizados los cómputos distritales de la Elección. Sin embargo, mediante la teoría del muestreo probabilístico,se selecciona una muestra aleatoria de$\boldsymbol{n}$ casillas, de un total de $\boldsymbol{N}$ instaladas, y con la información recuperada en la tarde-noche del día de la elecciónse calcula el estimador $\widehat{\boldsymbol{p}}$.Utilizando las herramientas del muestreo, es posible definir **estrategias de selección,tamaños de muestra** y **estimadores** para asegurar que

$$\left|\boldsymbol{p} - \widehat{\boldsymbol{p}}\right| \leq \boldsymbol{\epsilon}, \qquad (\mathbf{1})$$

con un 95% de confianza. A la cantidad $\boldsymbol{\epsilon}$ se le conoce comúnmente como **precisión** o **error máximo aceptable** en la estimación y se fija de acuerdo a las exigencias de la Elección y a la capacidad operativa de campo. La expresión (1) se puede escribir de forma equivalente en términos de intervalos de confianza

$$\widehat{\boldsymbol{p}} - \boldsymbol{\epsilon} \leq \boldsymbol{p} \leq \widehat{\boldsymbol{p}} + \boldsymbol{\epsilon}. \qquad (2)$$

Lo anterior, se pueden interpretar tomando como referencia el siguiente ejemplo: si se extraen 100 muestras distintas e independientes una de otra (cada una siguiendo la misma estrategia de selección,usando el mismo tamaño de muestra y el mismo estimador) y con cada muestra se genera una estimación, de tal manera que se obtengan 100 estimaciones (independientes)

$$\widehat{\boldsymbol{p}}_{\mathbf{1}}, \widehat{\boldsymbol{p}}_{\mathbf{2}}, \widehat{\boldsymbol{p}}_{\mathbf{3}}, \widehat{\boldsymbol{p}}_{\mathbf{4}}, \widehat{\boldsymbol{p}}_{\mathbf{5}}, \dots, \widehat{\boldsymbol{p}}_{\mathbf{100}},$$

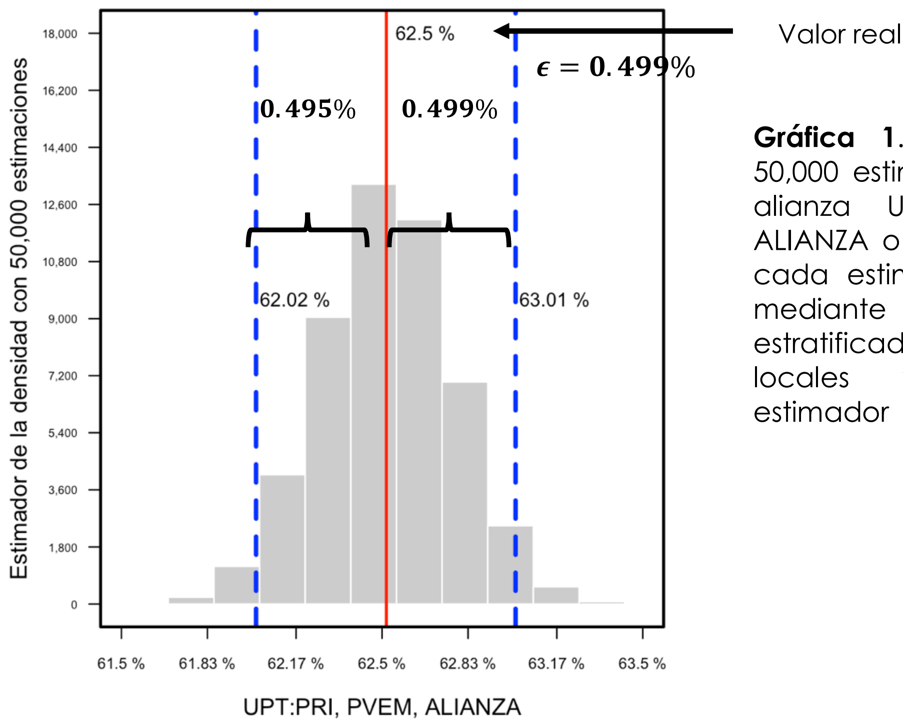
aproximadamente 95 cumplirán con que su distancia a $\boldsymbol{p}$, es menor o igual a $\boldsymbol{\epsilon}$.

*Precisión en la estimación*

La idea anterior nos brinda un procedimiento mediante simulación Monte Carlo para obtener la precisión de varias estrategias de selección, con diferentes tamaños de muestra, y así decidir cuál es el que mejor se ajusta a los objetivos y condiciones del

[1]Esto se hará para cada contendiente así como para la estimación de la votación en favor de candidatos no registrados, votación nula y participación en la elección, sin embargo para efectos de la explicación nos concentraremos en la estimación para un solo contendiente.

conteo rápido en el Estado de México. Por ejemplo, para generar la Gráfica 1 se seleccionaron 50,000 muestras aleatorias independientes de casillas de la base de datos de los Cómputos Distritales para la Elección de Gobernador en el Estado de México en 2011. Cada muestra fue de 1,000 casillas y se siguió un muestreo estratificado por distritos locales (45 estratos).




**Gráfica 1**. Histograma de 50,000 estimaciones para la alianza UPT (PRI, PVEM, ALIANZA o PANAL) en 2011, cada estimación se realizó mediante un muestreo estratificado por distritos locales y usando el estimador común de razón.

Con cada una de las 50,000 muestras se calculó el valor del estimador de razón:

$$\widehat{\boldsymbol{p}}_{\boldsymbol{i}} = \frac{\sum_{h=1}^{L} N_h \overline{y}_{hi}}{\sum_i \sum_{h=1}^{L} N_h \overline{y}_{hi}} \qquad (3)$$

En donde $\boldsymbol{i}$ es el candidato i-ésimo (incluyendo también el número de votos por candidatos no registrados y nulos), $\boldsymbol{h}$ es el número de estrato, $\boldsymbol{L}$ el número total de estratos, $\boldsymbol{N_h}$ es el número de casillas instaladas en el estrato $\boldsymbol{h}$ y $\overline{\boldsymbol{y}}_{\boldsymbol{hi}}$ es el promedio de votos, en las casillas seleccionadas en muestra, para el candidato a gobernador por el partido $\boldsymbol{i}$ en el estrato $\boldsymbol{h}$.

Las estimaciones se realizaron para todos los candidatos (incluyendo candidatos no registrados y votos nulos), sin embargo, en este apartido sólo se mostrará el análisis para el candidato que registró mayor varianza en su votación y que por lo tanto requería un tamaño de muestra mayor para alcanzar la misma precisión que las estimaciones para los demás candidatos.

Las líneas punteadas (azules), en la Gráfica 1, señalan los cuantiles $q_{0.025}$ y $q_{0.975}$:

- Por debajo del primero se encuentra el 2.5% de las estimaciones y por arriba el 97.5%.
- Por debajo del segundo se encuentra el 97.5% de las estimaciones y por arriba el 2.5%.
- Entre ambos cuantiles se encuentra el 95% de las estimaciones.

El valor real, $p = 62.5\%$, se señala con la línea roja, por lo que las 50,000 estimaciones ajustan de forma adecuada este valor.

Regresando a las expresiones (1) y (2), se tiene que

$$q_{0.025} = p - \epsilon \ y \ q_{0.975} = p + \epsilon.$$

Por lo tanto, la precisión se obtiene mediante la expresión

$$\epsilon = max\{p - q_{0.025}, q_{0.975} - p\}.$$

Entonces, con la estratificación y tamaño de muestra para generar la Gráfica 1, se obtiene una precisión de $\epsilon = 0.499\%$. Siguiendo este procedimiento se obtuvieron las precisiones para varias estratificaciones y tamaños de muestra.

*Estrategias de selección*

Existen muchas estrategias de selección, pero de forma general podemos hablar de cuatro que son la base para las demás: el muestreo aleatorio simple sin reemplazo (MASSR), muestreo aleatorio simple con reemplazo (MASCR), el muestreo estratificado y el muestreo por conglomerados o polietápico.

- El MASSR es el muestreo básico y suele usarse como parte de los demás esquemas, en este caso se tiene poco control de las casillas seleccionadas. Para explicar este tipo de estrategia, se recurre a una urna con $N$objetos numerados y el objetivo es seleccionar una muestra de $n$, con$n < N$. En cada selección, se revuelve la urna y se toma unobjeto al azar. El objeto seleccionado se deja afuera de la urna para la siguiente selección (sin reemplazo), en la selección j-ésima quedan $N - (j - 1)$objetos en la urna.

- El MASCR es muy similar al MASSR, pero después de cada selección, el objeto se regresa a la urna, de forma que en cada extracción se obtienen $\boldsymbol{N}$ objetos. En este caso es necesario registrar el número del objeto seleccionado.

- En el muestreo estratificado en cambio, se divide a la población en estratos/grupos[2] y en cada estrato se hace una selección de casillas.[3] En este caso se tiene mayor control de la muestra,ya que en cada estrato siempre se tendrán casillas seleccionadas, logrando una mejor distribución de la misma. Otra ventaja, es que permite obtener una mejor estimación: si los estratos se definen de manera que, al interior de cada estrato, la variable de interés tenga un comportamiento más homogéneo que si se considerara la población completa, entonces la estimación tendrá más precisión que el MASSR o el MSCR.

- Por último, tenemos el muestreo por conglomerados, en donde también se divide a la población en conglomerados/grupos y la selección se realiza en etapas, primero se seleccionan conglomerados (generalmente mediante MASCR) y posteriormente, y sólo de los conglomerados seleccionados en la primera etapa, se eligen casillas (generalmente mediante MSSR). El muestreo por conglomerados se usa para reducir costos y facilitar la logística operativa de campo, sin embargo, se sacrifica precisión en la estimación.

Por lo descrito en el párrafo anterior, lo que se hizo para determinar el tamaño de muestra en para el Conteo Rápido del Estado de México fue probar distintas estratificaciones y tamaños de muestra eligiendo la combinación que permitiera obtener la mejor precisión, pero cuidando no rebasar la capacidad operativa de campo.

*Poblaciones de referencia para realizar las simulaciones Monte Carlo*

En todos los ejercicios realizados se tomaron como referencia los cómputos distritales de tres votaciones:

1. Elección para Gobernador en el Estado de México del 3 de julio de 2011.
    - 17,489 casillas en total $\Rightarrow$ 17,472 casillas para análisis (17 casillas).
2. Elección Presidencial, tomando en cuenta sólo los resultados en el Estado de México, del 1 de julio de 2012.
    - 17,350 casillas en total $\Rightarrow$ 17,229 casillas para análisis (121 casillas).

[2] los estratos son ajenos entre sí y su unión conforma la población completa
[3] La selección en cada estrato usualmente se realiza mediante un MASS.

3. Elección Federal de Diputados del 7 de junio de 2015. En este caso se asumió que los candidatos fueron los mismos en los 40 Distritos Federales y que hubo dos coaliciones: PRI-PVEM y PRD-PT.
    - 18,250 casillas en total $\Rightarrow$ 18,115 casillas para análisis (135 casillas).

La diferencia entre las casillas totales y las casillas para análisis se debe a que no fue posible determinar el DISTRITO LOCAL de algunas casillas. Sin embargo, esto no modifica en manera alguna el análisis para determinar el tamaño de muestra.

*Estratificaciones y tamaños de muestra*

Las estratificaciones consideradas en los ejercicios de simulación, para cada elección, fueron seis:

1. Distritos Federales $\Rightarrow$ 40 estratos.
2. Distritos Federales y al interior de cada distrito se dividió por sección urbana no urbana $\Rightarrow$ 64 estratos.
3. Distritos Locales $\Rightarrow$ 45 estratos.
4. Distritos Locales y al interior de cada distrito se dividió por sección urbana y no urbana $\Rightarrow$ 75 estratos.
5. Municipios $\Rightarrow$ 125 estratos.
6. Municipios y al interior de cada municipio se dividió por sección urbana y no urbana $\Rightarrow$ 206 estratos.

Para distribuir el tamaño de muestra en cada uno de los estratos, se utilizó lo que se conoce como "afijación proporcional":

$$\boldsymbol{n_h = n\ \frac{N_h}{N}},$$

en donde $\boldsymbol{n_h}$es el tamaño de muestra que se seleccionará en el estrato ***h***, ***n*** el tamaño de muestra total, $\boldsymbol{N_h}$ el número total de casillas instaladas en el estrato ***h*** y $\boldsymbol{N}$el número total de casillas instaladas[4] en el Estado de México. En palabras comunes, el tamaño de muestra total, se distribuirá proporcionalmente de acuerdo al número de casillas instaladas en cada estrato. En estratos con mayor número de casillas se seleccionará mayor muestra.

*Resultados de las simulaciones*

Lo que se busca es la estratificación que alcance la mayor precisión (o error máximo

[4]Se toman las casillas para análisis descritas en la sección anterior.

aceptable) con el menor tamaño de muestra posible. Experiencias anteriores indican que una precisión de

$$\epsilon = 0.5 \text{ (medio punto porcentual)}$$

es una buena meta esperada toda vez que, aunque requiere una muestra relativamente grande, ésta es factible y manejable dado el número de personas en campo el día de la jornada electoral.

En todas las simulaciones, su utilizó el estimador de razón dado por la expresión (3), los resultados obtenidos con cada una de las tres Elecciones, se muestran en las Tablas 2, 3 y 4 (para obtener las precisiones en cada celda de la tabla se seleccionaron 50,000 muestras independientes y con cada muestra se hizo una estimación, ver Gráfica 1).

**Tabla 2**. Elección para Gobernador 2011: precisiones obtenidas usando el estimador (3), para distintos tamaños de muestra y estratificaciones.

| TAMAÑO DE MUESTRA | DISTRITOS FEDERALES | DISTRITOS FEDERALES URBANO - NO URBANO | DISTRITOS LOCALES | DISTRITOS LOCALES URBANO - NO URBANO | MUNICIPIOS | MUNICIPIOS URBANO - NO URBANO |
|---|---|---|---|---|---|---|
| **700** | 0.62 | 0.62 | 0.60 | 0.60 | 0.58 | 0.56 |
| **750** | 0.60 | 0.59 | 0.58 | 0.57 | 0.56 | 0.54 |
| **800** | 0.58 | 0.57 | 0.56 | 0.55 | 0.55 | 0.53 |
| **850** | 0.56 | 0.55 | 0.55 | 0.53 | 0.53 | 0.52 |
| **900** | 0.54 | 0.53 | 0.53 | 0.52 | 0.51 | 0.50 |
| **950** | 0.53 | 0.53 | 0.51 | 0.51 | 0.49 | 0.48 |
| **1,000** | 0.52 | 0.51 | 0.50 | 0.49 | 0.48 | 0.47 |
| **1,050** | 0.50 | 0.50 | 0.48 | 0.49 | 0.47 | 0.46 |
| **1,100** | 0.49 | 0.48 | 0.47 | 0.47 | 0.46 | 0.45 |
| **1,150** | 0.48 | 0.48 | 0.46 | 0.46 | 0.45 | 0.44 |
| **1,200** | 0.47 | 0.47 | 0.45 | 0.45 | 0.44 | 0.43 |
| **1,250** | 0.46 | 0.45 | 0.44 | 0.43 | 0.43 | 0.42 |
| **1,300** | 0.45 | 0.44 | 0.43 | 0.43 | 0.42 | 0.41 |
| **1,350** | 0.44 | 0.44 | 0.42 | 0.42 | 0.42 | 0.41 |
| **1,400** | 0.44 | 0.43 | 0.42 | 0.41 | 0.40 | 0.39 |

**Tabla 3**. Elección Presidencial 2012: precisiones obtenidas usando el estimador (3), para distintos tamaños de muestra y estratificaciones.

| TAMAÑO DE MUESTRA | DISTRITOS FEDERALES | DISTRITOS FEDERALES URBANO - NO URBANO | DISTRITOS LOCALES | DISTRITOS LOCALES URBANO - NO URBANO | MUNICIPIOS | MUNICIPIOS URBANO - NO URBANO |
|---|---|---|---|---|---|---|
| **700** | 0.55 | 0.54 | 0.52 | 0.52 | 0.51 | ALGUNOS ESTRATOS CON 1 ÚNICA CASILLA INSTALADA |
| **750** | 0.53 | 0.52 | 0.50 | 0.50 | 0.49 | |
| **800** | 0.51 | 0.50 | 0.49 | 0.48 | 0.48 | |
| **850** | 0.49 | 0.48 | 0.48 | 0.46 | 0.46 | |
| **900** | 0.48 | 0.47 | 0.46 | 0.45 | 0.45 | |
| **950** | 0.46 | 0.46 | 0.44 | 0.44 | 0.43 | |
| **1,000** | 0.45 | 0.45 | 0.44 | 0.43 | 0.42 | |
| **1,050** | 0.44 | 0.43 | 0.42 | 0.42 | 0.42 | |
| **1,100** | 0.43 | 0.42 | 0.41 | 0.41 | 0.40 | |
| **1,150** | 0.42 | 0.41 | 0.40 | 0.40 | 0.40 | |
| **1,200** | 0.41 | 0.40 | 0.39 | 0.39 | 0.39 | |
| **1,250** | 0.40 | 0.39 | 0.38 | 0.38 | 0.38 | |
| **1,300** | 0.39 | 0.39 | 0.38 | 0.37 | 0.37 | |
| **1,350** | 0.38 | 0.38 | 0.37 | 0.36 | 0.36 | |
| **1,400** | 0.37 | 0.37 | 0.36 | 0.36 | 0.35 | |

**Tabla 4**. Elección Federal de Diputados 2015: precisiones obtenidas usando el estimador (3), para distintos tamaños de muestra y estratificaciones.

| TAMAÑO DE MUESTRA | DISTRITOS FEDERALES | DISTRITOS FEDERALES URBANO - NO URBANO | DISTRITOS LOCALES | DISTRITOS LOCALES URBANO - NO URBANO | MUNICIPIOS | MUNICIPIOS URBANO - NO URBANO |
|---|---|---|---|---|---|---|
| **700** | 0.68 | 0.67 | 0.66 | 0.66 | 0.59 | 0.56 |
| **750** | 0.67 | 0.65 | 0.64 | 0.63 | 0.57 | 0.55 |
| **800** | 0.64 | 0.62 | 0.61 | 0.61 | 0.55 | 0.53 |
| **850** | 0.62 | 0.61 | 0.59 | 0.59 | 0.52 | 0.51 |
| **900** | 0.60 | 0.59 | 0.58 | 0.57 | 0.51 | 0.50 |
| **950** | 0.58 | 0.58 | 0.56 | 0.56 | 0.50 | 0.48 |
| **1,000** | 0.57 | 0.55 | 0.55 | 0.54 | 0.49 | 0.48 |
| **1,050** | 0.55 | 0.54 | 0.54 | 0.52 | 0.48 | 0.47 |
| **1,100** | 0.53 | 0.53 | 0.52 | 0.51 | 0.47 | 0.45 |
| **1,150** | 0.52 | 0.52 | 0.51 | 0.50 | 0.45 | 0.44 |
| **1,200** | 0.51 | 0.51 | 0.50 | 0.49 | 0.44 | 0.44 |
| **1,250** | 0.51 | 0.49 | 0.48 | 0.47 | 0.43 | 0.42 |
| **1,300** | 0.50 | 0.48 | 0.48 | 0.47 | 0.43 | 0.42 |
| **1,350** | 0.48 | 0.48 | 0.46 | 0.46 | 0.42 | 0.40 |
| **1,400** | 0.47 | 0.47 | 0.46 | 0.45 | 0.41 | 0.40 |

Estos resultados muestran que, de las tres elecciones, la que requiere el menor tamaño de muestra para alcanzar una mejor precisión es la Presidencial, le sigue la Elección de Gobernador y la que exige un mayor tamaño de muestra es la Elección de Diputados. Esto es debido a que la varianza de la votación es mayor en la Elección de Diputados, en nuestro análisis supusimos que hubo 8 candidatos (con 2 coaliciones) en los 40 Distritos Federales que componen es Estado de México, sin embargo, en la realidad en cada distrito hubo candidatos distintos. Esto introduce una gran variabilidad en la votación para todo el estado. Entonces, el tamaño de muestra que se derive del análisis de la votación para Diputados necesariamente será suficiente para las otras votaciones.

En la Tabla 3, la mejor estratificación es la que contempla a los municipios, ya que el tamaño de muestra requerido para alcanzar un error de $\boldsymbol{\epsilon = 0.5}$es de $\boldsymbol{n = 950}$ casillas. Sin embargo, existe un problema práctico con esta estratificación. El detalle se resume en la Tabla 5.

**Tabla 5**. Elección Federal de Diputados 2015: Estadísticas básicas para cada estratificación y las casillas instaladas en cada estrato.

| Variable de interés | DISTRITOS FEDERALES | DISTRITOS FEDERALES URBANO - NO URBANO | DISTRITOS LOCALES | DISTRITOS LOCALES URBANO - NO URBANO | MUNICIPIOS | MUNICIPIOS URBANO - NO URBANO |
|---|---|---|---|---|---|---|
| Número de estratos | 40 | 64 | 45 | 75 | 125 | 194 |
| Promedio de casillas instaladas | 453 | 283 | 403 | 242 | 145 | 93 |
| Máximo casillas instaladas | 678 | 606 | 522 | 522 | 1,966 | 1,966 |
| Mínimo casillas instaladas | 350 | 3 | 302 | 2 | 6 | 2 |

El mínimo de casillas instaladas en uno de los municipios es 6, de hecho, en 23 municipios se instalaron menos de 20 casillas. Esto representa un gran riesgo para el conteo rápido, de elegirse los municipios como estratos, de cada municipio se extraería una muestra de casillas y con la información que se recupere de esta muestra, se realiza una estimación. Finalmente, se conjuntan las estimaciones de todos los estratos para obtener la estimación global, ver ecuación (3).

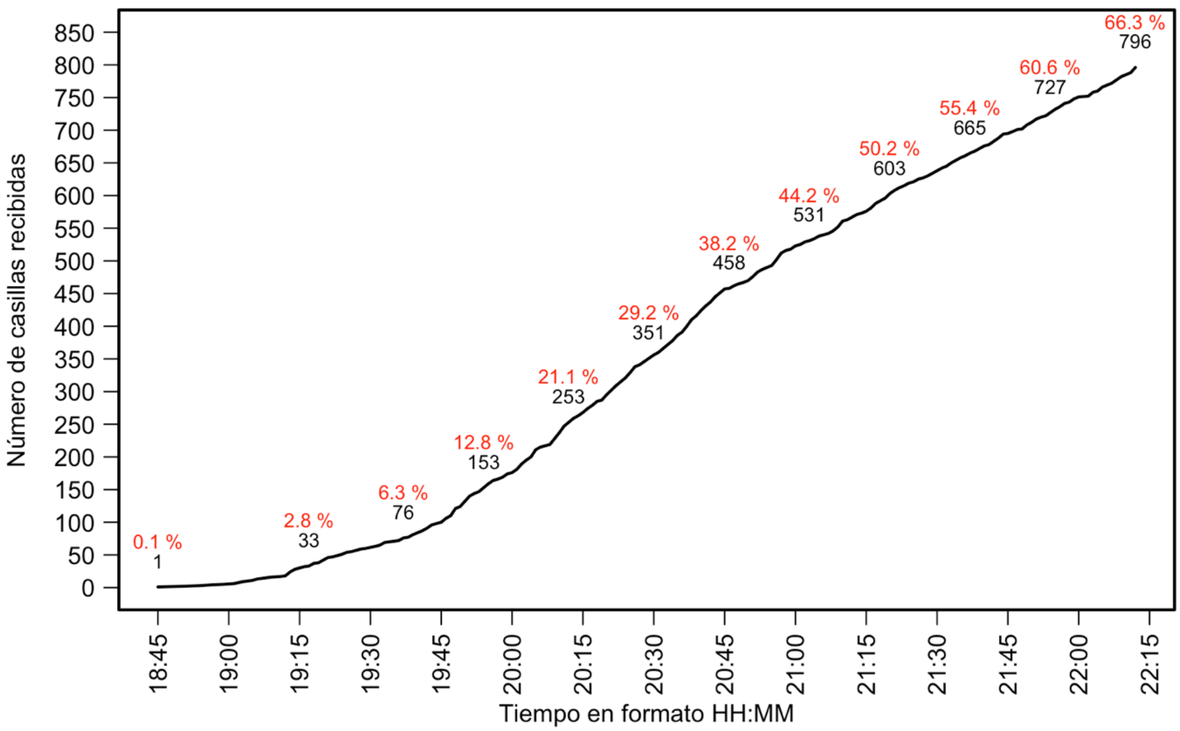
Con estratos con tan pocas casillas instaladas, aunque se extraiga una muestra grande (proporcionalmente), la probabilidad de que no se recupere ninguna casilla en estos estratos es muy alta. Si no se tiene información en algún estrato, será imposible realizar la estimación. Lo mismo ocurre si se utiliza alguna estratificación que contemple la parte urbana y no urbana, el número de casillas instaladas en algunos estratos es muy pequeño. Este riesgo se reduce de manera sustancial, si se toman los Distrito Locales como estratos, en este caso el mínimo de casillas instaladas es de 302, el máximo es de 522 y en promedio se instalan 403 casillas por Distrito Local. Además, también se alcanza una buena precisión con un tamaño de muestra aceptable.

Tomando los Distritos Locales como estratos, se obtendría una precisión de $\boldsymbol{\epsilon = 0.5}$ con $\boldsymbol{n = 1,200}$ casillas, además, utilizando la afijación proporcional el promedio de casillas en muestra por estrato sería de 27, con un máximo de 35 y mínimo de 20.

*No respuesta o casillas no reportadas*

En cualquier encuesta existe un porcentaje de no respuesta, en el caso del Conteo Rápido serían casillas seleccionadas en muestra, pero que por diversos motivos su información no se recibe a tiempo para hacer la estimación. En el Conteo Rápido para la Elección Federal de Diputados de 2015, la muestra en el Estado de México fue de 1,200 casillas (30 casillas por cada Distrito Federal), sin embargo, a las 10:00 pm sólo se había recibido el 60 % y a las 10:15 pm, cuando se hizo la última estimación, se tenía aproximadamente el 66 % de la muestra total. El número de casillas recibidas, así como la hora de arribo se muestra en la Gráfica 2.

**Gráfica 2**. Conteo Rápido INE 2015: Serie de tiempo para el número de casillas recibidas del Estado de México.




Si el mismo comportamiento se mantiene para el Conteo Rápidos en la Elección de Gobernador en 2017, para realizar la estimación final a las 10:00 pm se necesitaría un tamaño de muestra total, considerando la tasa de no respuesta, de

$$\boldsymbol{n_{Total} = \frac{1,200}{0.66} = 1,818\ casillas.}$$

Con este tamaño de muestra total, se esperaría contar con un tamaño de muestra efectivo de $\boldsymbol{n = 1,200}$ casillas alrededor de las 10:00 pm del 4 de junio.

*Casillas a reportar por cada CAE*

Uno de los retos logísticos del levantamiento de datos es la proporción de CAEs que participarán en la muestra, por un lado, y el número de casillas que deberá reportar cada CAE. Con una muestra total de 1,818 casillas y 3,352 CAEs en campo el día de la jornada electoral, se pueden estimar, tentativamente, ambas proporciones.

- Como sugiere la Tabla 6, una muestra de 1818 casillas implica que alrededor de 1452 CAEs (43% del total) deban reportar al menos una casilla.
- La gran mayoría de los CAEs seleccionados en la muestra deberán reportar tan solo una casilla (78%).
- Se estima que 22% de los CAEs deban reportar dos o más casillas, para lo cual se podría contar con el apoyo de los supervisores electorales (SE).
- En un escenario límite, si todos los CAEs seleccionados en la muestra reportarán tan solo una casilla, se podría recabar una muestra de 1,452 casillas.

**Tabla 6.** Número estimado de casillas a reportar por CAE

| Casillas | CAEs | % | Muestra | %2 |
|---|---|---|---|---|
| 1 | 1,131 | 77.9% | 1,131 | 62.2% |
| 2 | 280 | 19.3% | 560 | 30.8% |
| 3 | 38 | 2.6% | 114 | 6.3% |
| 4 | 3 | 0.2% | 12 | 0.7% |
| | **1,452** | **100.0%** | **1,817** | **100.0%** |

*Consideraciones finales del diseño muestral*

Por último, es importante tomar en cuenta los siguientes puntos

1. La precisión se maneja tanto en la etapa de planeación del conteo rápido, como en la presentación de los resultados finales del mismo. En la etapa de planeación, se usa para determinar un tamaño de muestra capaz (teóricamente) de alcanzar el nivel de precisión y confianza deseadas en la estimación. Sin embargo, debido a que este tamaño de muestra se fija tomando como referencia los resultados de una o varias elecciones previas, una vez concluida la elección y con toda la información disponible se determinará la precisión realmente obtenida.

2. Los resultados de las simulaciones presentadas son únicamente para el método de estimación usado (ecuación 3), y tienen como objetivo fijar un tamaño de muestra. Sin embargo, se pueden emplear métodos de estimación alternativos (usando la misma estrategia de selección) que arrojarán precisiones distintas.

3. Para la elección en el Estado de México, se espera contar con 3,352 capacitadores asistentes electorales (CAES), mismos que recabarán la información de las casillas seleccionadas en muestra y procederán a su transmisión. Es conveniente apuntar que 1,818 casillas como tamaño de muestra se encuentra dentro de la capacidad operativa de campo.

## 1.3 Estimación e integración de los intervalos de confianza

Para dar mayor certeza a la estimación, los integrantes del COTECORA utilizarán tres enfoques estadísticos distintos para estimar los resultados de la elección:

- Método Bayesiano, a cargo del Dr. Arturo Erdely Ruiz.
- Método Clásico vía las fórmulas para estimadores de razón, derivadas mediante las técnicas de muestreo probabilístico, a cargo de la Dra. Karla Beatriz Valenzuela Ocaña y el Dr. Alberto Alonso y Coria.
- Método Clásico vía métodos de remuestreo, a cargo del Dr. Carlos Erwin Rodríguez.

Dado que se aplicarán diversos métodos de estimación, no se espera que los límites inferior y superior de cada intervalo sean estrictamente idénticos, toda vez que los métodos que se utilizarán introducen elementos aleatorios. Sin embargo, el objetivo de los procedimientos de estimación será el mismo, las estimaciones se consolidarán para presentar resultados únicos en forma de intervalos de confianza.

Las estrategias de estimación que implementarán los miembros del COTECORA estarán basadas en la misma muestra. Durante toda la jornada, el COTECORA monitoreará que las tendencias de votación indicadas por los intervalos de confianza estimados vayan en el mismo sentido.

Para integrar las estimaciones en una sola, se aplicará el siguiente procedimiento: Sea $\boldsymbol{I_k} = [\boldsymbol{a_k}, \boldsymbol{b_k}]$ una estimación por intervalo de un porcentaje de votación, donde el subíndice $\boldsymbol{k} = \mathbf{1}, \dots, \boldsymbol{r}$ representa cada una de las $\boldsymbol{r}$estimaciones que realizarán los miembros del COTECORA. El intervalo combinado $\boldsymbol{I} = [\boldsymbol{a}, \boldsymbol{b}]$ que reportará el COTECORA se obtendrá mediante las fórmulas:

$$a = \frac{promedio\{a_1,\ldots,a_r\} + mediana\{a_1,\ldots,a_r\}}{2} \quad \text{y} \quad b = \frac{promedio\{b_1,\ldots,b_r\} + mediana\{b_1,\ldots,b_r\}}{2}$$

Utilizar solo promedios tiene la ventaja de incorporar directamente información de todos los intervalos, pero la desventaja de que un solo valor atípico podría tener demasiada influencia. Utilizar solo medianas tiene la ventaja de ser inmune ante un valor atípico, pero la desventaja de que varios intervalos podrían quedar totalmente fuera de consideración. Es por lo anterior que se escoge un término medio entre ambas posibilidades.

## 1.4 Procedimientos de estimación

### A. Propuesta del Dr.Arturo Erdely Ruiz

# Estimación bayesiana por intervalos de porcentajes de votación

Dr. Arturo Erdely Ruiz

Abril de 2017

**Resumen**

Se propone la utilización de la distribución de probabilidad multinomial, y la correspondiente inferencia bayesiana de sus parámetros, para la estimación por intervalo de los porcentajes del voto emitido que obtendrán los candidatos en la elección para gobernador del Estado de México del año 2017, a partir de un conteo rápido de actas electorales por casilla.

## 1. Introducción

Si una variable o vector aleatorio $X$ con una función de densidad o masa de probabilidades $f(x\,|\,\theta)$ cuantifica adecuadamente la incertidumbre sobre un fenómeno aleatorio de interés, pero su parámetro (o vector de parámetros) $\theta \in \mathfrak{S}$ es desconocido, donde $\mathfrak{S}$ es el conjunto de valores admisibles para $\theta$ (espacio paramétrico), el enfoque bayesiano de la estadística propone cuantificar la incertidumbre sobre $\theta$ mediante una variable aleatoria $\Theta$ con una distribución de probabilidad sobre $\mathfrak{S}$, misma que puede basarse en información inicial con que se cuente en un momento dado y que se denomina *distribución a priori,* y que al combinarse con información muestral se convierte en una *distribución a posteriori.*

Sea $(X_1,\ldots,X_m)$ una muestra aleatoria con función de densidad o masa de probabilidades conjunta $f(y_1,\ldots,y_m\,|\,\theta)$. Si $(x_1,\ldots,x_m)$ son los valores observados de dicha muestra, y se denota mediante $\pi(\theta)$ a la función de densidad de probabilidades a priori, y mediante $p(\theta\,|\,x_1,\ldots,x_m)$ a la función de densidad de probabilidades a posteriori, entonces mediante la *Regla de Bayes* se obtiene la siguiente fórmula:

$$p(\theta\,|\,x_1,\ldots,x_m) \,=\, \frac{1}{B}\,\pi(\theta)f(x_1,\ldots,x_m\,|\,\theta)\,, \quad \theta \in \mathfrak{S}, \tag{1}$$

donde la constante $B = \int_{\mathfrak{S}} \pi(\theta)f(x_1,\ldots,x_m\,|\,\theta)\,d\theta$. Mediante (1) se realiza cualquier inferencia estadística de interés sobre el parámetro (o vector de parámetros) $\theta$. Para el caso de *estimación puntual,* misma que consiste en escoger un valor $\widehat{\theta} \in \mathfrak{S}$ con base en toda la información disponible, puede calcularse alguna medida de tendencia central (la media o la mediana, por ejemplo) por medio de $p(\theta\,|\,x_1,\ldots,x_m)$.

En el caso de *estimación por intervalo* de probabilidad $0 < \alpha < 1$, si $\theta$ es de dimensión 1 entonces se busca un intervalo $I(\alpha) := [a,b] \subset \mathfrak{S}$ tal que:

$$\mathbb{P}[\theta \in I(\alpha)] \,=\, \int_a^b p(\theta\,|\,x_1,\ldots,x_m)\,d\theta \,=\, \alpha. \tag{2}$$

Usualmente la solución para $I(\alpha)$ en (2) no es única, por lo que de entre el conjunto de soluciones posibles $\{[a,b] : \int_a^b p(\theta \,|\, x_1,\ldots,x_m)\, d\theta = \alpha\}$ se escoge aquel intervalo cuya longitud $b-a$ sea mínima. En caso de que la dimensión de $\mathfrak{S}$ sea $J > 1$ entonces la densidad de probabilidades a posteriori (1) de $\theta = (\theta_1,\ldots,\theta_J)$ es multivariada y por tanto deberá marginalizarse para cada componente $\theta_j$ , y mediante $p(\theta_j \,|\, x_1,\ldots,x_m)$ hacer estimación puntual y por intervalo para cada $\theta_j$ con $j \in \{1,\ldots,J\}$. Si $p(\theta_j \,|\, x_1,\ldots,x_m)$ no es simétrica respecto a algún valor $\theta_j = \theta_0$, obtener la estimación por intervalo $I_j(\alpha)$ que sea de longitud mínima pudiera resultar un tanto laborioso o computacionalmente intensivo, y por tanto es común ocupar la solución pseudo-óptima que consiste en $I_j(\alpha) = [a_j, b_j]$ tal que:

$$\int_{-\infty}^{a_j} p(\theta_j \,|\, x_1,\ldots,x_m)\, d\theta_j \;=\; \frac{1-\alpha}{2} \;=\; \int_{b_j}^{+\infty} p(\theta_j \,|\, x_1,\ldots,x_m)\, d\theta_j\,. \tag{3}$$

## 2. Notación

Se adopta prácticamente tal cual la notación utilizada en Mendoza y Nieto-Barajas (2016). El *marco muestral* disponible consiste en un total de $K$ casillas electorales que se instalarán el día de la elección para que un total de $n$ posibles votantes (tamaño de la *lista nominal*) estén en condiciones de ejercer su derecho al voto, mismos que pueden decidir sufragar a favor de algún candidato registrado, o a favor de alguno no registrado (NOREG), o anular su voto (NULOS) o simplemente abstenerse de acudir a votar (ABST). Al conjunto de posibilidades anteriores les asignaremos las categorías $1, 2, \ldots, J$ reservando siempre $J-2$ para NOREG, $J-1$ para NULOS y $J$ para ABST, esto es, los candidatos registrados serán $1,\ldots,J-3$, con $J \geq 4$. En lo sucesivo cuando se haga referencia a un *candidato* $j \in \{1,\ldots,J\}$ deberá entenderse en un sentido amplio, es decir considerando como "candidatos" incluso a las categorías NOREG, NULOS y ABST.

El total de $K$ casillas electorales se reparten en $N$ *estratos* o subconjuntos (distritos locales en este caso), y se denotará $K_i$ al número de casillas en el distrito $i \in \{1,\ldots,N\}$ de modo que $\sum_{i=1}^N K_i = K$, y mediante $n_i$ al tamaño de la lista nominal en el distrito $i$ de modo que $\sum_{i=1}^N n_i = n$. Al momento del cierre de las casillas se procede al conteo de votos emitidos en cada una y su clasificación en las categorías 1 a la $J-1$. Se representará mediante las variables aleatorias $X_{i,j}^{(k)}$ el número de votos que obtendrá el candidato $j \in \{1,\ldots,J\}$ en el distrito $i \in \{1,\ldots,N\}$ y casilla $k \in \{1,\ldots,K_i\}$. A su vez, las variables aleatorias $X_{i,j} = \sum_{k=1}^{K_i} X_{i,j}^{(k)}$ representan el número total de votos que obtendrá el candidato $j$ en el distrito $i$, y las variables aleatorias $\Theta_{i,j} = X_{ij}/n_i$ representan la proporción de votos a favor del candidato $j$ en el distrito $i$. De lo anterior es inmediato que $\sum_{j=1}^J X_{i,j} = n_i$ y $\sum_{j=1}^J \Theta_{i,j} = 1$ para todo distrito $i$.

Con lo anterior es posible definir variables aleatorias $X_j$ como el número total de votos que obtiene el candidato $j$ y en consecuencia $X_j = \sum_{i=1}^N X_{i,j} = \sum_{i=1}^N n_i \Theta_{i,j}$ . La proporción total de votos que obtendrá el candidato $j$ se obtiene mediante las variables aleatorias:

$$\Theta_j \;=\; \frac{X_j}{n} \;=\; \sum_{i=1}^N \frac{n_i}{n}\,\Theta_{i,j}\,, \qquad j \in \{1,\ldots,J\} \tag{4}$$

de donde es inmediato que $\sum_{j=1}^J X_j = n$ y en consecuencia $\sum_{j=1}^J \Theta_j = 1$. La proporción de votos $\Theta_j$ es respecto a la lista nominal, pero también es de interés la proporción de votos respecto al total de votos emitidos en las casillas (esto es, sin considerar la $J-$ésima categoría ABST), misma que puede representarse mediante las variables aleatorias:

$$\Lambda_j = \frac{X_j}{\sum_{\ell=1}^{J-1} X_\ell} = \frac{\Theta_j}{\sum_{\ell=1}^{J-1} \Theta_\ell}, \qquad j \in \{1, \ldots, J-1\}, \tag{5}$$

y la proporción de *participación* en la elección se calcula como complemento del abstencionismo (ABST), esto es como $1 - \Theta_J$ .

Se realizará un *muestreo aleatorio estratificado* consistente en muestreo aleatorio simple de casillas por distrito (estrato) proporcional a la lista nominal de cada distrito. Esto es, si en el distrito $i$ hay $K_i$ casillas, de entre éstas se seleccionará al azar un subconjunto de $c_i$ casillas donde $c_i \ll K_i$ y de forma que si $c = \sum_{i=1}^{N} c_i$ representa el número total de casillas en la muestra estratificada entonces $c_i \approx \frac{n_i}{n} c$ .

# 3. Distribución de probabilidad multinomial

En virtud de que la *unidad de muestreo* para el conteo rápido es el resultado de la votación en cada casilla electoral, se propone representar los resultados de las distintas casillas mediante un vector aleatorio $(X_{i,1}^{(k)}, \ldots, X_{i,J}^{(k)})$ donde $i \in \{1, \ldots, N\}$ y $k \in \{1, \ldots, K_i\}$. Los resultados que se asientan en el acta de cada casilla corresponden a votos efectivamente emitidos, esto es a los valores que se observarán para la variables aleatorias $X_{i,1}^{(k)}, \ldots, X_{i,J-1}^{(k)}$, excluyendo la categoría $J$ (ABSTenciones) ya que esta última se deduce restando al tamaño de la lista nominal de la casilla ($n_{i,k}$) el total de la votación efectivamente emitida $\sum_{j=1}^{J-1} X_{i,j}^{(k)}$ de modo que $\sum_{j=1}^{J} X_{i,j}^{(k)} = n_{i,k}$. Esto último implica una dependencia entre las variables aleatorias $X_{i,j}^{(k)}$ ya que sin importar el resultado de cada casilla su suma siempre debe ser igual (teóricamente) al tamaño de la lista nominal de dicha casilla; así, por ejemplo, un valor muy elevado de alguna de ellas implica necesariamente valores modestos para las demás. Esto es, los valores de unas influyen en las demás porque el máximo número de votos obtenible por casilla ($n_{i,k}$) ya está predefinido (salvo en unas pocas casillas denominadas *especiales*).

La idea de aplicar un muestreo aleatorio estratificado radica en el beneficio inferencial que se obtiene por establecer o escoger una estratificación tal que en cada estrato sea razonable suponer cierta homogeneidad en estas subpoblaciones, en contraste con la heterogeneidad que se tendría al obtener muestras directamente de la población total sin estratificar, en deterioro de la precisión de las inferencias que se desea realizar, véase por ejemplo Särndal *et al.* (1992). Con lo anterior, el vector aleatorio $(X_{i,1}^{(k)}, \ldots, X_{i,J}^{(k)})$ puede asociarse a un *experimento aleatorio multinomial* consistente en $n_{i,k}$ repeticiones independientes (votantes en el distrito $i$ y casilla $k$) de un procedimiento que en cada repetición puede arrojar uno de entre $J$ resultados posibles con una probabilidad $0 < \theta_{i,j} < 1$, $j \in \{1, \ldots, J\}$, con $\sum_{j=1}^{J} \theta_{i,j} = 1$ (nótese que aquí se está aplicando el supuesto de homogeneidad por estrato, ya que las probabilidades teóricas $\theta_{i,j}$ se suponen razonablemente homogéneas para toda casilla $k$ dentro de un mismo distrito $i$), y por tanto las variables aleatorias $X_{i,j}^{(k)}$ representan el número de veces que se obtuvo el resultado $j$ en las $n_{i,k}$ repeticiones (número de votos a favor del candidato $j$). Bajo los supuestos anteriores, la función de masa de probabilidades conjunta del vector aleatorio $(X_{i,1}^{(k)}, \ldots, X_{i,J}^{(k)})$ para cada distrito $i$ y casilla $k$, véase por ejemplo Johnson *et al.*(1997), sería $\mathbf{Multinomial}(\theta_{i,1},\ldots,\theta_{i,J},n_{i,k})$ :

$$
\begin{aligned}
f_i^{(k)}(z_1,\ldots,z_J\,|\,\theta_{i,1},\ldots,\theta_{i,J},n_{i,k}) &= \mathbb{P}(X_{i,1}^{(k)}=z_1,\ldots,X_{i,J}^{(k)}=z_J) \\
&= \binom{n_{i,k}}{z_1,\ldots,z_J}\prod_{j=1}^{J}\theta_{i,j}^{z_j}\,\mathbf{1}_{\{\sum_{j=1}^{J}z_{i,j}=n_{i,k}\}}
\end{aligned} \tag{6}
$$

Con lo anterior, el vector aleatorio de votos por candidato en cada distrito se obtiene sumando los votos obtenidos por cada candidato en cada casilla del distrito:

$$
(X_{i,1},\ldots,X_{i,J}) = \Big(\sum_{k=1}^{K_i}X_{i,1}^{(k)},\ldots,\sum_{k=1}^{K_i}X_{i,J}^{(k)}\Big), \qquad i\in\{1,\ldots,N\} \tag{7}
$$

donde de acuerdo a Johnson *et al.*(1997) el vector aleatorio (7) tiene distribución de probabilidad $\mathbf{Multinomial}(\theta_{i,1},\ldots,\theta_{i,J},n_i)$ con $n_i=\sum_{k=1}^{K_i}n_{i,k}$ .

# 4. Inferencia bayesiana

En cada distrito (estrato) $i$ con $K_i$ casillas se obtendrá una muestra de $c_i \ll K_i$ casillas por muestreo aleatorio simple, información muestral que puede denotarse:

$$
\mathcal{M}_i = \{(x_{i,1}^{(k)},\ldots,x_{i,J}^{(k)}) : k=1,\ldots,c_i\}, \qquad i\in\{1,\ldots,N\}, \tag{8}
$$

a partir de lo cual puede obtenerse la votación total por candidato de cada distrito $i$, esto es $(x_{i,1},\ldots,x_{i,J})$ donde $x_{i,j}=\sum_{k=1}^{K_i}x_{i,j}^{(k)}$. Con lo anterior se desea realizar inferencias respecto a los vectores de parámetros $(\theta_{i,1},\ldots,\theta_{i,J})$, que desde un enfoque bayesiano requiere una densidad de probabilidades a posteriori como en (1):

$$
p(\theta_{i,1},\ldots,\theta_{i,J}\,|\,x_{i,1},\ldots,x_{i,J}) = \frac{1}{B}\,\pi(\theta_{i,1},\ldots,\theta_{i,J})f_i(x_{i,1},\ldots,x_{i,J}\,|\,\theta_{i,1},\ldots,\theta_{i,J})\,. \tag{9}
$$

Para una distribución de probabilidad Multinomial normalmente se utiliza una distribución a priori **Dirichlet**, véase por ejemplo Bernardo y Smith (1994), por dos razones esencialmente: primero, representa un vector de variables aleatorias cada una con soporte en el intervalo abierto $]0,1[$ y tales que su suma es siempre igual a 1, tal cual se requiere para los parámetros de la distribución Multinomial; segundo, es *familia conjugada* para la distribución de probabilidad Multinomial, esto es que la distribución a posteriori resulta ser también Dirichlet (aunque con parámetros actualizados de acuerdo a la información muestral):

$$
\pi(\theta_{i,1},\ldots,\theta_{i,J}) = \frac{\Gamma\big(\sum_{j=1}^{J}\alpha_{i,j}\big)}{\prod_{j=1}^{J}\Gamma(\alpha_{i,j})}\prod_{j=1}^{J}\theta_{i,j}^{\alpha_{i,j}-1}\,\mathbf{1}_{\{\sum_{j=1}^{J}\theta_{i,j}=1\}} \tag{10}
$$

donde los *hiperparámetros* $\alpha_{i,j}>0$ deben especificarse de acuerdo a *información a priori.* Como lo que se busca es una inferencia basada más en la información muestral que en información a priori puede optarse por utilizar $\alpha_{i,j}=1$ ya que en tal caso se obtiene una distribución de probabilidad *uniforme* sobre el conjunto $\{(\theta_{i,1},\ldots,\theta_{i,J})\in\,]0,1[^2 : \sum_{j=1}^{J}\theta_{i,j}=1\}$ :

$$
\pi(\theta_{i,1},\ldots,\theta_{i,J}) = (J-1)!\,\mathbf{1}_{\{\sum_{j=1}^{J}\theta_{i,j}=1\}}\,, \tag{11}
$$

y por lo tanto:

$$\pi(\theta_{i,1},\ldots,\theta_{i,J})f_i(x_{i,1},\ldots,x_{i,J}\,|\,\theta_{i,1},\ldots,\theta_{i,J}) \;\propto\; \prod_{j=1}^{J}\theta_{i,j}^{x_{i,j}+1-1}\,\mathbf{1}_{\{\sum_{j=1}^{J}\theta_{i,j}=1\}}\,, \tag{12}$$

de lo cual se deduce que la distribución de probabilidad a posteriori en cada distrito $i\in\{1,\ldots,N\}$ para $(\theta_{i,1},\ldots,\theta_{i,J})$ es $\mathbf{Dirichlet}(1+x_{i,1},\ldots,1+x_{i,J})$, donde la distribución marginal a posteriori para un candidato $j*$ del distrito $i$ resulta ser $\mathbf{Beta}(1+x_{i,j*}\,,\,\sum_{j=1}^{J}x_{i,j}-x_{i,j*}+J-1)$.

# 5. Estimación por intervalo

Con lo obtenido en la sección anterior se obtiene la distribución de probabilidad a posteriori para cada variable aleatoria $\Theta_{i,j}$, la proporción de votos a favor del candidato $j$ en el distrito $i$, y por tanto la proporción total de votos a favor del candidato $j$ en (4) es una combinación lineal convexa de la forma $\Theta_j=\sum_{i=1}^{N}\beta_i\Theta_{i,j}$ donde $\beta_i=n_i/n>0$ y $\sum_{i=1}^{N}\beta_i=1$. Definiendo $\mu_{i,j}:=\mathbb{E}(\Theta_{i,j})$ y $\sigma_{i,j}^2:=\mathbb{V}(\Theta_{i,j})$, por la propiedad de linealidad de la esperanza es inmediato obtener que:

$$\mu_j\,:=\,\mathbb{E}(\Theta_j)\,=\,\sum_{i=1}^{N}\beta_i\mu_{i,j}\,,\qquad j\in\{1,\ldots,J\}, \tag{13}$$

lo cual deja claro que la esperanza de $\Theta_j$ solo depende de las esperanzas marginales de las variables aleatorias $\{\Theta_{i,j}\,:\,i=1,\ldots,N\}$ sin importar la posible dependencia que pudiera existir entre ellas. Sin embargo, en el cálculo de la varianza de $\Theta_j$ sí se encuentra implicada dicha dependencia ya que aparecen covarianzas:

$$\begin{aligned}
\sigma_j^2:=\mathbb{V}(\Theta_j) \;&=\; \sum_{i=1}^{N}\beta_i^2\sigma_{i,j}^2\,+\,2\sum_{1\leq i<}\sum_{r\leq N}\beta_i\beta_r\mathbb{C}\mathrm{ov}(\Theta_{i,j},\Theta_{r,j}) &(14)\\
&=\; \sum_{i=1}^{N}\sum_{r=1}^{N}\beta_i\beta_r\mathbb{C}\mathrm{ov}(\Theta_{i,j},\Theta_{r,j})\\
&\leq\; \sum_{i=1}^{N}\sum_{r=1}^{N}\beta_i\beta_r\sigma_{i,j}\sigma_{r,j}=\left(\sum_{i=1}^{N}\beta_i\sigma_{i,j}\right)^2 &(15)
\end{aligned}$$

en donde la desigualdad (15) es consecuencia inmediata de la *Desigualdad de Cauchy-Schwarz,* ver por ejemplo Casella y Berger (2002), y proporciona el valor máximo posible de la varianza $\sigma_j^2$. Sea $I_j(\alpha)$ un intervalo de probabilidad $0<\alpha<1$ para $\Theta_j$, esto es tal que $\mathbb{P}[\Theta_j\in I_j(\alpha)]=\alpha$. Supongamos que para algún valor $z_\alpha$ dicho intervalo puede expresarse en la forma:

$$I_j(\alpha)\,=\,[\,\mu_j-z_\alpha\sigma_j\,,\,\mu_j+z_\alpha\sigma_j\,] \tag{16}$$

y lo análogo para $\Theta_{i,j}$ mediante $I_{i,j}(\alpha)=[\mu_{i,j}-z_\alpha\sigma_{i,j},\mu_{i,j}+z_\alpha\sigma_{i,j}]$. Si se construye un intervalo $I_j^*(\alpha)$ como combinación lineal convexa de los intervalos $I_{i,j}(\alpha)$ mediante las poderaciones $\beta_i=n_i/n$ se obtiene lo siguiente:

$$
\begin{aligned}
I_j^*(\alpha) &= \sum_{i=1}^{N} \beta_i I_{i,j}(\alpha) \\
&= \left[\, \sum_{i=1}^{N} \beta_i \mu_{i,j} - z_\alpha \sum_{i=1}^{N} \beta_i \sigma_{i,j} \,,\, \sum_{i=1}^{N} \beta_i \mu_{i,j} + z_\alpha \sum_{i=1}^{N} \beta_i \sigma_{i,j} \right] \\
&= [\, \mu_j - z_\alpha \sigma_j^* \,,\, \mu_j + z_\alpha \sigma_j^* \,] \qquad (17)
\end{aligned}
$$

donde la desviación estándar $\sigma_j^* = \sum_{i=1}^{N} \beta_i \sigma_{i,j}$ es el valor máximo de $\sigma_j$ de acuerdo a (15). En caso de que fuese válido un supuesto de independencia entre distritos entonces $\mathbb{C}\text{ov}(\Theta_{i,j}, \Theta_{r,j}) = 0$ y de acuerdo a (14) la desviación estándar resultante sería $\sigma_j^{\perp} = \sqrt{\sum_{i=1}^{N} \beta_i^2 \sigma_{i,j}^2}$, cantidad que al ser menor que $\sigma_j^*$ da origen a un intervalo de menor longitud $I_j^{\perp}(\alpha) = [\mu_j - z_\alpha \sigma_j^{\perp}, \mu_j + z_\alpha \sigma_j^{\perp}] \subset I_j^*(\alpha)$. Lo anterior deja en claro que si el supuesto de independencia entre distritos no es correcto entonces el intervalo resultante $I_j^{\perp}$ podría ser de menor longitud de la que debiera y, por lo tanto, de menor probabilidad al nivel $\alpha$ deseado. Pero utilizar el intervalo $I_j^*$ resultante de suponer covarianzas máximas entre distritos podría ser el otro extremo con una longitud mayor a la que debiera y por tanto con un nivel de probabilidad mucho mayor al deseado. Lo ideal sería obtener $I_j(\alpha)$ como en (16) con el valor preciso de $\sigma_j$ mediante (14) pero esto requería estimar $N(N-1)/2$ covarianzas distintas con base en una cantidad de información con la que no se cuenta. Suponiendo que el valor buscado $\sigma_j$ pertenece al intervalo $[\,\sigma_j^{\perp}\,,\, \sigma_j^*\,]$ entonces debe existir un valor $0 \leq \delta_j \leq 1$ tal que

$$
\sigma_j \,=\, (1-\delta_j)\sigma_j^{\perp} \,+\, \delta_j \sigma_j^* \qquad (18)
$$

y a partir de lo anterior el intervalo de probabilidad $\alpha$ para $\Theta_j$ sería:

$$
\begin{aligned}
I_j(\alpha) &= \mu_j \,\pm\, z_\alpha[(1-\delta_j)\sigma_j^{\perp} \,+\, \delta_j \sigma_j^*] \\
&= (1-\delta_j)\mu_j \,+\, \delta_j \mu_j \,\pm\, z_\alpha(1-\delta_j)\sigma_j^{\perp} \,\pm\, z_\alpha \delta_j \sigma_j^* \\
&= (1-\delta_j) I_j^{\perp}(\alpha) \,+\, \delta_j I_j^*(\alpha)\,. \qquad (19)
\end{aligned}
$$

En (19) persiste el problema de estimar $\delta_j$ pero ya solo serían $J$ valores (uno por candidato) además de que el valor $0 \leq \delta_j \leq 1$ nos proporcionaría una forma de medir del grado de dependencia entre distritos por candidato (con $\delta_j = 0$ representando independencia y $\delta_j = 1$ covarianzas máximas), mismo que al menos podría estimarse respecto a elecciones recientes, por medio del siguiente:

**Algoritmo 1**

1. Calcular los intervalos extremos $I_j^{\perp}(\alpha)$ e $I_j^*(\alpha)$ para cada $j \in \{1, \ldots, J\}$.
2. Simular una cantidad elevada de muestras estratificadas del tamaño que se haya determinado a partir del marco muestral de una elección pasada.
3. Con base en dichas simulaciones determinar cada valor $\delta_j$ de modo que el intervalo (19) efectivamente capture el verdadero valor $\theta_j$ en $100\alpha\,\%$ de las simulaciones del paso 2.

Se aplicó el algoritmo anterior a los resultados de las siguientes elecciones:

| Candidato | Gob2011 | Pre2012 | Dip2015 | *promedio* | *mediana* | *combinado* |
|---|---|---|---|---|---|---|
| Participación | 0.4638 | 0.3722 | 0.5609 | 0.4656 | 0.4638 | **0.4647** |
| NULOS | 0.2622 | 0.0976 | 0.0632 | 0.1410 | 0.0976 | **0.1193** |
| NOREG | 0.3000 | 0.2960 | 0.1534 | 0.2498 | 0.2960 | **0.2729** |
| PAN | 0.3521 | 0.4474 | 0.5683 | 0.4559 | 0.4474 | **0.4517** |
| PRI | 0.2248 | 0.2604 | 0.3345 | 0.2732 | 0.2604 | **0.2668** |
| PRD | 0.2987 | 0.3058 | 0.5329 | 0.3791 | 0.3058 | **0.3425** |
| MORENA | NA | NA | 0.3272 | 0.3272 | 0.3272 | **0.3272** |
| PT | 0.2987[a] | 0.3058[a] | 0.3430 | 0.3158 | 0.3058 | **0.3108** |
| MC | 0.2987[a] | 0.3058[a] | 0.5868 | 0.3971 | 0.3058 | **0.3515** |
| PANAL | 0.2248[b] | 0.0852 | 0.3856 | 0.2319 | 0.2248 | **0.2283** |
| PES | NA | NA | 0.2056 | 0.2056 | 0.2056 | **0.2056** |
| PH | NA | NA | 0.1482 | 0.1482 | 0.1482 | **0.1482** |
| Independientes | NA | NA | NA | NA | NA | **0.1769**[c] |

Cuadro 1: Valores de $\delta_j$ por candidato de acuerdo al Algoritmo 1. NA = no aplica. [a] Valor imputado por su alianza con PRD. [b] Valor imputado por su alianza con PRI. [c] Se imputó el promedio de los valores combinados de PES y PH por ser partidos pequeños de reciente creación y que posiblemente serían el referente más cercano a lo que podría resultar de candidatos independientes que por primera vez participan en una elección de gobernador del Estado de México en la historia reciente.

- Gobernador del Estado de México en el año 2011 (Gob2011),
- Presidente de México en el año 2012, restringiéndose estrictamente a la votación en el Estado de México (Pre2012),
- Diputados federales por el Estado de México en el año 2015 (Dip2015).

En los tres casos anteriores se mapearon los resultados a la estructura de distritos locales que se utilizará en la elección de gobernador del Estado de México en al año 2017, con el objeto de que los resultados sean comparables entre sí y útiles para la elección de 2017, obteniéndose los resultados que se resumen en el Cuadro 1 con muestras estratificadas de tamaño $c = 1,200$ proporcionales a la lista nominal por distrito y con $\alpha = 0.95$ (intervalos de probabilidad 95 %) para intervalos de proporción efectiva de votos de acuerdo a (5). La proporción de *participación* se calculó como 1 menos la proporción de abstencionismo (ABST).

En el Cuadro 1 el valor *combinado* es simplemente el punto medio entre el *promedio* y la *mediana*. Como puede apreciarse en todos los casos los valores de $\delta_j$ son positivos lo cual significa que un supuesto de independencia entre distritos (estratos) en dichas elecciones no era realista y que en dado caso hubiese conducido a subestimaciones, en mayor o menor grado, de las varianzas $\sigma_j^2$ y por tanto a la estimación con intervalos de probabilidad menor a la deseada de 95 %, como se resume en el Cuadro 2 donde se estiman las probabilidades de cobertura de intervalos de probabilidad 95 % bajo el supuesto de independencia entre distritos.

| Candidato | Gob2011 | Pre2012 | Dip2015 |
|---|---|---|---|
| Participación | 39.23 | 46.67 | 33.10 |
| NULOS | 65.03 | 81.17 | 82.60 |
| NOREG | 75.10 | 32.07 | 69.63 |
| PAN | 47.07 | 40.27 | 35.57 |
| PRI | 51.37 | 51.93 | 47.27 |
| PRD | 51.27 | 50.70 | 35.33 |
| MORENA | | | 48.87 |
| MC | | | 37.03 |
| PANAL | | 81.67 | 47.70 |
| PES | | | 64.13 |
| PH | | | 70.03 |

Cuadro 2: Probabilidades (en %) de que intervalos de 95 % de probabilidad hubiesen capturado el verdadero resultado de la elección (cobertura) bajo el supuesto de independencia entre distritos.

## Referencias

J.M. Bernardo y A.F.M Smith (1994) *Bayesian Theory.* Wiley (Chichester).

G. Casella y R.L. Berger (2002) *Statistical Inference.* Duxbury (Pacific Grove).

N.L. Johnson, S. Kotz y N. Balakrishnan (1997) *Discrete Multivariate Distributions.* Wiley (New York).

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

C.-E. Särndal, B. Swensson y J. Wretman (1992) *Model Assisted Survey Sampling.* Springer (New York).

**B. Propuesta de la Dra. Karla Beatriz Valenzuela Ocaña y el Dr. Alberto Alonso y Coria**

Para estimar el porcentaje de votación que corresponde a cada partido y el porcentaje de participación ciudadana en la elección, se utilizarán los procedimientos clásicos de estimación para un muestreo aleatorio estratificado.

Como se menciona en la sección de Diseño de Muestreo, se determinó utilizar como estratos la conformación distrital local, así se tienen 45 estratos a analizar.

**I.** Análisis por estrato**h**.

a. Estimación del número de votos por candidato

Como primer paso se estima el número de votos que obtendría cada una de las fuerzas políticas en el estrato **h**.

Sea $\boldsymbol{v_{i,j}}$la votación por el candidato **i** en la casilla **j** y sea $\boldsymbol{ln_j}$el listado nominal para la casilla **j**. El estimado total de participación para el candidato **i**para el estrato **h** analizado está dado por:

$$\widehat{\boldsymbol{R}}_i = \frac{\sum_{j=1}^{n} \boldsymbol{v_{i,j}}}{\sum_{j=1}^{n} \boldsymbol{ln_j}}$$

donde **n** es el número de casillas en la muestra seleccionadas del estrato **h.**

El cálculo de la varianza para el estimador $\widehat{\boldsymbol{R}}_i$se obtiene de la siguiente forma:

$$V(\widehat{R}_i) \cong \frac{1-f}{n\overline{LN}} \cdot \frac{\sum_{j=1}^{n}(v_{i,j} \; - \; \widehat{R}_i ln_j)^2}{n-1}$$

donde $\boldsymbol{f = \frac{n}{N}}$, **N** representa el número total de casillas en el estrato y $\overline{\boldsymbol{LN}}$el listado nominal promedio de las casillas en el estrato analizado.

El intervalo de confianza para la estimación del total de participación $\widehat{\boldsymbol{R}}_i$ para el candidato **i**para el estrato analizado está dado por:

$$\widehat{R}_i \pm z_{\alpha/2} * \sqrt{V(\widehat{R}_i)}$$

donde $\boldsymbol{1-\alpha}$es el nivel de confianza establecido. Para la estimación se ha definido el nivel de confianza como 95% por lo que $\boldsymbol{z_{\alpha/2} = 1.96}$

El total de votos estimado para el candidato **i** en el estrato **h** es simplemente $\widehat{\boldsymbol{R}}_i\boldsymbol{LN}$

donde **LN** es el listado nominal del estrato **h.**

b. Estimación de la participación de ciudadanos.

Es necesario estimar la participación de los ciudadanos en la elección para el estrato analizado, esto es, el número total de votos generados en el estrato.

Sea $\boldsymbol{v_{TV,j}}$el total de votos emitidos en la casilla **j** incluyendo votos nulos y votos de candidatos no registrados, tenemos que$\boldsymbol{\hat{R}_{TV}}$, el estimador de la participación ciudadana en el estrato analizado, es:

$$\boldsymbol{\hat{R}_{TV} = \frac{\sum_{j=1}^{n} v_{TV,j}}{\sum_{j=1}^{n} ln_j}}$$

Este procedimiento se repite para cada uno de los **h** estratos.

**II.** Integración de la información de los estratos.

Para obtener la estimación total para cada candidato, se procede a dar el peso correspondiente a cada estrato analizado. Así pues, se agrega la información correspondiente a cada estrato **h**, teniendo la estimación total de votos obtenidos por el candidato **i**:

$$\boldsymbol{v_{T,i} = \sum_{h}^{L} \hat{R}_{i,h} LN_h}$$

donde $\boldsymbol{LN_h}$ es la lista nominal del estrato **h** y $\boldsymbol{\hat{R}_{i,h}}$es la participación total para el partido **i** en el estrato **h** y **L**es el número total de estratos.

La estimación del total de votos emitidos en el estado se calcula de manera semejante, considerando el peso correspondiente a cada estrato. Así:

$$\boldsymbol{v_{T,total\ de\ votos} = \sum_{h} \hat{R}_{TV,h} LN_h}$$

Finalmente se tiene el intervalo de confianza para $\boldsymbol{P_i}$ como

$$\boldsymbol{P_i \pm z_{\alpha/2} * \sqrt{V_{T,i}}}$$

El análisis se realiza para cada candidato **i**.

**Referencias**

Cochran, W. G. (1977) Sampling Techniques. 3ed. Wiley&Sons, Inc.

**C. Propuesta del Dr. Carlos Erwin Rodríguez**

La estimación se realizará mediante el estimador derazón combinado y técnicas de remuestreo(*Bootstrap)*.

El Bootstrap es un método estadístico para estimar la distribución de muestreo de un estimador, se obtiene seleccionando $\boldsymbol{B}$ sub-muestras de la muestra original y calculando el estimador de interés con cada sub-muestra. Si se realiza un histograma con los $\boldsymbol{B}$ estimadores generados (en el caso de un estimador de una dimensión), se observaría la aproximación resultante. El número de réplicas $\boldsymbol{B}$, debe ser suficientemente grande para lograr una buena aproximación. Una vez obtenidas las réplicas Bootstrap es posible calcular los intervalos de confianza.

Existen muchas variantes del Boostrap, el trabajo de Efron (1979) sentó las bases, sin embargo, esta estrategia ha sido explorada por muchos autores. A continuación se describe el enfoque del Bootstrap que se implementará en el Conteo Rápido para la Elección para Gobernador en el Estado de México. En particular se usarán las ideas de Sitter (1992A) y (1992B). El primer artículo describe el método que se implementará el día de la elección, mientras que el segundo describe varias comparaciones, incluso con re-muestreo Jackknife. Estas ideas ya han sido aplicadas con éxito para realizar la estimación de la composición de la cámara de diputados en la Elección Federal de 2015, así como para la Elección Extraordinaria de Colima en 2016.

Las ideas de Sitter (1992A) aplicadas a la Elección de Gobernador en el Estado de México se describen a continuación. Sean$\boldsymbol{N_h}$el número total de casillas instaladas en el estrato $\boldsymbol{h}$ y $\boldsymbol{n_h}$ la muestra de casillas seleccionadas (o recuperadas) en el mismo estrato. Entonces, la probabilidad de selección de una casilla en el estrato $\boldsymbol{h}$ está dada por $\boldsymbol{f_h} = \frac{\boldsymbol{n_h}}{\boldsymbol{N_h}}$ .

*Esbozo del algoritmo*

Si $\boldsymbol{N_h} = \boldsymbol{n_h}\,\boldsymbol{k_h} \Leftrightarrow \boldsymbol{k_h} = \frac{\boldsymbol{N_h}}{\boldsymbol{n_h}} = \boldsymbol{1}/\boldsymbol{f_h}$y $\boldsymbol{m_h} = \boldsymbol{f_h}\boldsymbol{n_h}$(suponiendo que $\boldsymbol{k_h}$y $\boldsymbol{m_h}$son enteros), entonces

1. De la muestra recuperada del primer estrato (Distrito Local del Estado de México) se selecciona una sub-muestra SIN reemplazo de $\boldsymbol{m_1}$casillas (de las $\boldsymbol{n_1}$ recuperadas).
2. Se repite el paso 1, $\boldsymbol{k_1}$veces, obteniendo una muestra de $\boldsymbol{n_1} = \boldsymbol{m_1 k_1}$casillas en total.
3. Se calcularán las siguientes estadísticas:
   - ✓ Estimador común del total de votos en el primer estrato para el i-ésimo candidato a gobernador

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{1i}} = \boldsymbol{N_1}\overline{\boldsymbol{y}}_{\boldsymbol{1i}}$$

4. Se repiten los pasos 1 a 3 para cada uno de los $\boldsymbol{L}$ estratos restantes.
5. Se calculan
   - ✓ El estimador del total de votos a favor del candidato a gobernador del partido i-ésimo en todo el estado (sumando sobre todos los estratos)

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} = \sum_{\boldsymbol{h=1}}^{\boldsymbol{L}} \widehat{\boldsymbol{Y}}_{\boldsymbol{hi}}$$

   - ✓ Se estiman los porcentajes mediante la siguiente expresión

$$\widehat{\boldsymbol{p}}_{\boldsymbol{i}} = {}^{\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}}\!\Big/_{\sum \widehat{\boldsymbol{Y}}_{\boldsymbol{i}}}$$

El paso 5 concluye una iteración Boostrap, el objetivo es realizar $\boldsymbol{B}$ iteraciones con $\boldsymbol{B}$ suficientemente grande. Para el 4 de junio de 2017 y con cada remesa de información que se reciba se tiene pensado realizar $\boldsymbol{B} = \mathbf{5,000}$iteraciones.

Es importante observar que en el paso 1, en el muestreo SIN reemplazo, la probabilidad de selección es la misma que en el diseño de muestreo original $\boldsymbol{f_h}' = \frac{\boldsymbol{m_h}}{\boldsymbol{n_h}} = \boldsymbol{f_h}$, esto se repite $\boldsymbol{k_h}$ veces de manera independiente para obtener una muestra de $\boldsymbol{n_h}$casillas. Esta estrategia asegura obtener estimaciones in-sesgadas para la varianza y para el tercer momento.

En caso de que $\boldsymbol{k_h}$o $\boldsymbol{m_h}$no sean enteros se sigue un proceso de aleatorización para alcanzar la consistencia del algoritmo.

El algoritmo anterior es computacionalmente demandante. Sin embargo, las iteraciones del Bootstrap son independientes, así que es posible correr el algoritmo mediante procesamiento paralelo.

**Referencias**


Efron, B. (1979). "Bootstrap methods: Another look at the jackknife". The Annals of Statistics, Vol 7. No 1, pp. 1-26.

Sitter, R. R. (1992A). “A Resampling algorithm for complex survey data”. Journal of the American Statistical Association. Vol. 87. No. 419, pp. 755-765.

Sitter, R. R. (1992B). “Comparing Three Bootstrap Methods for Survey Data”. The Canadian Journal of Statistics. Vol. 20. No. 2, pp 135-154.


## 1.5 Jornada Electoral

El Sistema de Captura del Conteo Rápido (SICCORA) generará cortes de información cada 10 minutos. Una vez que se tenga información para cada uno de los estratos, es decir al menos 2 casillas por Distrito Local, se realizarán estimaciones con cada una de los tres métodos propuestos para monitorear que éstas vayan en el mismo sentido y que las estimaciones sean consistentes.

Considerando tanto el avance en la cobertura de la muestra recibida por estrato, como el seguimiento de las estimaciones, el COTECORA determinará el momento oportuno en el que esté en posibilidades de pronunciar sus estimaciones finales en forma de intervalos de confianza. En este momento se generará un informe con formato previamente definido, mismo que se imprimirá y será firmado por cada uno de los miembros del COTECORA para hacerlo llegar al Presidente del Consejo General del IEEM.

A partir de las 21:00 horas del día de la jornada electoral, el COTECORA rendirá un informe de avance de la integración de la muestra al Consejo General del IEEM. Este informe se actualizará como máximo cada hora y hasta la entrega de los resultados finales de la estimación.

Sea cual fuere la muestra recabada y los resultados obtenidos, y a más tardar a las 11:30 pm del domingo 4 de junio de 2017, el COTECORA entregará al Presidente del Consejo General del IEEM un informe para su inmediata difusión.

# III. Selección y resguardo de la muestra

## 1.1 Consideraciones generales para la selección de la muestra

La muestra con la que se estimará la votación a favor de los candidatos a la gubernatura del Estado de México se realizará en un acto público el viernes 2 de junio de 2017 a las 10:00 horas. En el acto estará presente un notario público que dará fe del desarrollo del protocolo desde la instalación del software requerido hasta la obtención y resguardo de la muestra definitiva.

Para seleccionar la muestra se hará uso de un equipo de cómputo debidamente blindado y habilitado con el software estadístico. El marco muestral será el listado final de las casillas aprobado para las elecciones y la selección se realizará de acuerdo con el diseño de muestra establecido anteriormente.

## 1.2 Protocolo de selección y resguardo de la muestra

Para la selección y resguardo de la muestra se realizarán las siguientes actividades:

**Instalación**

1. Personal de la Unidad de Informática y Estadística (UIE) del IEEM realizará ante notario público la validación del equipo de cómputo donde se instalará el programa para la obtención de la muestra.
2. El COTECORA entregará un disco compacto con el software necesario para la selección de la muestra (disco de insumos) junto con sus códigos de integridad (programa para la selección de la muestra y la base de datos de casillas así como sus correspondientes códigos de integridad).
3. Personal de la UIE obtendrá los códigos de integridad los archivos entregados por el COTECORA y los comparará con los entregados. El Notario Público validará que se trata del mismo código de integridad.
4. A la vista de los asistentes se instalarán el software R y R Studio.
5. Se solicitará a un miembro del COTECORA iniciar la etapa de ejecución de la selección de la muestra.

**Selección de la muestra**

1. Para seleccionar la muestra se requiere de un número aleatorio denominado semilla.
2. La semilla se construirá con tres números de seis dígitos.
3. Para construir y capturar los tres números se requiere la participación de seis personas elegidas entre los asistentes. Tres, anotarán un número en un formato diseñado para tal efecto y lo ingresarán en el programa. Para asegurar el correcto ingreso, el programa solicitará la confirmación de los números, los cuales serán nuevamente ingresados por las tres personas restantes. Esto permitirá garantizar que el número semilla generado sea igual al ingresada en la computadora.
4. A la vista del Notario Público, los números se ingresarán en el programa de selección de la muestra. Los demás asistentes no conocerán estos números.
5. Una vez ingresados los números para construir la semilla, un miembro del COTECORA ejecutará el programa para seleccionar la muestra. Esta última quedará grabada en el disco duro de la computadora.
6. Se generará un código de integridad de la muestra. El código de integridad será impreso y se entregará al Notario Público y a los asistentes que lo soliciten.

**Resguardo y entrega de la muestra**

1. La muestra será grabada en dos discos compactos no regrabables.
2. En forma inmediata y en presencia de Notario Público se enviarán electrónicamente los cuarenta archivos cifrados a las cuarenta juntas distritales del INE.
3. El disco de insumos junto con uno de los discos con la muestra, los códigos de integridad y los formatos en los que se anotaron los números para construir la semilla serán guardados en un sobre sellado.
4. El segundo disco con la muestra será guardado en un sobre distinto debidamente etiquetado.
5. Asimismo, el equipo de cómputo en el cual fue generada la muestra se resguardará en un sobre.
6. Todos los sobres serán sellados y rubricados por el Notario Público y por los asistentes que así lo deseen.
7. El Notario Público entregará al Secretario Ejecutivo los sobres a los que se refiere el numeral anterior para su resguardo.
8. El disco con la muestra referido en el punto 4será entregado al titular de la UIE del IEEM el sábado 3 de junio a las 12:00 horas, con el fin de que se preparen

los trabajos de acopio de datos de las actas de escrutinio y cómputo el día de la jornada electoral.

## IV. Procedimiento logístico operativo

*Esquema general*

El IEEM, en coordinación con el COTECORA y el INE, elaborará una "guía de procedimientos de la operación logística" para el Conteo Rápido.

Para poder brindar información oportuna sobre las tendencias de votación para la elección de Gobernador en el Estado de México, se requiere planear y ejecutar una serie coordinada de actividades que tendrán como base el siguiente procedimiento general que tendrá verificativo el día de la jornada electoral:

1) El personal en campo, CAE o SE, se encargará de recabar en el formato diseñado para tal fin, los datos de votación contenidos en las actas de escrutinio y cómputo de la(s) casilla(s) seleccionada(s) de la muestra que se hayan asignado a su Área de Responsabilidad Electoral (ARE). De manera inmediata llamará, a través del medio de comunicación que le fue asignado, al Centro de Captura del IEEM para reportar la información recopilada.

2) En el Centro de Captura del IEEM, los/las capturistas de cómputo recibirán las llamadas y capturarán directamente en el Sistema de Captura del Conteo Rápido (SICCORA) los datos que le sean comunicados, para su transferencia inmediata a la sede del COTECORA.

3) Los integrantes del COTECORA procesarán la información proporcionada por el sistema y realizarán las estimaciones estadísticas correspondientes. A partir de ello, elaborarán un informe sobre los resultados obtenidos y lo enviarán al presidente del Consejo General del IEEM para su difusión.

4) El presidente del Consejo General del IEEM dará a conocer a la opinión pública, la noche del 4 de junio de 2017, los resultados del Conteo Rápido.

# V. Conclusiones

El objetivo principal del conteo rápido para la elección de gobernador en el Estado de México es abonar a la certidumbre, confianza y transparencia en el proceso electoral, ofreciendo una estimación estadísticamente confiable de las tendencias de votación en la noche misma de la jornada electoral.

La muestra a utilizar en las estimaciones del Conteo Rápido será de 1818 casillas, con un diseño estratificado basado en los 45 distritos locales del Estado de México.

La muestra se generará en un acto protocolario y público el día viernes 2 de junio por la mañana.

Para estimar el porcentaje de participación electoral y el porcentaje de votos de cada aspirante a la gubernatura del Estado de México se utilizarán tres métodos estadísticos:

- Método Bayesiano
- Método Clásico vía las fórmulas para estimadores de razón, derivadas mediante las técnicas de muestreo probabilístico.
- Método Clásico vía métodos de remuestreo.

Cada estimación realizada por el COTECORA estará dada en forma de intervalos de confianza del 95%. Las tres estimaciones se combinarán para producir una estimación final la noche del 4 de junio de 2017.

Al día siguiente de la jornada electoral, el COTECORA entregará al IEEM los siguientes documentos para su publicación y difusión en internet:

a) El protocolo de selección de la muestra.
b) Las fórmulas de cálculo utilizadas para cada método establecido.
c) El reporte de los resultados estimados por el conteo rápido.
d) La base de datos utilizada en las estimaciones de los conteos rápidos, incluyendo la siguiente información:
    I. Lista de casillas que fueron seleccionadas en la muestra, y
    II. Lista de casillas que se integraron en el cálculo final, cada una con el resultado de la elección del 4 de junio de 2017.
e) Un informe en lenguaje sencillo con el objetivo de facilitar la comprensión y utilidad de la realización del conteo rápido y sus resultados en la elección de gobernador del Estado de México.