Modelos de regresión multinivel

para el conteo rápido de las elecciones 2018

Michelle Anzarut, Felipe González, Teresa Ortiz

2018/10/02

Idea general

Objetivo: Estimar resultados de la elección a partir de una muestra aleatoria de casillas.

Inferencia basada en modelos: Modelo de regresión multinivel, en función de covariables asociadas a las casillas, que estima el número de votos a favor del candidato en cada casilla.

Calibración: Evaluamos modelos con remuestreo y datos de elecciones pasadas, revisando cobertura de intervalos de confianza/credibilidad.

Experiencia: Mostramos resultados de elección 2018.

Contexto: Conteo rápido

- En México, las elecciones tienen lugar un domingo, los resultados oficiales del proceso se presentan a la población una semana después. A fin de evitar proclamaciones de victoria injustificadas durante ese período el INE organiza un conteo rápido.
- **Conteo rápido**: Procedimiento para estimar el porcentaje de votos a favor de los candidatos en el día de la elección, consiste en seleccionar una muestra aleatoria de las casillas de votación y analizar sus resultados para presentar intervalos con una probabilidad de al menos 0.95.
- La rapidez y precisión del conteo favorece un ambiente de confianza y sirve como una herramienta contra el fraude.

Documento del conteo rápido de Guanajuato 2018

INE
Instituto Nacional Electoral

INE
Instituto Nacional Electoral
Firmado digitalmente por Instituto Nacional Electoral
Fecha: 2018.07.01 22:08:36 -05'00'

Conteos Rápidos 2018

Estimación de los resultados de la votación para la elección de la Gubernatura del Estado de Guanajuato 2018

Con los datos recibidos a las 21:45 hora del centro, del 1 de julio de 2018, el Comité Técnico Asesor informa lo siguiente:

1. De las 500 casillas que integran la muestra, se recibió información de 357 casillas, las cuales representan el 71.4 % de la muestra total.
2. De los 22 estratos considerados para definir el diseño muestral, se contó con información de 22 estratos.

Con la información recibida y con un nivel de confianza de al menos 95 % se estima lo siguiente:

3. La participación ciudadana se encuentra entre 51.8 % y 54.6 %.
4. El porcentaje de votos para cada candidatura a la Gubernatura del estado de Guanajuato se presenta a continuación:

NOMBRE	PARTIDO / COALICIÓN	INTERVALO % LÍMITE INFERIOR	INTERVALO % LÍMITE SUPERIOR
Diego Sinhue Rodríguez Vallejo	PAN_PRD_MC	49.5	51.5
Gerardo Sánchez García	PRI	11.5	12.9
Felipe Arturo Camarena García	PVEM	6.2	7.2
Francisco Ricardo Sheffield Padilla	MORENA_PT_ES	23.2	25.2
María Bertha Solórzano Lujano	NA	2.6	3.5

Atentamente
Comité Técnico Asesor del Conteo Rápido

Dr. Alberto Alonso y Coria
Dra. Michelle Anzarut Chacalo
Dr. Carlos Hernández Garciadiego
Dr. Manuel Mendoza Ramírez
Dr. Luis Enrique Nieto Barajas
Dr. Gabriel Núñez Antonio
Dr. Carlos Erwin Rodríguez Hernández-Vela
Mtra. Patricia Isabel Romero Mares
Dr. Raúl Rueda Díaz del Campo

1 de 1

Elecciones 2018

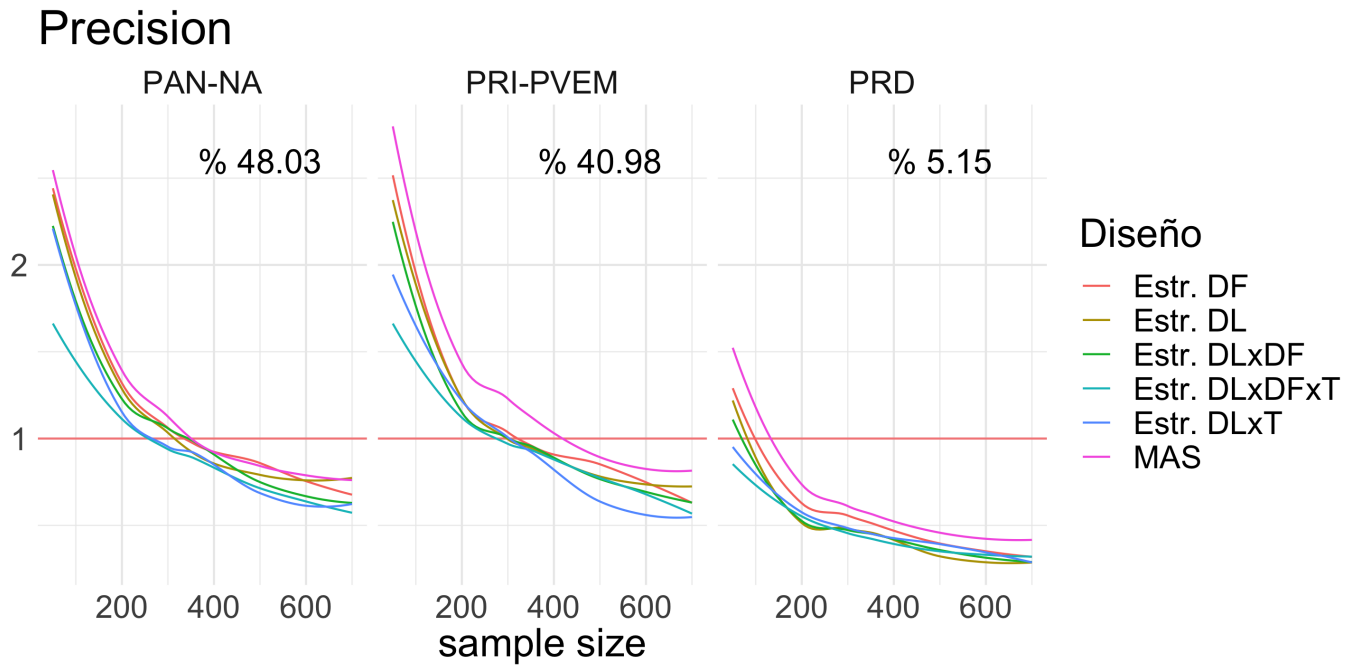
- La elección del 2018 fue la más grande que se ha vivido en México, con 3,400 puestos en disputa. Se realizaron conteos rápidos para **9 elecciones estatales** simultáneas a un conteo rápido para la **elección de presidente**.
- El día de la elección, el sistema de información comienza a las 6 p.m. y, cada 5 minutos, produce una secuencia de archivos acumulativos. Estas muestras parciales se analizan con los métodos de estimación para rastrear la tendencia de los resultados.
- Las muestras parciales tienen un sesgo potencial asociado al patrón de llegada de la información.

Diseño de la muestra

El diseño de la muestra es probabilístico.

- **Diseño:** es unietápico y estratificado, donde la unidad de observación es la casilla.
    - En Guanajuato son 22 estratos definidos por distrito local.
- **Tamaño de muestra:** Se eligió el tamaño de muestra para lograr intervalos de 95% confianza con **longitud máxima de 2 puntos porcentuales.**
    - En Guanajuato el tamaño de muestra se fijó en 500 casillas.
- **Selección de la muestra:** La distribución de la muestra en las casillas se realizó proporcional al número de casillas en cada estrato, y se utilizó muestreo aleatorio simple dentro de cada estrato.
- **Consideraciones adicionales:** Se busco que todos los estratos tuvieran al menos 50 casillas, y que porcentaje de CAEs encargados de más de una casilla fuera menor a 20%.

Diseño de la muestra: Guanajuato 2018



- Tras considerar distintas alternativas de estratificación se decidió utilizar la distritación electoral local.
- Dando lugar a 22 estratos, con un promedio de 300 casillas cada uno.

Datos faltantes

En la práctica la muestra seleccionada no llega completa. Entre las posibles razones de faltantes están:

- El clima en ciertas regiones dificulta la transmisión de los resultados.
- El responsable de reportar los resultados está saturado de trabajo: contando votos a falta de funcionarios de casilla, retrasado por la dificultad de llegar a la casilla por malas condiciones de terreno,...

Usualmente los faltantes **no son completamente aleatorios**, esto es, la probabilidad de que una casilla no se reporte está asociada a la respuesta de la casilla.

Buscamos un modelo con **tratamiento consistente de datos faltantes:** en ausencia de respuesta.

Antecedentes

- **Manuel Mendoza, Luis E. Nieto-Barajas, 2016.** *Quick counts in the Mexican presidential elections: A Bayesian approach*.
  - Se ajusta un modelo de manera independiente **para cada candidato** en **cada estrato**.
  - Modelo normal para el total de votos X_k que recibe cada candidato en la k-ésima casilla.

$$X_k \sim N\Big(n_k\theta, \tau/n_k\Big)$$

  - θ: proporción de gente de la lista nominal a favor del candidato.
  - n_k: número de personas en la lista nominal de la k-ésima casilla.

- **Roderick Little, 2012**. *Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics*.

- **David K Park, Andrew Gelman, and Joseph Bafumi, 2004**. *Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls*.

Inferencia en muestreo

1. **Inferencia basada en diseño de muestras.** Los valores poblacionales son una cantidad fija, la inferencia se basa en la distribución de probabilidad asociada a la selección de la muestra.

2. **Inferencia basada en modelos.** Las variables provienen de un modelo estadístico:

    - **Modelos de superpoblaciones:** los valores poblacionales se consideran una muestra aleatoria de una **superpoblación**, y se les asigna una distribución de probabilidad.

    - **Modelos bayesianos:** agregamos distribuciones iniciales a los parámetros y se hace inferencia de las **cantidades poblacionales** y de los **parámetros** usando la distribución posterior.

Modelos bayesianos

Predecimos la variable de interés para las unidades de la población que **no se incluyeron en la muestra** y para las unidades **que no respondieron**, condicional a la muestra observada y al modelo.

Usamos modelos paramétricos:

$$p(y|x) = \int p(y|x,\theta)p(\theta|x)p(\theta)d\theta$$

- $p(y|x,\theta)$: modelo paramétrico de y.
- $p(\theta|x)$: distribución inicial para θ.

Se incluyen en el modelo las variables involucradas en el diseño de la muestra (estratos, conglomerados).

Modelos bayesianos

1. La inferencia de θ se obtiene de la distribución posterior:

$$p(\theta|y_{obs}, x) \propto p(\theta|x)L(\theta|y_{obs}, x)$$

2. La posterior de θ lleva a inferencia de las cantidades poblacionales no observadas usando la distribución predictiva posterior:

$$p(y_{falta}|y_{obs}, x) = \int p(y_{falta}|\theta, x)p(\theta|y_{obs}, x)d\theta$$

3. Utilizamos los datos observados y simulaciones de los *datos faltantes* para inferir cantidades poblacionales de interés.

Covariables

1. Tipo de sección (rural o urbana/mixta).
2. Tipo de casilla (básica/contigua/especial o extraordinaria).
3. Tamaño de sección (chica < 1000 votantes, mediana [1000, 5000], grande > 5000).
4. Región (oriente u occidente).
5. Distrito local.
6. Interacción de tipo de sección con tamaño de sección.

Modelo con distribución normal

Sea X_k el número de votos en favor del candidato en la k-ésima casilla:

- Nivel 1

$$X_k \sim \mathsf{N}(n_k \theta_k, n_k^{-1} \tau_k^{\text{distrito}})\, \mathcal{I}_{[0,750]},$$

donde n_k es la lista nominal y θ_k la proporción de personas en la lista nominal de la casilla k que votaron por el candidato,

$$\begin{aligned}\theta_k = \text{logit}^{-1}(&\beta^0 + \beta^{\text{rural}} \cdot \text{rural}_k + \beta^{\text{rural-tamañoM}} \cdot \text{rural}_k \cdot \text{tamañoM}_k \\ &+\beta^{\text{tamañoM}} \cdot \text{tamañoM}_k + \beta^{\text{tamañoL}} \cdot \text{tamañoL}_k + \beta^{\text{tipoSP}} \cdot \text{tipoSP}_k \\ &+\beta^{\text{distrito}}_{\text{distrito}(k)}),\end{aligned}$$

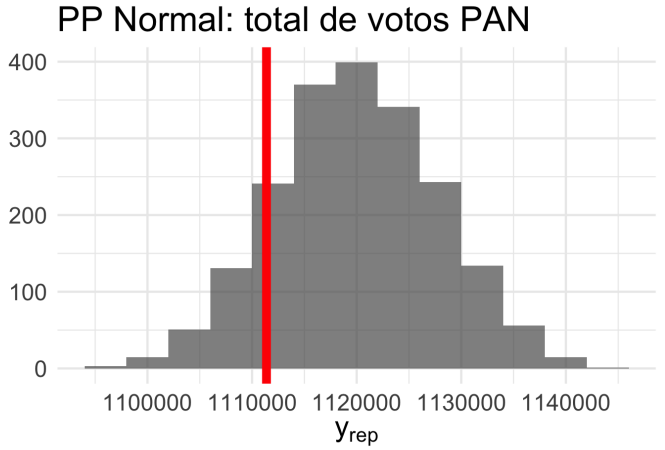
- Nivel 2

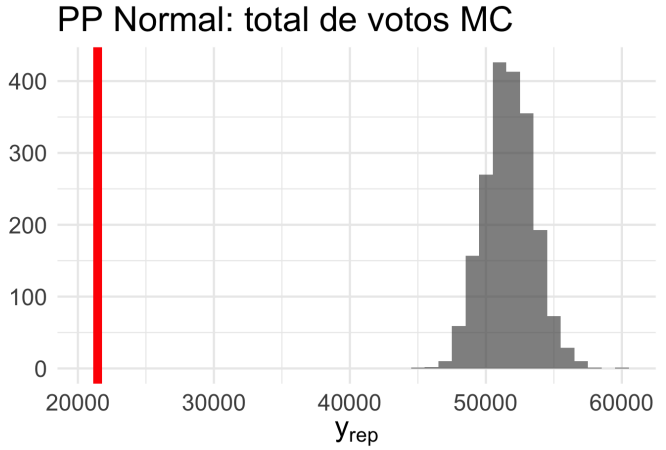
$$\beta_j^{\text{distrito}} \sim \mathsf{N}\left(\mu^{\text{distrito}}, \sigma^2_{\text{distrito}}\right).$$

Evaluación de ajuste

A total

La siguientes gráficas muestra la distribución predictiva posterior del total de votos para el PAN (partido ganador) y para Movimiento Ciudadano (partido chico).

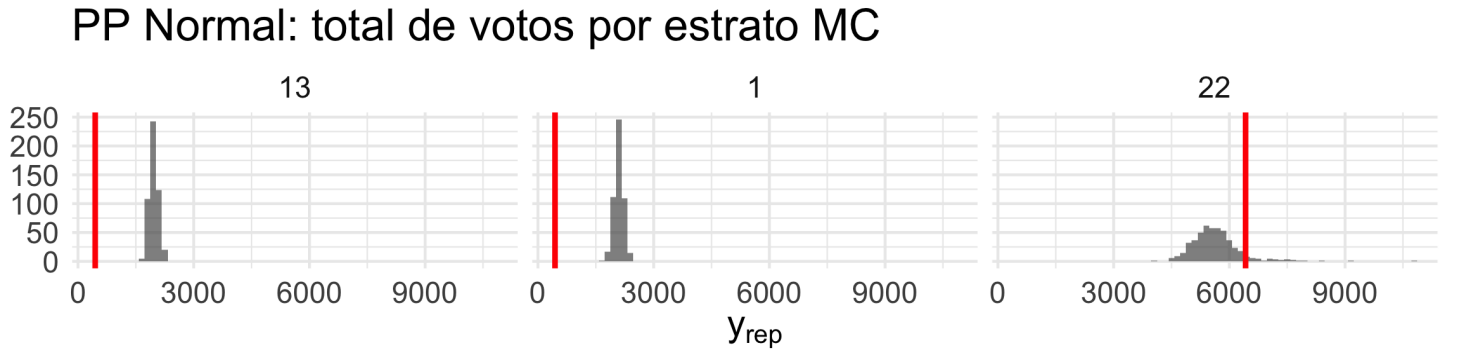




La línea roja indica el total de votos observado.

Por estrato

Examinamos otro nivel de desagregación: las distribuciones predictivas posteriores para el total de votos por estrato, mostramos las gráficas para 3 estratos.



Modelo con distribución mezcla

- Nivel 1

$$X_k \sim p_k \delta_0(x) + (1 - p_k)\mathsf{t}(n_k\theta_k, n_k^{-1}\tau_k^{\text{distrito}}, \nu_k^{\text{distrito}})\mathcal{I}_{[0,750]},$$

$$\begin{aligned}\theta_k = \text{logit}^{-1}(&\beta^0 + \beta^{\text{rural}} \cdot \text{rural}_k + \beta^{\text{rural-tamañoM}} \cdot \text{rural}_k \cdot \text{tamañoM}_k \\ &+\beta^{\text{tamañoM}} \cdot \text{tamañoM}_k + \beta^{\text{tamañoL}} \cdot \text{tamañoL}_k + \beta^{\text{distrito}}_{\text{distrito}(k)} \\ &+\beta^{\text{tipoSP}} \cdot \text{tipoSP}_k),\end{aligned}$$

$$\begin{aligned}p_k = \text{logit}^{-1}(&\beta_p^0 + \beta_p^{\text{rural}} \cdot \text{rural}_k + \beta_p^{\text{rural-tamañoM}} \cdot \text{rural}_k \cdot \text{tamañoM}_k \\ &+\beta_p^{\text{tamañoM}} \cdot \text{tamañoM}_k + \beta_p^{\text{tamañoL}} \cdot \text{tamañoL}_k + \beta^{\text{distrito-p}}_{\text{distrito}(k)} \\ &+\beta_p^{\text{tipoSP}} \cdot \text{tipoSP}_k).\end{aligned}$$

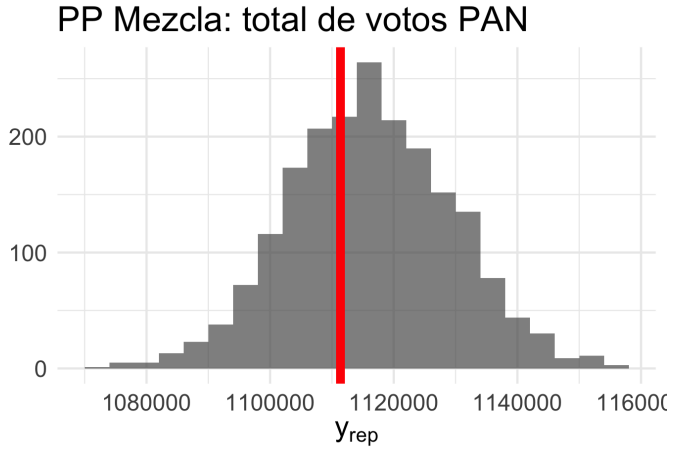
- Nivel 2

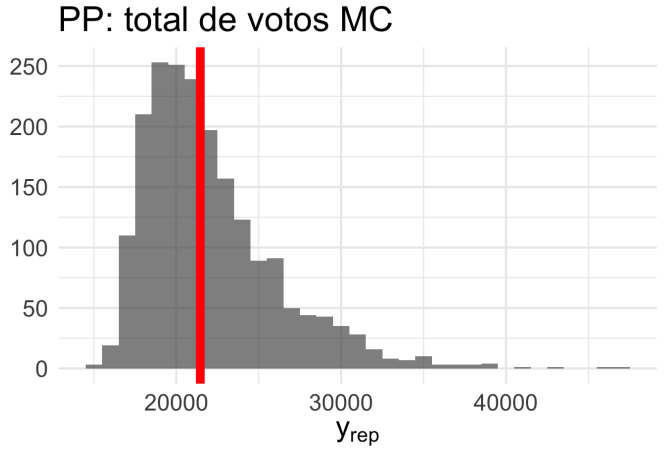
$$\beta_j^{\text{distrito}} \sim \mathsf{N}\left(\mu^{\text{distrito}}, \sigma^2_{\text{distrito}}\right).$$

Evaluación de ajuste

A total

La siguientes gráficas muestra la distribución predictiva posterior del total de votos para el PAN (partido ganador) y para Movimiento Ciudadano (partido chico).

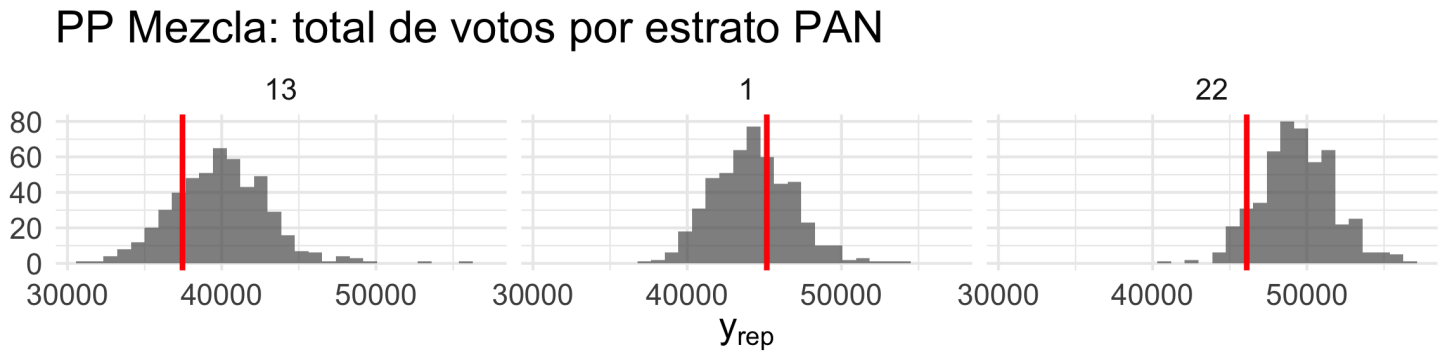


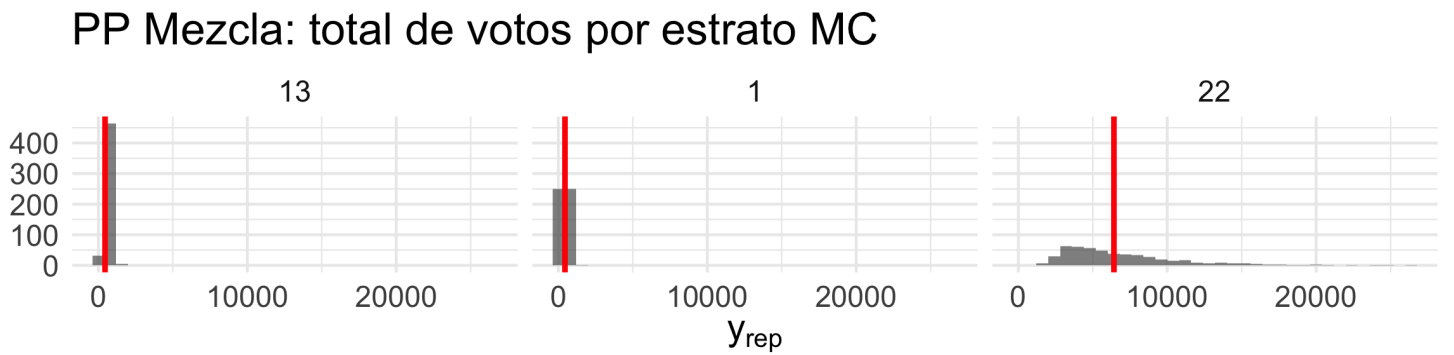


La línea roja indica el total de votos observado.

Por estrato

Mostramos las gráficas para los mismos 3 estratos que en el caso normal.





Calibración

Metodología:

1. Simulamos n muestras.
2. Para cada muestra creamos intervalos de 95% de probabilidad.
3. Revisamos el porcentaje de intervalos que contienen el valor observado.

Simulamos bajo los siguientes escenarios:

- Muestras completas.
- Censuramos las muestras completas usando patrones observados de la llegada de datos de cada distrito y ámbito (rural/urbano).
- Censuramos las muestras completas eliminando estratos.

Calibración

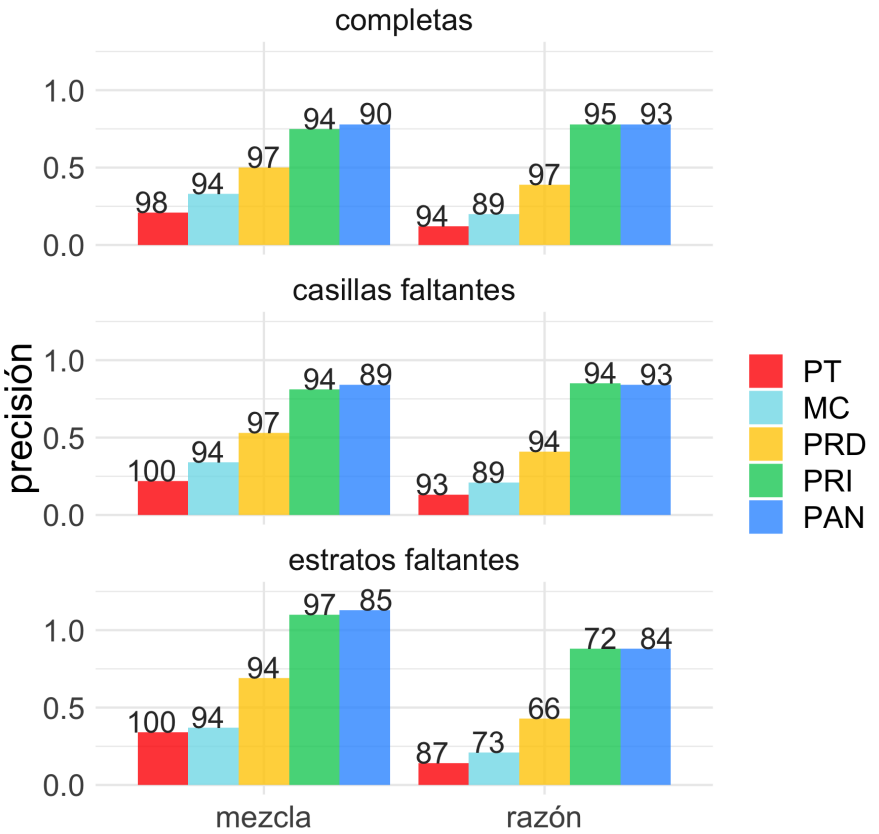
Estimador de razón combinado

Comparamos el desempeño del modelo a lo largo de las muestras simuladas con el estimador de razón combinado:

$$\hat{p_k} = \frac{\hat{X}_k}{\hat{Y}} = \frac{\sum_h \hat{X_{kh}}}{\sum_h \hat{Y}_h} = \frac{\sum_h \frac{N_h}{n_h} \sum_i X_{khi}}{\sum_h \frac{N_h}{n_h} \sum_i Y_{hi}}$$

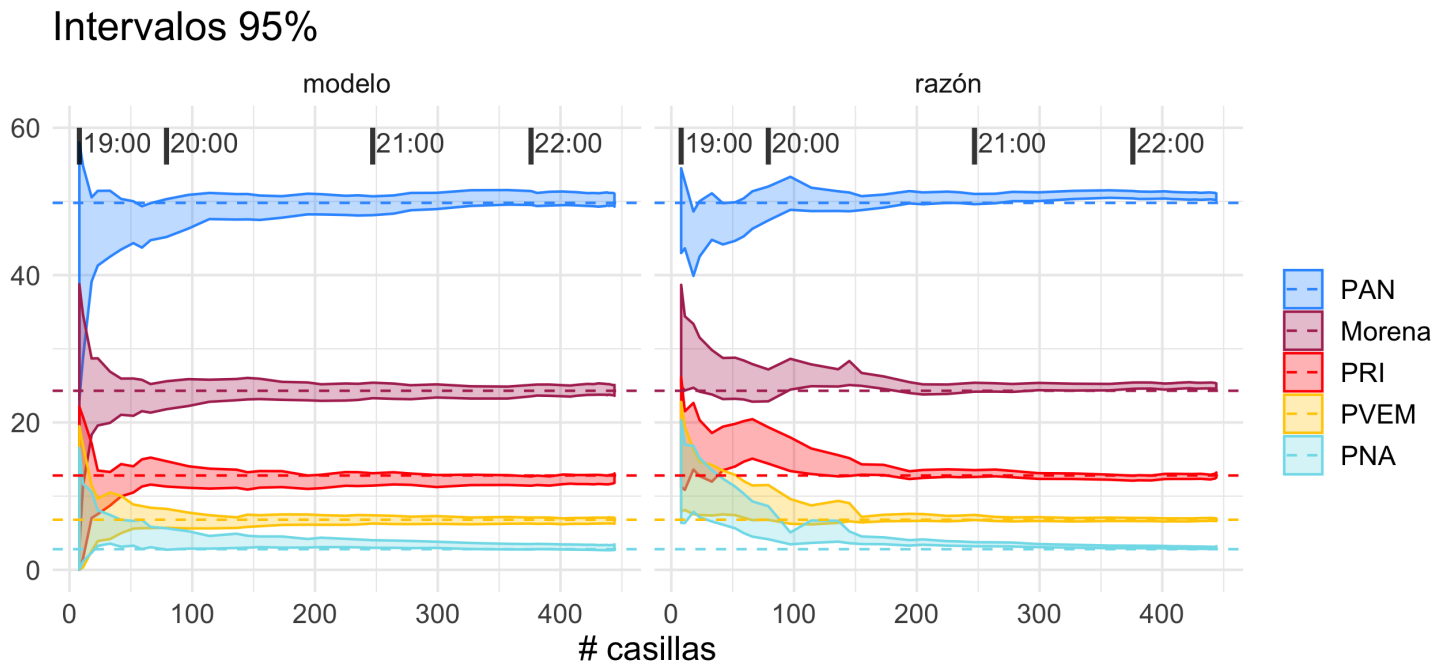
- Utilizamos bootstrap para estimar los errores estándar.
- En el caso de estratos faltantes se debe seleccionar una estrategia para utilizar este estimador.

Calibración



Elección Guanajuato 2018

Se reportaron los intervalos de probabilidad de las 9:45 pm, con 357 casillas.



Implementación

Implementamos en *JAGS*, la estimación se puede consultar y **reproducir completamente** con el paquete de R quickcountmx (Ortiz 2018).

- La reproducibilidad es crucial para examinar la **veracidad de las conclusiones** de un trabajo científico.
- La reproducibilidad ayuda a lograr la **transparencia en el procedimiento electoral**, fomenta la **confianza en las instituciones** y da **legitimidad al resultado del conteo rápido**.

Modelo nacional

El modelo multinivel con distribución de probabilidad mezcla resulta *muy lento* para la elección nacional.

- **División de datos:** Se estima un modelo de forma independiente para cada una de 7 regiones geográficas lo que nos permite paralelizar, pero no podemos usar información entre las regiones.
- Se modela utilizando una distribucióin **binomial negativa** (un parámetro menos).
- Se implementó con **Stan** en lugar de **JAGS** (el código está en el paquete de R quickcountmx).

Modelo nacional

Para cada región y para cada candidato:

- Nivel 1

$$X_k \sim \mathsf{NB}(n_k\theta_k, n_k\theta_k\nu_k^{\text{distrito}})\mathcal{I}_{[0,750]},$$

donde n_k es la lista nominal y θ_k la proporción de personas en la lista nominal de la casilla k que votaron por el candidato,

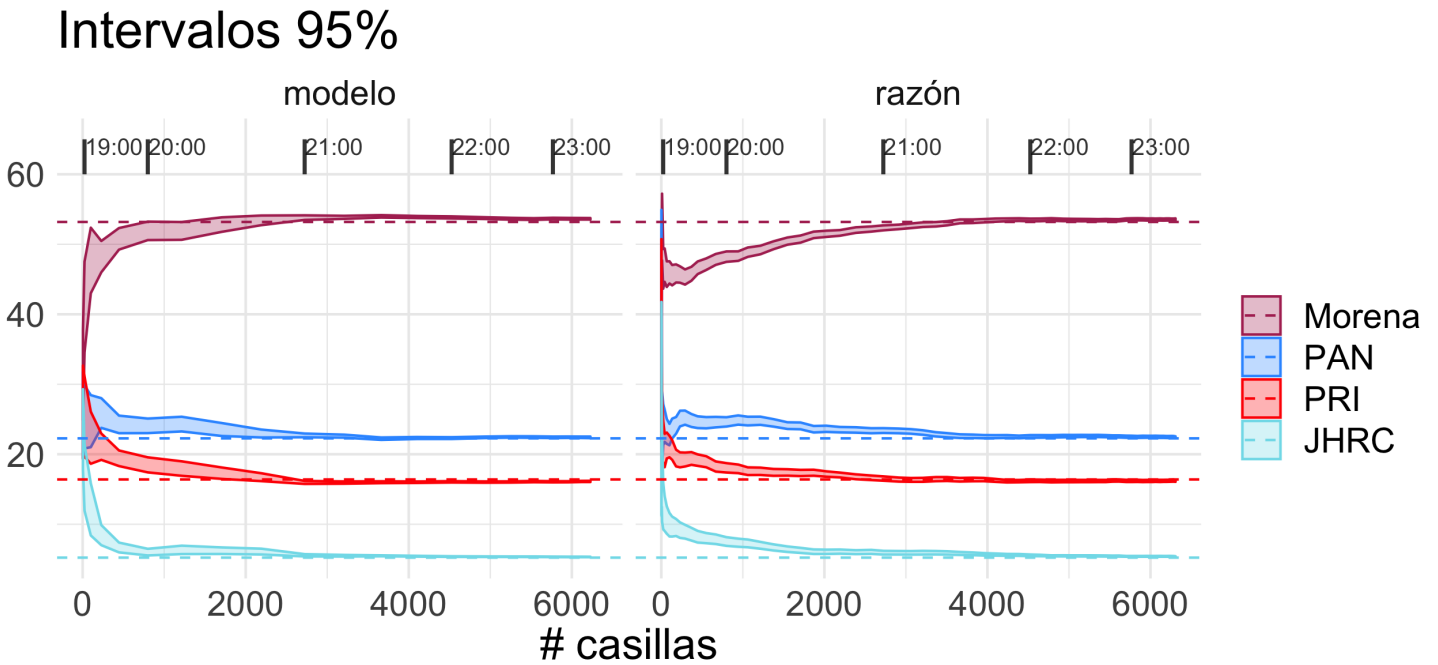
$$\begin{aligned}\theta_k = \text{logit}^{-1}(&\beta^0 + \beta^{\text{rural}} \cdot \text{rural}_k + \beta^{\text{rural\_tamañoM}} \cdot \text{rural}_k \cdot \text{tamañoM}_k \\ &+\beta^{\text{tamañoM}} \cdot \text{tamañoM}_k + \beta^{\text{tamañoL}} \cdot \text{tamañoL}_k + \beta^{\text{distrito}}_{\text{distrito}(k)} \\ &+\beta^{\text{tipoSP}} \cdot \text{tipoSP}_k),\end{aligned}$$

- Nivel 2

$$\beta_j^{\text{distrito}} \sim \mathsf{N}\left(\mu^{\text{distrito}}, \sigma^2_{\text{distrito}}\right).$$

Elección Presidencial 2018

Se reportaron los intervalos de confianza correspondientes a las 22:30 pm, con 7,787 casillas (67% de la muestra planeada).



Conclusiones

Ventajas de los modelos

Tratamiento consistente de datos faltantes: en ausencia de respuesta, la regresión atrae los parámetros hacia la media grupal,

- Comportamiento más estable de muestras parciales.
- Mejores coberturas ante problemas de sesgo.

Desventajas de los modelos

- Lentos comparado a estimador de razón o modelos más sencillos.