Proyecto Final: Ejercicio 2

Alejandra Lelo de Larrea 124433, Diego A. Estrada 165352, Victor Quintero 175897

Se estima el porcentaje de intención de voto para las elecciones del 2018 por candidato a partir de una muestra no probabilística asociada a una encuesta online. Para ello, se utiliza el método de Raking con el objetivo de calibrar los factores de expansión de tal forma que la muestra cumpla con la distribución poblacional por nivel socioeconómico (SEL), por tipo de servicio telefónico (TEL), por sexo (SEX) y por grupo de edad (AG).

1. Resultados de la calibración

Se realizaron 100 iteraciones del siguiente proceso de calibración: se calibran los factores de expansión para que la distribución muestral del TEL sea igual a la poblacional; utilizando los factores de expansión calibrados en paso anterior, se realiza una segunda calibración para que la distribución muestral de SEL sea igual a la poblacional; finalmente, utilizando los factores de expansión calibrados para SEL, se realiza una tercera calibración para que la distribución muestral conjunta SEX-AG sea igual a la poblacional. La tabla 1 muestra las distribuciones objetivo, las distribuciones de los datos muestrales previas a la post-estratificación (EFN) y las distribuciones post-estratificadas (EFC). Se logró calibrar de manera exacta todas las categorías excepto las correspondientes a la variable TEL; de éstas, la mayor diferencia se obtuvo para la categoría NEITHER.

Tabla 1: Comparación de la distribución muestral por variable.

Variable	Categoría	Objetivo	EFN	EFC
SEL	AB	3.90	33.00	3.90
	C+	9.30	30.40	9.30
	C	10.70	16.80	10.70
	C-	12.80	11.40	12.80
	D+	19.00	5.20	19.00
	D	31.80	2.60	31.79
	E	12.50	0.60	12.50
TEL	LANDLINE	2.50	5.80	2.38
	MOBILE	53.77	48.60	51.21
	BOTH	35.61	42.00	33.91
	NEITHER	8.12	3.60	12.50
SEX-AG	M 18-24	9.21	9.60	9.21
	M 25-34	10.79	13.00	10.79
	M 35-44	9.94	9.20	9.94
	M 45-59	10.46	11.60	10.46
	M 60+	7.17	5.80	7.17
	F 18-24	9.45	8.60	9.45
	F 25-34	11.87	15.60	11.87
	F 35-44	11.03	15.40	11.03
	F 45-59	11.75	8.20	11.75
	F 60+	8.33	3.00	8.33

Comparando la distribución de la muestra antes y después de realizar la post-estratificación, se puede notar que la mayor redistibución de los factores de expansión en la muestra se dio dentro de la variable SEL; mientras que servicios telefónicos, grupos de edad y sexo, tuvieron menores ajustes. Dentro del nivel socioeconómico la redistribución de los factores de expansión se dio principalmente de las clases AB y C+ a las clases D+, D y E; esto se debe en parte a que, al ser una encuesta en línea, es más factible que la población con nivel socioeconómico alto tenga acceso a servicios de internet (fijo y/o móvil) que la población en los niveles socioeconómicos más bajos. Dentro de los servicios telefónicos, destaca el aumento del factor de expansión para los individuos que no cuentan con ningún tipo de servicio telefónico; esto hace sentido ya que es menos factible que este grupo poblacional conteste la encuesta. Finalmente, de manera general se puede concluir que existe una redistribución de los factores de expansión de la población masculina a la población femenina; así como de la población entre 25 y 44 años a la población mayor a 45 años, esto último debido a que es menos probable que los adultos y, sobre todo, los adultos mayores, utilicen servicios de internet.

La tabla 2 muestra algunos estadísticos descriptivos de los factores de expansión originales y calibrados. Podemos observar que la calibración modificó los factores de expansión de tal manera que hay una gran diferencia entre el EFC mínimo y máximo, la media de la distribución es la misma que antes de la calibración, pero está muy alejada de la mediana y con una varianza muy alta. Esto hace pensar que existen algunas observaciones con un factor de expansión calibrado mucho más grande que el resto de las observaciones.

Tabla 2: Estadísticas descriptivas de los factores de expansión.

	EFN	EFC
Mínimo	81,710	7.200571e-21
Máximo	81,710	3,894,920
Suma	40,855,000	40,855,000
Mediana	81,710	5,104.068
Media	81,710	81,710
Varianza	0	81,418,677,977
Desviación Estándar	0	285,339.6

En la figura 1 se puede comprobar que los factores de expansión calibrados no descienden suavemente, pues se tiene que el factor de expansión calibrado más grande es 47.67 veces más grande que el promedio de los EFC's. Por su parte, el segundo factor de expansión calibrado más grande es 24.69 veces más grande que el promedio de los EFC's. Es importante considerar que estas dos observaciones podrían influenciar los resultados de la encuesta.

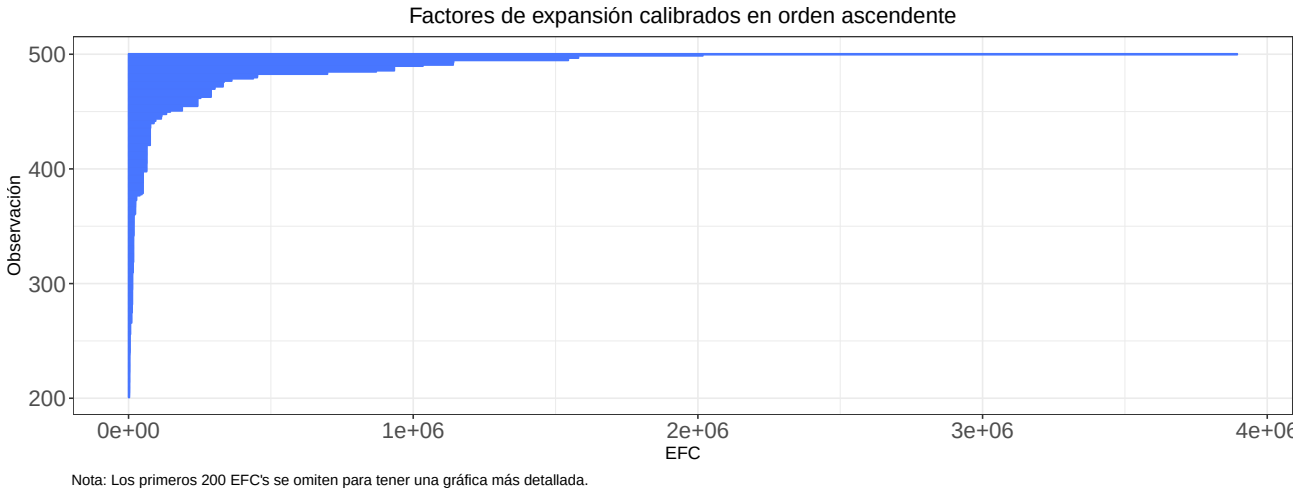

Figura 1: Factores de expansión para la muestra con post-estratificación.

2. Resultados de la encuesta

La tabla 3 muestra el porcentaje de intención de voto por candidato para la muestra antes y después de realizar la post-estratificación, así como el lugar que ocupa cada candidato. Independientemente de si se utilizan los EFN's o EFC's, AMLO resultaría ganador de la elección con el mayor porcentaje de intención de voto. Tanto Ricardo Anaya como José Antonio Meade, caen un lugar en la tabla al utilizar post-estratificación. De los candidatos independientes, Margarita Zavala es quien pierde mayor posición pasando del cuarto al séptimo lugar; mientras que el Bronco sube una posición y Jaguar se mantiene en último lugar de la tabla.

Es importante destacar que una vez calibrada la muestra, el porcentaje de indecisos pasó del quinto al segundo lugar en intención de voto; lo que significa que entre indecisos y votos nulos podrían darle la vuelta a los resultados electorales a pesar de la amplia diferencia que existe entre el porcentaje de intención de votos entre AMLO y RAC (primer y segundo lugar respectivamente sin considerar los votos indecisos).

Tabla 3: Porcentaje de intención de voto de los candidatos antes y después de post-estratificar

Candidato	EFN		EFN	
	% Int. Votos	Lugar	% Int. Votos	Lugar
AMLO	40.80	1	33.75	1
JAMK	14.60	3	11.21	4
RAC	17.40	2	14.31	3
BRONCO	2.80	7	7.26	6
JAGUAR	0.80	8	0.23	8
MZG	10.00	4	6.43	7
INDECISO	5.60	6	17.51	2
NULO	8.00	5	9.30	5

La figura 2 muestra la redistribución del porcentaje de intención de voto por candidato y por nivel socioeconómico para la muestra antes y después de la post-estratificación. Contrastando ambas gráficas, se puede notar que el porcentaje de votos de las clases socioeconómicas altas (AB, C+ y C) disminuye mientras que los de las clases socioeconómicas bajas (D+, D y E) aumentan; esto está en línea con los resultados de la sección anterior. Es importante destacar que, al utilizar factores de expansión calibrados, los votos indecisos corresponden en su mayoría a la clase E y que los votos nulos corresponden en su mayoría a la clase D+. Además, es evidente que la simpatía por AMLO es más homogénea entre todas las clases socioeconómicas; mientras que los demás candidatos son apoyados en su mayoría por un grupo socioeconómico específico.

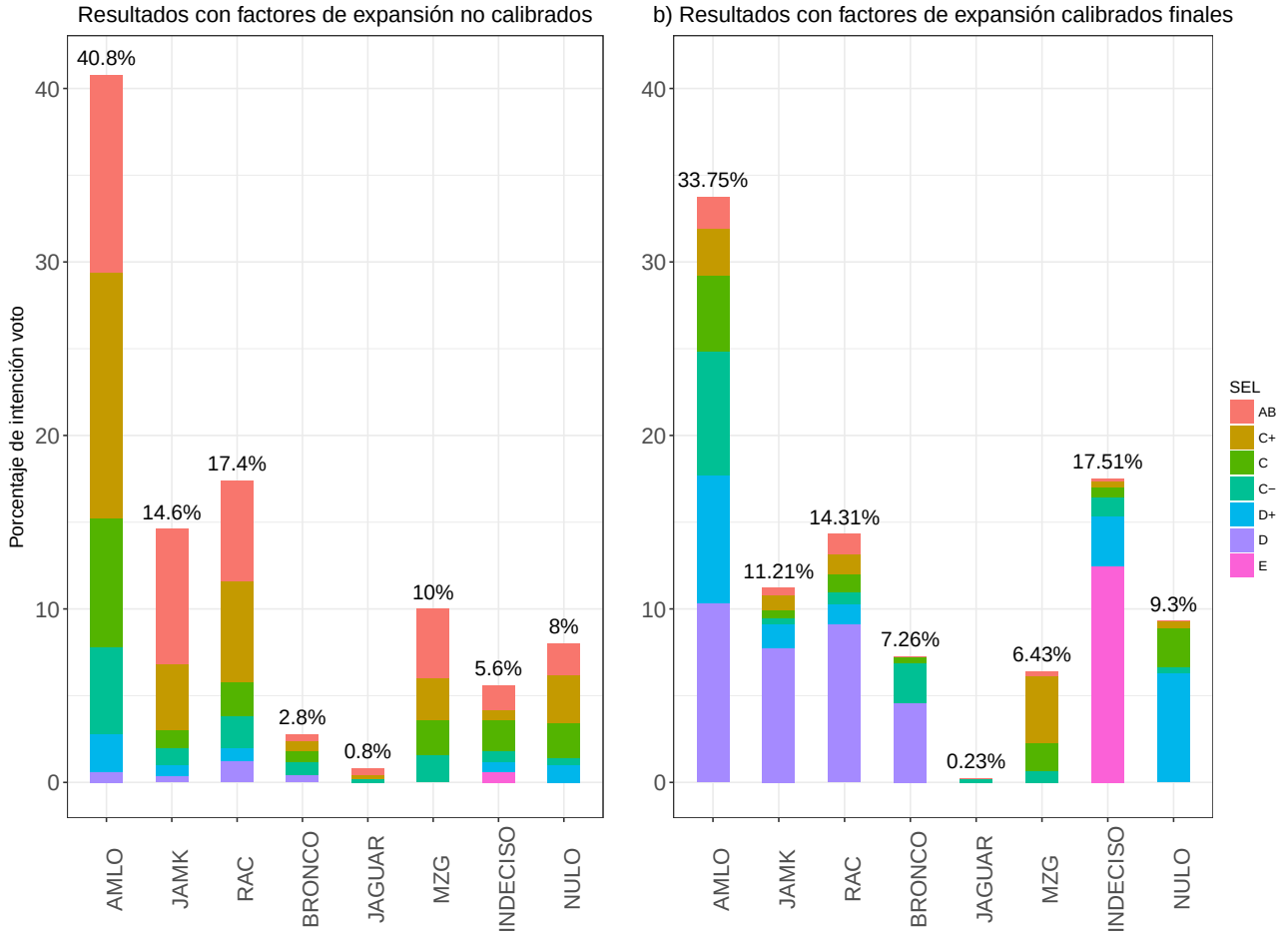

Figura 2: Porcentaje de intención de voto por candidato y por nivel socio económico antes y después de post-estratificar.

Finalmente sería interesante ver como cambian los resultados si se utiliza un podado de pesos para las dos observaciones que resultaron con un EFC muy grande; así como si se incluyeran otras variables al estudio como el grado de escolaridad o la region/estado/ciudad en que vive el encuestado con el fin de dar una mejor estimación del porcentaje de intención de voto.