Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Iztapalapa

Esquemas de muestreo para el Conteo Rápido bajo restricciones operativas

Jerónimo Hernández Mendoza

Universidad Autónoma Metropolitana unidad Iztapalapa
Maestría en Ciencias Matemáticas Aplicadas e Industriales
Ciudad de México
Octubre de 2019

Esquemas de muestreo para el Conteo Rápido bajo restricciones operativas

Jerónimo Hernández Mendoza

Tesis de grado presentada como requisito para obtener el título de:

Maestro en Ciencias
Matemáticas Aplicadas e Industriales

Co-Asesor:
Dr. Carlos Erwin Rodríguez Hernández-Vela

Co-Asesor:
Dr. Gabriel Nuñez Antonio

Universidad Autónoma Metropolitana unidad Iztapalapa
Maestría en Ciencias Matemáticas Aplicadas e Industriales
Ciudad de México
Octubre de 2019

Contenido

RESUMEN

Los conteos rápidos han estado presentes, desde hace mas de dos décadas, en los procesos electorales de México. El primer ejercicio de este tipo, tuvo por objetivo estimar los resultados de la elección presidencial de 1994 y comunicarlos a la población en la noche del mismo día de la elección. A partir de este momento, el árbitro electoral decidió implementar conteos rápidos para estimar las tendencias de la votación en cada elección presidencial. Sin embargo, debido al ambiente de desconfianza y tensión que rodea cada elección en nuestro país, en el año 2016 el Consejo General del Instituto Nacional Electoral, decidió que para generar confianza y transparencia en los procesos electorales, sería obligatorio organizar conteos rápidos tanto para elecciones federales como locales. Lo anterior, trajo nuevos retos de tipo metodológico y logístico. Uno de ellos, el problema que nos ocupa en esta tesis, se refiere a la presión que se ejerce sobre el personal que manda la información de la votación de las casillas seleccionadas en muestra; los Capacitadores y Asistentes Electorales, mejor conocidos como CAEs.

Las funciones principales de los CAEs, se resumen en dos puntos, primero, meses antes de la elección, los CAEs reclutan y capacitan a varios ciudadanos para que funjan cómo funcionarios de casilla. Segundo, durante la jornada electoral, cada CAE tiene que apoyar a estos funcionarios, en cualquier problema que se presente en las casillas. En promedio, cada CAE tiene bajo su responsabilidad 4 casillas, con sus respectivos funcionarios. Como una actividad adicional, entre otras tantas, a los CAES se les pide mandar información para el conteo rápido. Entonces, si no se contempla esta situación en el diseño de la muestra, con alta probabilidad habrá CAEs que deberán mandar información de más de una de las casillas que tienen a su cargo, por lo que tendrán que descuidar su responsabilidad principal o terminarán optando por no mandar información. En esta tesis se propone una solución a este problema.

Este trabajo está organizado de la siguiente manera: en el primer capítulo se brinda un visión de los conteos rápidos, desde sus inicios hasta llegar al proceso electoral de 2018. En el segundo capítulo, se describe la teoría básica del muestreo probabilístico. En el tercer capítulo, se aplica la teoría vista en el capítulo dos al conteo rápido. Aquí se compara el desempeño de varias estrategias de selección utilizando varias votaciones reales, tanto para gobernador como presidenciales. En el cuarto capítulo, se describe una estrategia para asegurar que en la muestra aleatoria se tenga como máximo una casilla por CAE, se define el método de estimación y el cálculo de su varianza. Además, se compara el desempeño de esta idea con el método empleado por el Comité Técnico para los Conteos Rápidos en 2018. Finalmente, en el quinto y último capítulo se presentan conclusiones y algunas ideas para trabajo futuro.

1. INTRODUCCIÓN

Los conteos rápidos (CR) son el procedimiento estadístico diseñado con la finalidad de estimar con oportunidad las tendencias de los resultados finales de una elección, a partir de la votación de una muestra aleatoria de casillas electorales. Las estimaciones se comunican a la población en la misma noche de la jornada electoral, en forma de intervalos de confianza, además se incluye la estimación del porcentaje de participación ciudadana en la elección. Para ser exactos, la Ley General de Instituciones y Procedimientos Electorales (Legipe) establece que el nivel de confianza debe ser del 95 % [Reglamento de Elecciones 2016].

Aún cuando los CR estiman, con un margen de error muy pequeño, los resultados finales de una elección, su aceptación ha enfrentado una infinidad de retos a través de los años. Basta mencionar que en sus inicios fue objeto de denuncias, quejas y controversias por parte de partidos, analistas, medios de comunicación, redes sociales y ciudadanos [Alonso y Coria]. Por este motivo, los CR han ido evolucionando a través del aprendizaje en cada proceso electoral.

ELECCIÓN PRESIDENCIAL DE 1994

El primer conteo rápido, del que se tiene registro, fue implementado en 1994 por el entonces Instituto Federal Electoral (IFE), en la elección presidencial de ese año. Los resultados del conteo fueron altamente coincidentes con los del cómputo final, probando así su valía para generar certidumbre sobre los resultados de una elección. Por tanto, se estableció su implementación en cada elección federal. En la tabla **1-1**, se muestran las estimaciones del CR, los resultados del PREP, así como los resultados definitivos de la elección.

	ESTIMACIONES (%)			PREP (%)	RESULTADO DE LA ELECCIÓN (%)
	MÍNIMO	MÁXIMO	PRESICIÓN		
PRI	49.3	50.7	0.7	50.1	50.2
PAN	26.8	28.2	0.7	28.8	26.7
PRD	15.8	17.1	0.65	17.1	17.1

Tabla 1-1.: Comparación de resultados entre el conteo rápido, PREP y cómputo final, 1994.

ELECCIÓN PRESIDENCIAL DEL AÑO 2000

Para la elección federal del año 2000, los resultados del conteo rápido fueron esenciales la noche de la jornada electoral, ya que el PREP no tuvo resultados hasta 23 horas después del cierre de las casillas. En ese entonces, el comité técnico para los conteos rápidos, ahora COTECORA, se conformó con tres de los integrantes del comité de 1994 más el coordinador del PREP. Sin embargo, y al igual que en 1994 los ejercicios de estimación no los realizó directamente el IFE, si no que se contrató a tres empresas encuestadoras para realizar este proceso.

Al dar a conocer las estimaciones, el factor primordial para la pronta aceptación de los resultados que arrojó el conteo rápido fue que el presidente de la República los aceptó públicamente, lo que redujo la tensión y contribuyó a que los demás actores políticos también aceptaran las estimaciones. La comparación entre las estimaciones del CR, PREP y resultados definitivos se muestran en la tabla **1-2** [Alonso y Coria].

	BERUMEN (%) 622 casillas = 43.65% de su muetsra				GALLUP MÉXICO (%) 1511 casillas = 97.61% de su muetsra				INTEGRADO POR EL INSTITUTO (%)			PREP FINAL (%)	RESULTADO FINAL (%)
	Estimación	Mínimo	Máximo	Precisión	Estimación	Mínimo	Máximo	Precisión	Mínimo	Máximo	Precisión		
Alianza por el Cambio	43.2	41.2	45.2	2.0	42.1	40.8	43.3	1.3	39.0	45.0	3.0	42.7	42.5
PRI	34.7	33.3	36.2	1.5	36.6	35.5	37.6	1.1	35.0	38.0	1.5	35.8	36.1
Alianza por México	16.8	15.5	18.0	1.3	16.4	15.5	17.2	0.8	15.1	18.0	1.5	16.5	16.6

Tabla 1-2.: Comparación de resultados entre el conteo rápido, PREP y cómputo final, 2000.

Una vez más, los resultados del conteo rápido fueron coincidentes con los resultados finales de la elección. Entonces, para las elecciones intermedias de 2003, se decidió implementar el conteo rápido para estimar la conformación de la Cámara de Diputados.

ELECCIONES LOCALES 2003

Esta elección fue un verdadero reto, ya que la composición de los 500 diputados que conforman la cámara, se determina de la siguiente manera:

1. 300 diputados por elección directa, en cada uno de los 300 distritos electorales en los que se divide el país.

2. 200 diputados mediante representación proporcional, pero aplicando algunas reglas de no sobre-representación descritas en el [Reglamento de Elecciones 2016].

Entonces, primero, era necesario que la muestra tuviera información de cada uno de los 300 distritos electorales. Ya que en cada distrito se tendría que hacer una estimación. Segundo,

no sería posible utilizar métodos convencionales para realizar la estimación de la conformación debido a las reglas de no sobre-representación. En virtud de lo anterior, el IFE creó un nuevo Comité de Conteo Rápido, conformado por científicos que estuvieron en el ejercicio de 2000. Adicionalmente, en esta ocasión no se convocó a ninguna empresa; el IFE decidió emplear sus propios recursos, los Capacitadores y Asistentes Electorales (CAE), fueron los encargados de transmitir la información de las diferentes casillas en muestra.

Los resultados de esta elección se presentan en la tabla **1-3**. Se observa que la conformación de la cámara definitiva es estimada con gran precisión por el CR [Alonso y Coria].

	ESTIMACIÓN CONTEO RÁPIDO (%)				NÚMERO DE DIPUTADOS		PREP (%)	CÓMPUTOS DISTRITALES (%)	NÚMERO DE DIPUTADOS	
	Puntual	Mínimo	Máximo	Precisión	Mínimo	Máximo			Asignación con base en PREP	Real
PAN	30.5	30.0	31.0	0.5	148	158	30.6	30.8	155 (154)*	151
PRI	34.4	33.9	34.9	0.5	222	227	34.4	34.6	2	222 (224)
PRD	17.1	16.6	17.6	0.5	93	100	17.7	16.6	1	96
PT	2.4	1.9	2.9	0.5	5	8	2.4	2.4	6 (5)*	5
PVEM	6.2	5.9	6.5	0.3	14	16	6.1	6.1	15 (17)*	17**
CONV	2.3	2.1	2.5	0.2	5	6	2.3	2.3	5	5
PSN	0.3	0.2	0.40	0.1			0.3	0.3		
PAS	0.7	0.6	0.80	0.1			0.7	0.7		
MP	1.0	0.9	1.10	0.1			0.9	0.9		
PLM	0.4	0.3	0.50	0.1			0.4	0.5		
FC	0.5	0.4	0.60	0.1			0.5	0.5		

*En dos de los distritos que ganó Alianza para todos de acuerdo al convenio, al momento de registro se cambiaron por dos de PVEM. Eso ocasionó que se redujera en uno de los pluronominales del PAN y del PT.
**En la tabla de asignación final faltan cuatro diputados en razón de la nulidad en dos eleciones

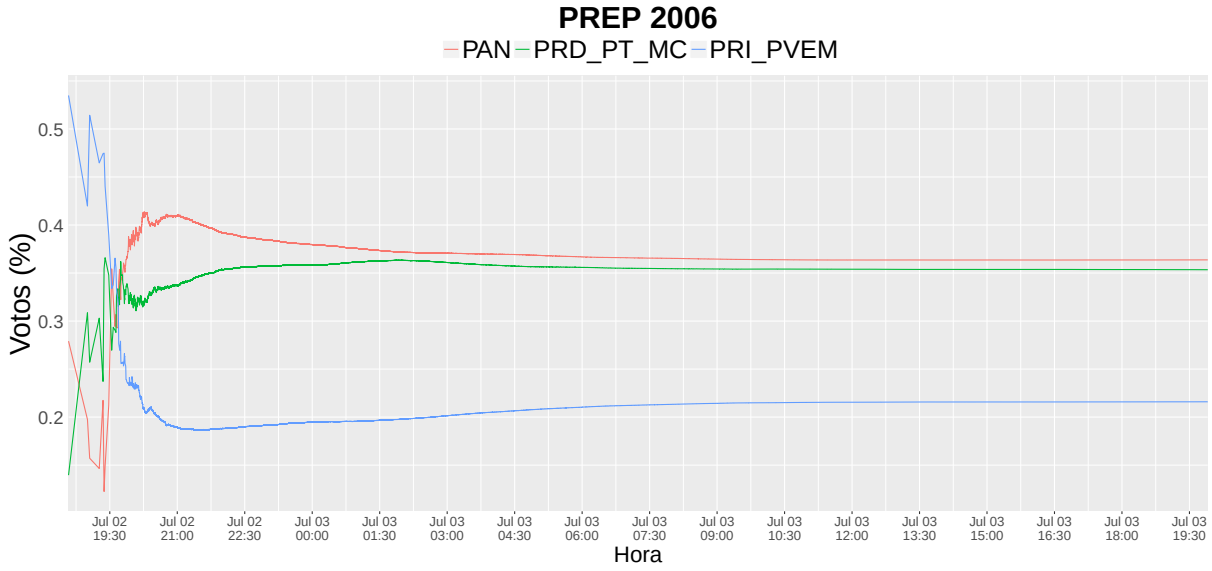
Tabla 1-3.: Comparación de resultados entre el conteo rápido, PREP y cómputo final, 2003.

ELECCIÓN PRESIDENCIAL DE 2006

Esta elección presidencial ha sido una de las más competidas de la historia de nuestro país, con un margen de diferencia muy pequeño entre el primer y segundo lugar. En esta elección el IFE decidió hacer el CR con sus propios recursos, es decir, empleó a los CAEs. Además, se determinó utilizar un tamaño de muestra de $7,636$ casillas, distribuidas en 481 estratos. Con este tamaño, se estimarían las proporciones de votos con un margen de error de $0.5\,\%$ y con una confiabilidad mayor o igual a $95\,\%$. Para ello, se utilizaron tres métodos estadísticos: clásico, bayesiano y robusto. Sin embargo, a pesar de que al momento del corte se había recibido el $95.12\,\%$ de la muestra total, se observó un traslape en los intervalos de confianza para las estimaciones de los dos candidatos punteros.

Bajo esta situación, el entonces Consejero Presidente del IFE (Luis Carlos Ugalde), anunció, conforme al acuerdo del Consejo General (IFE 2006), que los resultados del procedimiento de conteo rápido no permitían anunciar a un ganador y por lo tanto no se darían a conocer las estimaciones del conteo rápido a la población. Por su puesto, esto creo incertidumbre

acerca de la transparencia del CR y aumentó las dudas sobre la certeza de la elección, pues dejó al PREP como único instrumento para conocer el resultado de la elección antes de los cómputos distritales.



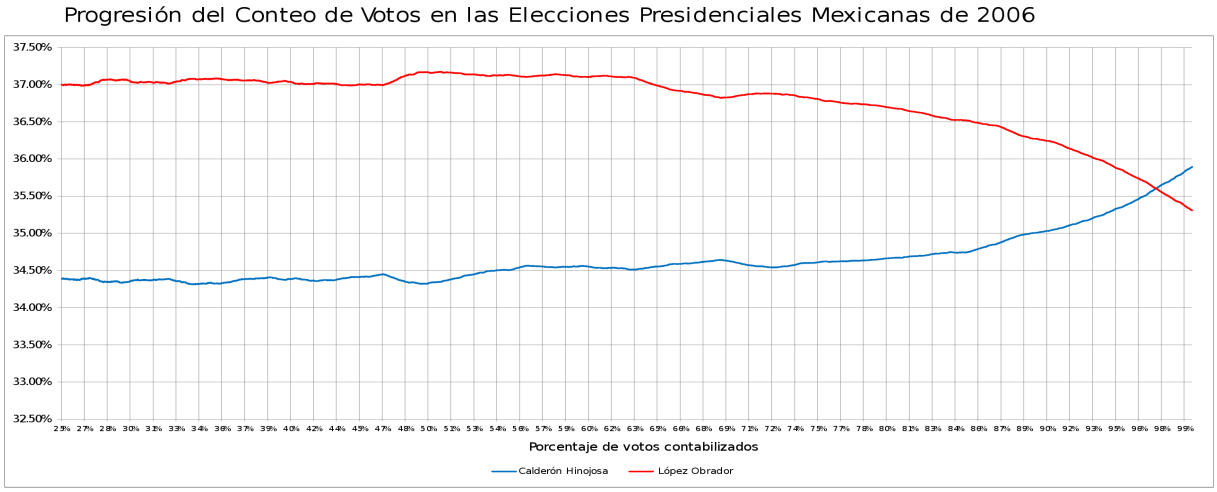
Gráfica 1-1.: Captura de votos PREP 2006.

La captura de votos realizada por el PREP comenzó a las 18 : 00 horas del 2 de julio de 2006, el porcentaje de avance en la recepción de votos para los tres partidos con mayor votación se presenta en la gráfica **1-1**[1] donde se observa que en las primeras 2 horas, el candidato de la colación PRI_PVEM Roberto Madrazo Pintado estaba en primer lugar, seguido por Andrés Manuel López Obrador de la coalición PRD_PT_MC. Por su parte, el candidato del PAN Felipe Calderón Hinojosa estaba en tercer lugar. Posterior a estas dos horas, los lugares entre el primero y el tercero se invierten colocando a Felipe Calderón en primer lugar, posición que mantendría al 3 de julio al cierre del PREP a las 20 : 00 horas con el 92.16 % de casillas computadas. No obstante, la diferencia entre los dos primeros lugares fue de apenas el 1.03 % de votos, esto es, el 36.37 % de votos para Felipe Calderón y 35.34 % para Andrés Manuel[2].

El 5 de julio dio inicio el conteo oficial en los 300 distritos electorales, el cual duró más de 30 horas. Al comienzo del conteo López Obrador estaba a la cabeza, seguido por Felipe Calderón, con una diferencia de 2.59 % al llevar el 25 % de actas computadas, minutos después se daría un apagón general en las pantallas que mostraban los resultados del sistema de cómputo por espacio de 5 segundos.

[1] Gráfica construida con datos oficiales del PREP 2006. `https://portalanterior.ine.mx/documentos/proceso_2005-2006/prep2006/bd_prep2006/bd_prep2006.htm`

[2] Estos resultados no consideran las actas computadas que presentaban inconsistencia en su llenado



Gráfica 1-2.: Resultados de casillas computadas por el IFE (Cómputos Distritales).

Finalmente, el jueves 6 de julio a las 3:56 horas con un 97.70 % de las casillas computadas, López Obrador pasó al segundo lugar, siendo aventajado por Felipe Calderón (ver Gráfica **1-2**). El conteo concluyó a las 15 : 20 horas con el 35.89 % de los votos para Felipe Calderón, y el 35.31 % para López Obrador [3].

En la tabla **1-4** se presentan los resultados de esta elección, en los intervalos de confianza se observa una intersección entre primer y segundo lugar (método robusto y clásico). No obstante, obsérvese de nueva cuenta la coincidencia de resultados entre el conteo rápido, el PREP y los cómputos distritales [Alonso y Coria].

PARTIDO COALICIÓN	ROBUSTO (%)			CLÁSICO (%)			BAYESIANO (%)			PREP (%)	CÓMPUTOS DISTRITALES (%)
	Mínimo	Máximo	Precisión	Mínimo	Máximo	Precisión	Mínimo	Máximo	Precisión		
PAN	35.25	37.40	1.08	35.68	36.53	0.43	35.77	36.40	0.31	36.40	35.89
APMx	20.85	22.70	0.93	21.66	22.26	0.30	21.72	22.24	0.26	21.48	22.26
CPBT	34.24	36.38	1.07	34.97	35.70	0.36	35.07	35.63	0.28	35.41	35.31
NA	0.75	1.19	0.22	0.93	1.03	0.05	0.94	1.05	0.06	0.99	0.96
ASDC	2.40	3.13	0.37	2.60	2.80	0.10	2.60	2.80	0.10	2.82	2.70

Tabla 1-4.: Comparación de resultados entre el conteo rápido, PREP y cómputo final, 2006.

ELECCIONES LOCALES 2009

En las elecciones intermediase del 2009, para la renovación de la Cámara de Diputados, no se realizó un conteo rápido. Las razones de esta decisión no se conocen en su totalidad, pero es lógico pensar que fue consecuencia de lo sucedido en 2006.

[3] https://es.wikipedia.org/wiki/Elecciones_federales_de_México_de_2006

ELECCIÓN PRESIDENCIAL DE 2012

El conteo rápido en la elección presidencial de 2012 fue posible gracias a que el Consejo General, de aquel entonces, especificó que los resultados del Conteo Rápido, así como los rangos de votación por candidato, cualesquiera que sean las diferencias entre ellos, se difundirían a la opinión pública[4]. En esta ocasión, el número de casillas en muestra fue de $7,597$, con una estratificación semejante al de 2006 (con 483 estratos). A pesar de que al momento del corte sólo se disponía del $82.4\,\%$ de las casillas en muestra [5], no hubo traslape alguno entre las estimaciones. Mas aún, los intervalos de confianza resultaron en una coincidencia total con los resultados del PREP y el cómputo final de la elección [Alonso y Coria].

CANDIDATO	MÍNIMO (%)	MÁXIMO (%)	PRECISIÓN (%)	PREP (%)	CÓMPUTOS DISTRITALES (%)
Josefina Vázquez Mota	25.10	26.03	0.46	25.40	25.41
Enrique Peña Nieto	37.93	38.55	0.31	38.15	38.21
Andrés Manuel López Obrador	30.90	31.86	0.48	31.64	31.59
Gabriel R. Quadri de la Torre	2.27	2.57	0.15	2.30	2.29

Tabla 1-5.: Comparación de resultados entre el conteo rápido, PREP y cómputo final, 20012.

LEY GENERAL DE INSTITUCIONES Y PROCEDIMIENTOS ELECTORALES 2014

Para el año 2014, se promulga la nueva Ley General de Instituciones y Procedimientos Electorales (LEGIPE), en donde se definen y describen con mayor rigor los programas de resultados preliminares. Destaca el énfasis en su carácter meramente informativo y que no son definitivos. Con referencia a los conteos rápidos, la ley faculta al Instituto Nacional Electoral (INE), antes IFE, y a los Organismos Públicos Locales Electorales (OPLE) para ordenar su realización. Finalmente, en el reglamento de elecciones emitido por el Consejo General en 2016 se establece la obligación de contar con conteos rápidos para las elecciones de gobernador y de jefe de gobierno [Reglamento de Elecciones 2016].

[4] Acuerdo del Consejo General CG297/2012 del 16 de mayo de 2012
[5] https://portalanterior.ine.mx/documentos/proceso_2011-2012/alterna/conteo-rapido.html

ELECCIONES LOCALES DE 2015-2017

Debido a este cambio en la legislación electoral, desde 2015, se han realizado conteos rápidos para estimar las tendencias de la votación en diferentes elecciones. Particularmente, en la elección intermedia de 2015, el INE organizó un conteo rápido para conocer la composición de la Cámara de Diputados. La muestra se diseñó aleatoriamente a partir de un número fijo de casillas por cada distrito (treinta) con un ligero incremento en aquellos distritos que, por su diferencia horaria, pudieran tener problemas en el acopio de los resultados. Al final, la muestra fue de 9,450 casillas, de las cuales solo se tuvo información del 74.24 %, al momento de realizar las estimaciones [Alonso y Coria]. La Tabla **1-6**, muestra que los intervalos de confianza del conteo rápido arrojaron una coincidencia casi total con el PREP.

Contendientes	PREP corte de las 20:10 horas, avance del 98.63 %			Conteo Rápido (estimación COTECORA)		
	MR	RP	CONF	MIN	MAX	Contiene al valor del PREP
PAN	56	53	109	105	116	SÍ
PRI	156	41	197	196	203	SÍ
PRD	28	27	55	51	60	SÍ
PVEM	27	18	45	41	48	SÍ
PT	6	7	13	3	12	NO
MC	11	15	26	24	29	SÍ
PANAL	1	10	11	9	12	SÍ
MORENA	14	21	35	34	40	SÍ
PH	0	0	0	0	1	SÍ
ES	0	8	8	8	10	SÍ
CI	1	0	1	1	1	SÍ
TOTAL	300	200	500			

Tabla 1-6.: Comparación de resultados entre conteo rápido y PREP, 2015.

Como se mencionó anteriormente, desde 2015, se empezaron a implementar conteos rápidos para estimar los resultados de varias elecciones para gobernador y para Jefe de Gobierno (CDMX 2015). El comparativo entre las estimaciones del CR y los cómputos finales de cada elección se resumen en la tabla **1-7** [Alonso y Coria] (únicamente se presenta el comparativo para los tres candidatos punteros en cada elección).

	PRIMER LUGAR (%)			SEGUNDO LUGAR (%)			TERCER LUGAR (%)		
	Mínimo	Máximo	Cómputos estatales	Mínimo	Máximo	Cómputos estatales	Mínimo	Máximo	Cómputos estatales
Sonora	46.2	48.3	47.6	39.2	41.4	40.6	2.8	3.6	3.4
Veracruz	33.3	34.8	34.4	29.0	30.4	30.3	26.5	28.2	26.4
Colima	42.7	43.9	43.2	39.0	40.3	39.7	11.5	12.2	12.0
Oaxaca	30.5	33.7	32.1	22.7	25.5	24.9	22.3	25.7	22.9
Zacatecas	37.1	39.4	37.4	26.3	29.0	27.3	18.0	20.3	17.8
Nayarit	38.0	41.4	38.6	24.8	28.2	26.5	10.3	12.7	12.0
México	32.8	33.6	33.7	30.7	31.5	30.9	17.6	18.3	17.9
Coahuila	34.7	37.3	38.2	36.6	39.1	35.8	11.2	12.4	12.0

Tabla 1-7.: Conteo rápido en elecciones de gobernador de $2015 - 2017$.

Cómo puede observarse, en la mayoría de las elecciones locales, la estimación vía los intervalos de confianza del CR coinciden con el cómputo estatal. La única excepción fue COAHUILA, en donde:

- Hubo un traslape entre los intervalos para los dos candidatos punteros, lo que no necesariamente indica un problema.
- Los intervalos para los dos candidatos punteros, no cubrieron a los valores de los cómputos estatales. Esto tampoco tendría que ser un problema mayor, sin embargo, el error o distancia entre el intervalo y el valor real es considerable.
- Se hizo la estimación con sólo el 54.61 % de la muestra.

Estos detalles, y algunos otros, crearon gran inconformidad con el desempeño del CR en Coahuila[6].

NUEVOS RETOS

Las modificaciones a la Ley General de Elecciones trajo consigo nuevos retos de tipo metodológico y logístico, particularmente en elecciones locales en donde en algunos estados el total de casillas instaladas es mucho menor a 1,000. En estos casos, el tamaño de muestra necesaria para garantizar un margen de error aceptable en la estimación (entre 0.5 % y 1 %), es muy grande en términos porcentuales. Además, dada la selección aleatoria de casillas, existen problemas de tipo operativo que repercuten directamente en las estimaciones finales. El problema en el que nos concentramos en esta tesis aparece cuando un mismo CAE tiene que reportar la votación de más de una casilla. Cada CAE es responsable de, en promedio, 4 casillas; al conjunto de casillas que son responsabilidad de un mismo CAE se le llama ARE

[6]https://www.jornada.com.mx/2017/06/25/politica/009n2pol

(Área de Responsabilidad Electoral). Para entender el problema, debemos tener en cuenta que, el día de la elección, la principal tarea de los CAEs es apoyar a los funcionarios de casilla para el correcto funcionamiento del proceso electoral, y que contribuir al CR es una actividad extra. Entonces, si esto no se considera para definir la estrategia de selección de casillas, pueden suceder dos cosas:

- Generar problemas en las casillas a cargo de CAEs con sobrecarga de casillas para el CR.
- Que los CAEs decidan no reportar las casillas del CR, aumentando la No respuesta.

Este segundo punto debe analizarse con cuidado, ya que en los estados en los que se ha realizado CR, para estimar los resultados de la elección local, se han observado los siguientes porcentajes de no respuesta[7]:

Estado	**No Respuesta**	**Hora de corte**
SONORA (2015)	29.53 %	01 : 22 horas, siguiente día.
VERACURZ (2016)	9.71 %	23 : 30 horas
ZACATECAS (2016)	10.66 %	22.40 horas
COLIMA (2016)	27.8 %	20 : 31 horas
OAXACA (2016)	36.74 %	01 : 30 horas, siguiente día.
NAYARIT (2017)	50 %	00 : 41 horas, siguiente día.
COAHUILA (2017)	45.39 %	02 : 05 horas, siguiente día.
MÉXICO (2017)	25.9 %	21 : 10 horas

En los Conteos Rápidos de 2016 y 2017 el porcentaje de no respuesta fue siempre mayor al 25 %, llegando en algunos casos al 50 %. Además, hay que tomar en consideración que para que un CR sea relevante, se deben comunicar las estimaciones en la noche del mismo día de la elección.

ELECCIONES LOCALES Y PRESIDENCIAL DE 2018

Teniendo en cuenta estos porcentajes de no respuesta, el COTECORA para las elecciones de 2018, puso una restricción para el tamaño de muestra. Se buscó garantizar que al menos el 80 % de los CAEs que participarían en el CR, reportaran información de una sola casilla. Entonces, se abordó sólo parcialmente el problema de sobrecarga de los CAEs, ya que equivalió a utilizar un diseño estratificado común (como siempre se ha hecho), simplemente se buscó la mejor estratificación hasta satisfacer la restricción. Otro acuerdo fue que el margen de error en ningún caso fuera mayor al 1 %.

Con la estrategia descrita en el párrafo anterior, los porcentajes de no respuesta estuvieron entre 20 % y 42 %. Sin embargo, esta no respuesta se debió a diferentes factores, por ejemplo:

[7]Información tomada de la página oficial de elecciones de cada estado.

en un gran número de casillas se tenían elecciones concurrentes (locales y federales), además, las coaliciones en varios estados dificultaban el cómputo de los votos y finalmente, se puso mucho énfasis en que se comunicaran estimaciones a la población antes de las 12 de la noche. Todo lo anterior, obligando al Comité a realizar la estimación final con muestras incompletas.

La tabla **1-8** muestra las estimaciones de los Conteos Rápidos para el proceso electoral 2018. Se presenta información únicamente para el primer y segundo lugar de la elección, los resultados del cómputo final y los porcentajes de no respuesta de cada elección[8]. Además, se incluye la hora de corte para las estimaciones y se ordena en orden descendente según el porcentaje de no respuesta.

	PRIMER LUGAR (%)			SEGUNDO LUGAR (%)			NO RESPUESTA (%)	CORTE
	Mínimo	Máximo	Cómputo final	Mínimo	Máximo	Cómputo final		
Morelos	51	53	52.6	13.4	16.1	14.1	20.0	22:10 pm
Jalisco	37.7	40	39	23	25.3	24.7	26.3	22:15 pm
Puebla	36.4	38.9	38.1	33.9	36.8	34.1	26.9	23:45 pm
Guanajuato	49.5	51.5	49.9	23.2	25.2	24.2	28.6	21:45 pm
Veracruz	43.9	45.9	44	37	38.7	38.4	29.9	23:10 pm
Presidencial	50	53.8	53.2	22.1	22.8	22.3	32.5	22:30 pm
Yucatán	38.4	40.8	39.6	34	36.5	36.1	37.3	23:55 pm
Chiapas	40.2	44.2	39.3	19.1	22.6	22.5	41.2	00.30 am (2 de julio)
Cdmx	46.6	47.7	47.1	30.4	31.2	31	41.6	22:15 pm
Tabasco	62.1	64.3	61.4	16.8	18.4	19.6	41.6	23:50 pm

Tabla 1-8.: Conteos rápidos en elecciones de gobernador y presidencial 2018

Como se puede observar, hay una coincidencia total entre los resultados del CR y el cómputo final para el primer lugar a excepción de CHIAPAS Y TABASCO donde las estimaciones sobrestimaron los resultados finales. Adicionalmente, para el segundo lugar, esta coincidencia es casi total, la excepción fue en el estado de TABASCO donde se subestimó por 1.2 % al porcentaje real (resaltado de color naranja). Adicionalmente, también se observa un traslape entre los intervalos (marcados en color rojo), que corresponde al estado de PUEBLA. A pesar de estos dos detalles, en general las estimaciones del CR de 2018 reflejaron correctamente las tendencias de las votaciones. Finalmente, nótese que los porcentajes de no respuesta son altos. Por ejemplo, para CDMX y TABASCO, esta fue del 40.6 % de las casillas en muestra con el último corte de las 22 : 15 y 23 : 50 horas respectivamente. Por su parte, el estado donde se observó una afluencia menor de casillas fue CHIAPAS, pues con el corte de las 00 : 30 horas del día siguiente a la jornada electoral tenía una no respuesta del 40.1 % de las casillas en muestra.

[8]https://www.ine.mx/voto-y-elecciones/resultados-electorales/

EL PROBLEMA A RESOLVER

Como ya se mencionó, los altos porcentajes de no respuesta no dependen únicamente de las casillas que debe reportar cada CAE al CR. Sin embargo, garantizar mediante el diseño de muestreo (que no castigue el tamaño de muestra) que cada CAE que colabore con el conteo rápido reporte información de una única casilla ayudaría en varios sentidos: no se distraería a los CAEs de su labor principal, no habría sobrecarga y se podrían obtener estimaciones con menor margen de error ya que no se castigaría el tamaño de muestra. Dado lo anterior, en este trabajo se propone una estrategia de selección probabilística de casillas que

1. Asegure que los CAEs que participen en el CR, reporten cómo máximo una casilla.

2. Alcance un margen de error, al menos tan bueno como el del muestreo estratificado común.

Para este fin, en el segundo capítulo de este texto se describen las estrategias de muestreo probabilístico más comunes: Muestreo Aleatorio Simple (MAS), Muestreo Aleatorio Estratificado (MAE), Muestreo por Conglomerados (MC) y Muestreo Estratificado con Conglomerados en Dos Etapas (MECDE), y se dan a conocer sus propiedades. En el tercer capítulo, se reescribe cada una de estas estrategias para aplicarlas al conteo rápido, lo anterior se hace utilizando bases de datos de diferentes elecciones, para conocer su desempeño. Se calculan tamaños de muestra, márgenes de error y se compara el margen de error del estimador combinado y separado para un MAE.

En el capítulo cuatro, se describe la estrategia de selección que resuelve el problema operativo sobre los CAEs, satisfaciendo los puntos anteriores. Se hace una comparación, vía simulación, entre los resultados del método que tradicionalmente se emplea en los conteos rápidos y el método propuesto. Para esto último, se usan las bases de datos de las elecciones locales de 2012 y 2016, y la presidencial de 2018.

En el capítulo cinco se despliega una gráfica comparativa entre el margen de error empleado en las elecciones del 2018, con los tamaños de muestra definidos por el Comité del CR 2018, y el margen que se hubiera obtenido con la estrategia propuesta. Finalmente, se presentan algunas conclusiones, así como ideas para trabajo futuro.

2. TEORÍA BÁSICA DEL MUESTREO PROBABILÍSTICO

El muestreo probabilístico es un método de selección y estimación en donde cada elemento en la población tiene una probabilidad de selección conocida y distinta de cero. Considerando estas probabilidades, se emplea un mecanismo aleatorio para elegir los elementos que se incluirán en la muestra.
Si el diseño del muestreo probabilístico se implementa bien, un investigador puede usar una muestra relativamente pequeña para hacer inferencias sobre una población arbitrariamente grande. Generalmente, el interés recae en la estimación de totales, promedios y porcentajes poblacionales.

El término muestreo probabilístico, hace referencia a un método con las siguientes características:

1. Se tiene una población de interés con un total de $N < \infty$ elementos, cuyos índices denotaremos como
$$U = \{1, 2, 3, \ldots, N\}.$$
A este conjunto se le denomina población o universo.

2. Nos interesa conocer alguna característica de la población, a la medición de esta característica en cada elemento de la población, se le denotará como
$$\{y_1, y_2, y_3, \ldots, y_N\}.$$

3. Las probabilidades de selección y el mecanismo aleatorio generan conjuntos $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_v$. Cada uno de estos conjuntos, está formado por índices de la población U. Estas serán las muestras posibles y cada muestra $\mathcal{S}_i$, tendrá una probabilidad conocida de selección, así como cada elemento de la población i.

4. Se construye un estimador del parámetro de interés y este considera las probabilidades de selección de cada elemento de la población.

Algunas de las ventajas y desventajas del muestreo probabilístico son:

VENTAJAS

- Ausencia de sesgo sistemático y de muestreo
- Mayor nivel de fiabilidad de los resultados de la investigación
- La posibilidad de hacer inferencias sobre la población
- Permite estimar los errores que se cometen.

DESVENTAJAS

- Mayor complejidad comparada con el muestreo no probabilístico
- Puede llevar más tiempo
- Normalmente más caro que el muestreo no probabilístico

La teoría de muestreo ofrece varias estrategias de selección. Las revisadas para el desarrollo de la tesis son las siguientes: muestreo aleatorio simple sin reemplazo, muestreo estratificado y muestreo por conglomerados.

2.1. MUESTREO ALEATORIO SIMPLE (MAS)

Una muestra de $n \leq N$ elementos de la población es llamada un muestreo aleatorio o muestreo aleatorio simple si los elementos se eligen de tal forma que todas las posibles selecciones de n elementos sean igualmente probables. Denotamos al conjunto de los n elementos por $\mathcal{S}$.

Esta es la estrategia más básica de muestreo probabilístico y proporciona la base teórica para las estrategias complejas. Hay dos maneras de seleccionar una muestra aleatoria simple: **con reemplazo**, que puede pensarse como la extracción independiente de muestras de tamaño 1 en la población, regresando este elemento antes de seleccionar el siguiente. Aquí cada elemento tiene probabilidad $\frac{1}{N}$ de ser seleccionado en una extracción. Sin embargo, es probable que el mismo elemento se incluya más de una vez en la muestra. En ocasiones, lo anterior no es deseable, por lo que se prefiere el MAS **sin reemplazo**, en donde cada elemento de la muestra es único o de otra manera, cada elemento de la población se puede seleccionar sólo una vez en muestra.

2.1.1. MUESTREO ALEATORIO SIMPLE SIN REMPLAZO (MASSR)

Sea Z_i una variable dicotómica que indica si el i-ésimo elemento de la población está en muestra o no.

$$Z_i = \begin{cases} 1, & \text{si el } i-\text{ésimo elemento está en muestra}, i = 1, ..., N \\ 0, & \text{en otro caso.} \end{cases}$$

Por combinatoria, sabemos que existen $\binom{N}{n}$ posibles maneras de seleccionar n elementos de una población de tamaño N. Además, es fácil ver que si la i-ésima observación está en muestra, entonces quedan $\binom{N-1}{n-1}$ posibles maneras de seleccionar a los $n-1$ elementos restantes del total $N-1$. Entonces, la probabilidad de seleccionar al i-ésimo elemento de la población en muestras es

$$P(Z_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

2.1.2. PROPIEDADES BÁSICAS

A menudo, se esta interesado en conocer algunos parámetros de la población, tales como el total y la media de los valores. Otro parámetro de especial interés es la razón entre los valores observados y alguna otra variable auxiliar. En la siguiente tabla se describen los parámetros mencionados, así como sus respectivos estimadores muestrales [Cochran]. Identificaremos a los parámetros poblacionales con letras mayúsculas y los estimadores con minúsculas.

	Total	Media	Razón
Valor Real:	$Y = \sum_{i=1}^{N} y_i$	$\overline{Y} = \frac{Y}{N}$	$R = \frac{\overline{Y}}{\overline{X}}$
Estimador:	$y = N\bar{y}$	$\bar{y} = \frac{1}{n}\sum_{i \in \mathcal{S}} y_i$	$r = \frac{\bar{y}}{\bar{x}}$

Tabla 2-1.: Estimadores por MAS

Además de estos parámetros, el cálculo de la **varianza poblacional**, así como su estimador son fundamentales en la teoría de muestreo.

	Poblacional	Muestral
Varianza:	$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2$,	$s^2 = \frac{1}{n-1}\sum_{i\in S}\left(y_i - \bar{y}\right)^2$.

En donde S es el subconjunto de índices que componen la muestra.

2.1.3. PROPIEDADES DE LOS ESTIMADORES

Un estimador es una variable aleatoria función de la muestra que se construye para aproximar al valor de un parámetro, y la estimación es el valor que toma el estimador una vez sustituidos los valores observados en la muestra en el estimador. Es de suma importancia conocer las propiedades de los estimadores, en muestreo nos interesan dos aspectos: esperanza y varianza. Además, siempre buscaremos (en la medida de lo posible) que el estimador sea insesgado.

Definición 2.1 Un estimador $\hat{\theta}$ del parámetro θ es **insesgado** si su valor esperado es θ, i.e.

$$\mathbb{E}(\hat{\theta}) = \theta.$$

A la diferencia $\mathbb{E}(\hat{\theta}) - \theta$ se le llama sesgo del estimador, se identifica con la letra b,

$$b(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

Con la notación y definiciones establecidas, es sencillo demostrar que tanto y como $\bar{y}$ son estimadores insesgados para Y y $\overline{Y}$, respectivamente. Simplemente, re-escribimos

$$\bar{y} = \sum_{i=1}^{n}\frac{y_i}{n} = \sum_{i=1}^{N} Z_i\frac{y_i}{n}.$$

De esta manera, podemos extender la suma sobre todos los elementos de la población, en donde las únicas variables aleatorias son los Z_i's porque las y_i's son cantidades fijas. Entonces, si se selecciona un MASSR de n elementos, las variables $\{Z_1, Z_2, ..., Z_n\}$ son variables aleatorias Bernoulli idénticamente distribuidas con

$$P(Z_i = 1) = \frac{n}{N} \qquad y \qquad P(Z_i = 0) = 1 - \frac{n}{N}.$$

Como consecuencia inmediata, $\mathbb{E}(Z_i) = \frac{n}{N}$ y

$$\mathbb{E}[\bar{y}] = \mathbb{E}\left[\sum_{i=1}^{N} Z_i\frac{y_i}{n}\right] = \sum_{i=1}^{N}\mathbb{E}[Z_i]\frac{y_i}{n} = \sum_{i=1}^{N}\frac{n}{N}\frac{y_i}{n} = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{Y}.$$

Esto muestra que $\bar{y}$ es un estimador insesgado de la media poblacional y consecuentemente y es insesgado del total poblacional Y. Esto es

$$\mathbb{E}[y] = \mathbb{E}[N\bar{y}] = N\mathbb{E}[\bar{y}] = N\overline{Y} = Y.$$

Definamos ahora las varianzas de los estimadores y y $\bar{y}$, y sus respectivas estimaciones.

Varianza Real	Varianza Estimada
$V(y) = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}$	$v(y) = N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}$
$V(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{S^2}{n}$	$v(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{s^2}{n}$

Tabla 2-2.: Estimadores de varianza

El factor $\left(1-\frac{n}{N}\right)$ se denomina **correción por población finita** (cpf). Intuitivamente, esta correción se hace porque con poblaciones pequeñas, cuanto mayor es la fracción de muestreo $\frac{n}{N}$, se tiene más información sobre la población y por tanto menor varianza [Lohr].

En la practica el cpf puede ser ignorado cuando la fración de muestreo no excede del 5 % y para muchos propósitos incluso si es tan alto como el 10 %. El efecto de ignorar la corrección es sobreestimar el error estándar de la estimación $\bar{y}$.[Cochran]

Para calcular la varianza, $V(\bar{y})$, se requiere primero obtener la covarianza entre Z_i y Z_j. Para esto, notemos que $\mathbb{E}\left[Z_i^2\right] = \frac{n}{N}$, luego

$$V[Z_i] = \mathbb{E}\left[Z_i^2\right] - (\mathbb{E}[Z_i])^2 = \frac{n}{N}\left(1-\frac{n}{N}\right). \tag{2.1}$$

Por otro lado, para $i \neq j$

$$\mathbb{E}[Z_iZ_j] = P(Z_j = 1|Z_i = 1)P(Z_i = 1) = \left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right).$$

Así, cuando $i \neq j$, la covarinza de Z_i y Z_j es

$$\begin{aligned} Cov(Z_i, Z_j) &= \mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j] \\ &= \left(\frac{n}{N}\right)\left(\frac{n-1}{N-1}\right) - \left(\frac{n}{N}\right)^2 = -\frac{n}{N(N-1)}\left(1-\frac{n}{N}\right). \end{aligned} \tag{2.2}$$

Aplicando estos resultados podemos hallar la varianza de $\bar{y}$

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_i\right) \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j Cov(Z_i, Z_j)\right] \\ &= \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j\right] \qquad (2.3) \\ &= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left[(N-1)\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2 + \sum_{i=1}^{N} y_i^2\right] \qquad (2.4) \\ &= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right] \qquad (2.5) \\ &= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left[N\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2\right] \qquad (2.6) \\ &= \left(1-\frac{n}{N}\right)\frac{S^2}{n}. \qquad (2.7) \end{aligned}$$

La igualdad (2.3) resulta de sustituir (2.1) y (2.2), y factorizar el término común. Luego (2.4) resulta de aplicar y simplificar la expresión: $\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j = \left(\sum_{i=1}^{N} y_i\right)^2 - \sum_{i=1}^{N} y_i^2$. Por otra parte, la varianza de y se deduce empleando (2.7). Es decir,

$$V(y) = N^2 V(\bar{y}) = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}.$$

Para mostrar que los estimadores $\hat{v}(\bar{y})$ y $\hat{v}(y)$ son insesgados, necesitamos probar que $\mathbb{E}[s^2] = S^2$, es decir, que s^2 es insesgado de la varianza poblacional. Obsérvese que s^2 puede reescri-

birse como

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i\in S}\left[(y_i-\overline{Y})-(\bar{y}-\overline{Y})\right]^2 \\ &= \frac{1}{n-1}\left[\sum_{i\in S}(y_i-\overline{Y})^2 - n(\bar{y}-\overline{Y})^2\right]. \end{aligned}$$

Esta forma facilita el calculo de la esperanza, pues

$$\mathbb{E}\left[\sum_{i\in S}(y_i-\overline{Y})^2\right] = \mathbb{E}\left[\sum_{i=1}^{N} Z_i(y_i-\overline{Y})^2\right] = \frac{n}{N}\sum_{i=1}^{N}(y_i-\overline{Y})^2 = \frac{n(N-1)}{N}S^2.$$

Mientras que

$$\mathbb{E}\left[n(\bar{y}-\overline{Y})^2\right] = nV(\bar{y}) = \left(1-\frac{n}{N}\right)S^2 = \frac{N-n}{N}S^2.$$

Combinando estos resultados se deduce que $\mathbb{E}[s^2] = S^2$ y por lo tanto que los estimadores $v(\bar{y})$ y $v(y)$ son insesgados de las varianzas para los estimadores poblacionales.

$$\begin{aligned} \mathbb{E}[s^2] &= \frac{1}{n-1}\left[\mathbb{E}[\sum_{i\in S}(y_i-\overline{Y})^2] - \mathbb{E}\left[n(\bar{y}-\overline{Y})^2\right]\right] \\ &= \frac{1}{n-1}\left[\frac{n(N-1)}{N}S^2 - \frac{N-n}{N}S^2\right] \\ &= S^2 \end{aligned} \tag{2.8}$$

2.1.4. ESTIMADOR DE RAZÓN

En la tabla (**2-1**) se define a la razón como un cociente entre el valor medio de la variable de interés y la media de otra variable auxiliar. Al ser la razón una función no lineal

$$r = f(\bar{y},\bar{x}) = \frac{\bar{y}}{\bar{x}},$$

para calcular su varianza, primero se debe linealizar. Para ello, se hace una aproximación por series de Taylor, de primer orden, al rededor del punto $\left(\overline{Y},\overline{X}\right)$.

$$r = f(\bar{y},\bar{x}) \approx f\left(\overline{Y},\overline{X}\right) + \left(\partial f_{\bar{y}}|_{\overline{Y},\overline{X}}\right)\left(\bar{y}-\overline{Y}\right) + \left(\partial f_{\bar{x}}|_{\overline{Y},\overline{X}}\right)\left(\bar{x}-\overline{X}\right),$$

Al calcular las derivadas parciales, se tiene que

$$\partial f_{\bar{y}}|_{\overline{Y},\overline{X}} = \frac{1}{\overline{X}} \qquad \partial f_{\bar{x}}|_{\overline{Y},\overline{X}} = -\frac{\overline{Y}}{\overline{X}^2} = -\frac{R}{\overline{X}} \qquad f\left(\overline{Y},\overline{X}\right) = \frac{\overline{Y}}{\overline{X}} = R.$$

Por lo tanto, al sustituir estos valores en la expresión anterior se obtiene

$$r \approx R + \frac{1}{\overline{X}}\left(\bar{y} - R\bar{x}\right) \qquad \Longrightarrow \qquad r - R \approx \frac{1}{\overline{X}}\left(\bar{y} - R\bar{x}\right).$$

donde R y $\overline{X}$ son constantes. Consecuentemente

$$\mathbb{E}\left[r\right] \approx R + \frac{1}{\overline{X}}\mathbb{E}\left[\bar{y} - R\bar{x}\right] = R$$

ya que

$$\begin{aligned}\frac{1}{\overline{X}}\mathbb{E}\left[\bar{y} - R\bar{x}\right] &= \frac{1}{\overline{X}}\left(\overline{Y} - R\overline{X}\right) \\ &= \frac{\overline{Y}}{\overline{X}} - R = 0\end{aligned}$$

Además, $\bar{y} - R\bar{x}$ puede reescribirse como

$$\bar{y} - R\bar{x} = \frac{1}{n}\sum_{i\in\mathcal{S}}\left(y_i - Rx_i\right) = \frac{1}{n}\sum_{i\in\mathcal{S}} d_i = \bar{d},$$

donde $d_i = y_i - Rx_i$ es una nueva variable tal que su media poblacional es $\overline{D} = 0$. Por lo tanto,

$$\begin{aligned}V\left(r\right) \approx \mathbb{E}\left[\left(r - R\right)^2\right] &= \frac{1}{\overline{X}^2}\mathbb{E}\left[\left(\bar{y} - R\bar{x}\right)^2\right] \\ &= \frac{1}{\overline{X}^2}\mathbb{E}\left[\left(\bar{d} - \overline{D}\right)^2\right] \\ &= \frac{1}{\overline{X}^2}V(\bar{d}).\end{aligned}$$

Entonces, aplicando la definición (2.7) a $\bar{d}$ se deduce que la varianza aproximada es:

$$V\left(r\right) \approx \left(1 - \frac{n}{N}\right)\frac{S_d^2}{n\overline{X}^2}. \tag{2.9}$$

con $S_d^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(d_i - \overline{D})^2$. Por otra parte, un estimador insesgado para $V\left(r\right)$ es

$$v(r) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n\bar{x}^2} \qquad \text{con} \qquad s_d^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\hat{d}_i - \hat{\bar{d}}\right)^2 \tag{2.10}$$

donde $\bar{x}$ es la media muestral y $\hat{\bar{d}}$ es la media de la variable $\hat{d}_i = y_i - rx_i$. La verificación de que $v(r)$ es insesgada, resulta de $\mathbb{E}[s_d^2] = S_d^2$, la prueba de esto es análoga a (2.8).
Adicionalmente, es importante mencionar que r es un estimador sesgado de la razón poblacional, R, pues tiene un sesgo de orden $1/n$ debido a la linealización. En la práctica, este valor generalmente no es importante en muestras de tamaño moderado [Cochran].

2.1.5. INTERVALOS DE CONFIANZA

Un intervalo de confianza es un rango de valores con el que estimamos al parámetro de la población. Se obtienen, primero utilizando variables aleatorias y probabilidades, pero cuando se tienen las estimaciones, se sustituyen en lugar de las variables aleatorias y por tal motivo no podemos hablar de probabilidades, sino de confianza que es un argumento basado en frecuencias. Esta confianza se fija de acuerdo a los intereses del investigador o del estudio en cuestión. Por lo general, los intervalos suelen construirse con una confianza del 95 %. Esto significa que si tomamos muestras de nuestra población una y otra vez, y construimos intervalos de confianza de cada muestra, esperamos que el 95 % de los intervalos resultantes contengan el valor real del parámetro poblacional.

Para la construcción de los intervalos hacemos uso del Teorema central del Limite para MASSR probado por Hájek en 1960. En términos prácticos, esta nos dice que bajo ciertas condiciones, y si n, N, y $N - n$ son todos suficientemente grandes, entonces

$$Z = \frac{\bar{y} - \overline{Y}}{\sqrt{V(\bar{y})}} \sim N(0, 1). \tag{2.11}$$

De lo anterior, se sigue que

$$P(-z_{\alpha/2} < \frac{\bar{y} - \overline{Y}}{\sqrt{V(\bar{y})}} < z_{\alpha/2}) = 1 - \alpha,$$

en donde $z_{\alpha/2}$ es el percentil $(1 - \alpha/2)$ de la distribución normal estándar, $\alpha \in [0, 1]$.
Aquí, estaríamos viendo a $\bar{y}$ como variable aleatoria (su valor depende de la muestra aleatoria) y por lo tanto las probabilidades son válidas. Pero al insertar los valores observados en la muestra y obtener un valor fijo (la estimación), tenemos que usar el argumento de frecuencias/confianza.
Así, un intervalo de confianza del $100 \times (1 - \alpha)\,\%$ para la media, el total y la razón es:

$$\text{Media:} \qquad \left[\bar{y} - z_{\alpha/2}\sqrt{V(\bar{y})}, \bar{y} + z_{\alpha/2}\sqrt{V(\bar{y})}\right]$$

$$\text{Total:} \qquad \left[y - z_{\alpha/2} N\sqrt{V(\bar{y})}, y + z_{\alpha/2} N\sqrt{V(\bar{y})}\right]$$

$$\text{Razón:} \qquad \left[r - z_{\alpha/2}\sqrt{V(r)}, r + z_{\alpha/2}\sqrt{V(r)}\right]$$

Usualmente las varianzas reales se desconocen, por lo que para calcular los intervalos se emplean sus estimadores, basta sustituir en las expresiones anteriores el estimador correspondiente. En la practica, los niveles de confianza más empleados son: 90 %, 95 % y 99 %, con valores α igual a 0.05, 0.025 y 0.005 respectivamente. Además, a menudo se substituye $z_{\alpha/2}$ por $t_{\alpha/2,n-1}$, el percentil $(1-\alpha/2)$ de una distribución t con $n-1$ grados de libertad. Para muestras grandes $t_{\alpha/2,n-1} \approx z_{\alpha/2}$.

Elegir el tamaño de muestra necesario para que la aproximación normal sea adecuada no es fácil, sin embargo, algunos autores proponen ciertas reglas. Por ejemplo, Sugden et al. (2000) recomiendan un tamaño mínimo de

$$n = 28 + 25\left(\frac{\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^3}{NS^3}\right) \tag{2.12}$$

$\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^3$ es la asimetría de la población; si la asimetría es grande, se necesita un tamaño de muestra grande para que la aproximación normal sea válida. Cabe resaltar que el número usual $n = 30$ citado como un número suficientemente grande a menudo no es suficiente en problemas de muestreo de poblaciones finitas [Lohr].

2.1.6. TAMAÑO DE MUESTRA PARA ALCANZAR UNA PRECISIÓN DADA

La precisión o margen de error deseado, a menudo se expresa en términos absolutos ligado a una confianza del $100 \times (1-\alpha)\,\%$, como

$$\begin{aligned}
1-\alpha &= P\left(|\bar{y} - \overline{Y}| \leq \epsilon\right), \\
&= P\left(-\epsilon \leq \bar{y} - \overline{Y} \leq \epsilon\right), \\
&= P\left(-\frac{\epsilon}{V(\bar{y})} \leq \frac{\bar{y} - \overline{Y}}{V(\bar{y})} \leq \frac{\epsilon}{V(\bar{y})}\right), \\
&= P\left(-\frac{\epsilon}{V(\bar{y})} \leq Z \leq \frac{\epsilon}{V(\bar{y})}\right), \\
&= F_Z\left(\frac{\epsilon}{V(\bar{y})}\right) - F_Z\left(-\frac{\epsilon}{V(\bar{y})}\right), \\
&= 2F_Z\left(\frac{\epsilon}{V(\bar{y})}\right) - 1,
\end{aligned}$$

en donde ϵ es la precisión o margen de error, la aproximación normal se obtiene de (2.11) y F_Z es la función de distribución de una $N(0,1)$. Despejando de la última expresión, se tiene que

$$F_Z\left(\frac{\epsilon}{V(\bar{y})}\right) = 1-\alpha/2 \Longleftrightarrow \frac{\epsilon}{V(\bar{y})} = z_{\alpha/2}.$$

De manera que,

$$\epsilon = z_{\alpha/2}\sqrt{V(\bar{y})} = z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}}.$$

Entonces, si se conoce la desviación estándar (S), el error deseado se alcanza con un tamaño de muestra igual a

$$n = \frac{z^2_{\alpha/2}S^2}{\epsilon^2 + \frac{z^2_{\alpha/2}S^2}{N}}, \tag{2.13}$$

y ese error se observará con una confianza del $100\times(1-\alpha)\,\%$.
Análogamente, para el estimador de razón

$$\epsilon = z_{\alpha/2}\sqrt{V(r)} = z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{S_d^2}{n\overline{X}^2}}.$$

Luego si S_d^2 es conocida, el tamaño de muestra necesario para alcanzar un error ϵ resulta de despejar n de la ecuación anterior:

$$n = \frac{Nz^2_{\alpha/2}S_d^2}{\epsilon^2\overline{X}^2 + z^2_{\alpha/2}S_d^2}. \tag{2.14}$$

2.2. MUESTREO ALEATORIO ESTRATIFICADO

En un muestreo estratificado, la población total de N elementos es dividida en H subpoblaciones llamados **estratos**, de $N_1, N_2, ..., N_H$ elementos, respectivamente. Los estratos no se traslapan, y su unión constituye la población total, de modo que cada unidad de muestreo pertenece exactamente a un estrato. Para poder trabajar correctamente con un muestreo estratificado se debe conocer previamente los elementos de la población y el número de ellos que pertenece a cada estrato, es decir, N_h, con $h = 1, 2, ..., H$.

En cada estrato se toman muestras independientes de n_h elementos seleccionados aleatoriamente. Si esta selección se hace vía MAS, entonces a todo el proceso se le conoce como **Muestreo Aleatorio Estratificado**. Se define $\mathcal{S}_h$ como el conjunto de los n_h elementos

tomados por MASSR en el estrato h. El tamaño total de la muestra es

$$n = \sum_{h=1}^{H} n_h.$$

NOTACIÓN

El subíndice h denota al estrato y j el elemento dentro de este. Todos los siguientes símbolos refieren a un estrato h.

N_h	número total de elementos
n_h	número de elementos en la muestra
y_{hj}	valor del j-ésimo elemento
$Y_h = \sum_{j=1}^{N_h} y_{hj}$	total
$\overline{Y}_h = \frac{Y_h}{N_h}$	media
$S_h^2 = \sum_{j=1}^{N_h} \frac{\left(y_{hj} - \overline{Y}_h\right)^2}{N_h - 1}$	varianza

Los correspondientes valores muestrales al tomar un MASSR en cada estrato son:

$$\overline{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}, \qquad y_h = \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} = N_h \bar{y}_h, \qquad s_h^2 = \sum_{j \in \mathcal{S}_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}.$$

En la siguiente tabla se presentan los parámetros poblacionales por estratificación, así como su estimador (el subíndice e denota estratificación).

	Población	Estimador
Total:	$Y_e = \sum_{h=1}^{H} Y_h = \sum_{h=1}^{H} N_h \overline{Y}_h$	$y_e = \sum_{h=1}^{H} y_h = \sum_{h=1}^{H} N_h \bar{y}_h$
Media:	$\overline{Y}_e = \frac{Y_e}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{Y}_h$	$\bar{y}_e = \frac{y_e}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$

Tabla 2-3.: Estimadores por estratificación

El valor $\frac{N_h}{N}$ representa la proporción de elementos de la población en el estrato h.

2.2.1. PROPIEDADES DE LOS ESTIMADORES

Las propiedades de los estimadores se siguen directamente de las propiedades del MASSR:

Insesgamiento: como en cada estrato se ha tomado un MASSR, entonces $\bar{y}_h$ es un estimador insesgado de la media real del estrato, es decir, $\mathbb{E}[\bar{y}_h] = \overline{Y}_h$. Consecuentemente $\bar{y}_e$ es insegado de la media pobalcional $\overline{Y}_e$.

$$\mathbb{E}\left[\,\bar{y}_e\,\right] = \mathbb{E}\left[\sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h\right] = \sum_{h=1}^{H} \frac{N_h}{N}\mathbb{E}[\bar{y}_h] = \sum_{h=1}^{H} \frac{N_h}{N}\overline{Y}_h = \overline{Y}_e.$$

Por las mismas razones que para la media poblacional, y_e es un estimador insesgado para el total poblacional

$$\mathbb{E}\left[\,y_e\,\right] = \mathbb{E}\left[\sum_{h=1}^{H} N_h\bar{y}_h\right] = \sum_{h=1}^{H} N_h\mathbb{E}[\bar{y}_h] = \sum_{h=1}^{H} N_h\overline{Y}_h = Y_e.$$

Varianza: Usando las propiedades de un MASSR para la $V(y_h)$ y del hecho que las muestras son independientes de los estratos, se deducen que

$$V\left(y_e\right) = \sum_{h=1}^{H} V\left(N_h\bar{y}_h\right) = \sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}. \tag{2.15}$$

con $S_h^2 = \dfrac{1}{N_h - 1}\displaystyle\sum_{i=1}^{N_h}\left(y_{hj} - \overline{Y}_h\right)^2$. Luego, un estimador de la varianza esta dada por:

$$v\left(y_e\right) = \sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h} \qquad \text{con} \qquad s_h^2 = \frac{1}{n_h - 1}\sum_{j\in\mathcal{S}_h}\left(y_{hj} - \bar{y}_h\right)^2. \tag{2.16}$$

Así, la varianza, real y estimada, para la media es:

$$V(\bar{y}_e) = \frac{1}{N^2}V(y_e) \qquad \text{y} \qquad v(\bar{y}_e) = \frac{1}{N^2}v(y_e).$$

Notar que, $v\left(y_e\right)$ es un estimador insegado ya que s_h^2 es insesgado para S_h^2. Esto último se sigue del muestreo aleatorio simple tomado en cada estrato. Consecuentemente, $v(\bar{y}_e)$ es insegado para $V(\bar{y}_e)$ [Lohr].

2.2.2. ESTIMADOR DE RAZÓN

En un diseño muestral por muestreo aleatorio estratificado existen dos métodos que normalmente son empleados para estimar la razón entre dos variables: **estimador separado** r_s y **estimador combinado** r_c.

PROPUESTA DE ESTIMADOR SEPARADO

En este caso, la estimación del total de la variable y es

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} X_h \frac{\bar{y}_h}{\bar{x}_h}, \tag{2.17}$$

$\bar{y}_h$ y $\bar{x}_h$ son las medias estimadas de cada variable en el estrato h, con $h = 1, 2, ..., H$. X_h es el total en el estrato h (debe ser conocido previamente). Luego el estimador de razón poblacional es

$$r_s = \frac{\hat{t}_y}{X}. \tag{2.18}$$

donde $X = \sum_{h=1}^{H} X_h$ es el total de la variable auxiliar y debe ser conocida previamente. La esperanza de este estimador depende fuertemente de la razón en cada estrato, $r_h = \dfrac{\bar{y}_h}{\bar{x}_h}$. Esto es,

$$\mathbb{E}[r_s] = \mathbb{E}\left[\frac{\hat{t}_y}{X}\right] = \frac{1}{X}\sum_{h=1}^{H} X_h \underbrace{\mathbb{E}\left[\frac{\bar{y}_h}{\bar{x}_h}\right]}_{*}$$

Nótese que en este caso X y X_h son constantes conocidas. Por otra parte, la expresión marcada como $*$ corresponde a la esperanza del estimador de razón para un muestreo aleatorio simple (Sección 2.1.4) en el estrato h, que en general es sesgado. Para notar esta característica, considérese la covarianza, en muestreo aleatorio simple de tamaño n_h en el h-ésimo estrato, de las cantidades r_h y $\bar{x}_h$. Se tiene

$$\begin{aligned} Cov(r_h, \bar{x}_h) &= \mathbb{E}[\frac{\bar{y}_h}{\bar{x}_h}\bar{x}_h] - \mathbb{E}[r_h]\mathbb{E}[\bar{x}_h] \\ &= \overline{Y}_h - \overline{X}_h\mathbb{E}[r_h], \end{aligned}$$

donde $\overline{Y}_h$ y $\overline{X}_h$ son las medias poblacionales en el estrato h. Por tanto,

$$\mathbb{E}[r_h] = \frac{\overline{Y}_h}{\overline{X}_h} - \frac{1}{\overline{X}_h} Cov(r_h, \bar{x}_h) = R_h - \frac{1}{\overline{X}_h} Cov(r_h, \bar{x}_h).$$

Así, el sesgo exacto en r_h es $-\frac{1}{\overline{X}_h} Cov(r_h, \bar{x}_h)$, con $R_h = \dfrac{\overline{Y}_h}{\overline{X}_h}$. En consecuencia

$$\begin{aligned}
\mathbb{E}[r_s] &= \frac{1}{X} \sum_{h=1}^{H} X_h \left[R_h - \frac{1}{\overline{X}_h} Cov(r_h, \bar{x}_h) \right] \\
&= \frac{1}{X} \sum_{h=1}^{H} Y_h - \frac{1}{X} \sum_{h=1}^{H} N_h Cov(r_h, \bar{x}_h) \\
&= R - \frac{1}{X} \sum_{h=1}^{H} N_h Cov(r_h, \bar{x}_h)
\end{aligned}$$

Esto muestra que la expresión (2.18) es un estimador sesgado de la razón poblacional, con sesgo igual a

$$B(r_s) = -\frac{1}{X} \sum_{h=1}^{H} N_h Cov(r_h, \bar{x}_h). \tag{2.19}$$

La varianza de (2.18) depende totalmente de la varianza de $\hat{t}_y$, la cual a su vez depende de las varianzas de las razones estimadas en cada estrato, $r_h = \dfrac{\bar{y}_h}{\bar{x}_h}$ (dadas por MAS). Así,

$$\begin{aligned}
V\left(\hat{t}_y\right) = \sum_{h=1}^{H} X_h^2 V\left(\frac{\bar{y}_h}{\bar{x}_h}\right) &= \sum_{h=1}^{H} X_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{dh}^2}{n_h \overline{X}_h^2} \\
&= \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{dh}^2}{n_h}.
\end{aligned} \tag{2.20}$$

Por lo tanto,

$$V\left(r_s\right) = \frac{1}{X^2} V\left[\hat{t}_y\right] = \sum_{h=1}^{H} \left(\frac{N_h}{X}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{dh}^2}{n_h}, \tag{2.21}$$

donde

$$S_{dh}^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}\left(d_{hj} - \overline{D}_h\right)^2 \quad \text{con} \quad \overline{D}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} d_{hj} \quad \text{y} \quad d_{hj} = y_{hj} - R_h x_{hj},$$

con varianza estimada dada por

$$v(r_s) = \sum_{h=1}^{H}\left(\frac{N_h}{X}\right)^2\left(1 - \frac{n_h}{N_h}\right)\frac{s_{dh}^2}{n_h}. \tag{2.22}$$

En esta expresión

$$s_{dh}^2 = \frac{1}{n_h - 1}\sum_{j\in\mathcal{S}_h}\left(\hat{d}_{hj} - \bar{d}_h\right)^2 \quad \text{con} \quad \bar{d}_h = \frac{1}{n_h}\sum_{j\in\mathcal{S}_h}\hat{d}_{hj} \quad \text{y} \quad \hat{d}_{hj} = y_{hj} - r_h x_{hj}.$$

$R_h = \dfrac{\overline{Y}_h}{\overline{X}_h}$ es la razón real en el estrato h y r_h el estimador. EL insesgamiento de s_{dh}^2 se sigue de las propiedades del muestreo aleatorio simple en cada estrato, es decir, $\mathbb{E}[s_{dh}^2] = S_{dh}^2$ y en consecuencia $v(r_s)$ es un estimador insesgado para $V(r_s)$.

ESTIMADOR COMBINADO

En este caso, los estimadores para los totales Y_e y X_e son:

$$y_e = \sum_{h=1}^{H} N_h \bar{y}_h \qquad\qquad x_e = \sum_{h=1}^{H} N_h \bar{x}_h.$$

Luego, el estimador combinado para la razón poblacional, $R_c = \dfrac{Y_e}{X_e}$, es

$$r_c = \frac{y_e}{x_e}. \tag{2.23}$$

Para calcular la esperanza de este estimador, considérese la covarianza de las cantidades r_c y x_e, es decir,

$$\begin{aligned} Cov(r_c, x_e) &= \mathbb{E}[\frac{y_e}{x_e}x_e] - \mathbb{E}[r_c]\mathbb{E}[x_e] \\ &= Y_e - X_e\mathbb{E}[r_c], \end{aligned}$$

Por lo tanto,

$$\mathbb{E}[r_c] = \frac{Y_e}{X_e} - \frac{1}{X_e}Cov(r_c, x_e) = R_c - \frac{1}{X_e}Cov(r_c, x_e).$$

Así, la expresión (2.23) representa un estimador sesgado para la razón poblacional, con sesgo

$$B(r_c) = -\frac{1}{X_e}Cov(r_c, x_e). \tag{2.24}$$

El cálculo de la varianza sigue un procedimiento similar al descrito en la sección (2.1.4).

$$\begin{aligned} V(r_c) &\approx \frac{1}{X^2} V(y_e - Rx_e) \\ &= \frac{1}{X^2} V\left(\sum_{h=1}^{H} N_h(\bar{y}_h - R\bar{x}_h)\right) \\ &= \frac{1}{X^2} V\left(\sum_{h=1}^{H} N_h \sum_{j\in\mathcal{S}_h} \frac{1}{n_h}(y_{hj} - Rx_{hj})\right) \\ &= \frac{1}{X^2} V\left(\sum_{h=1}^{H} N_h \bar{d}_h\right) \\ &= \frac{1}{X^2}\left[\sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_{dh}^2}{n_h}\right]. \end{aligned} \tag{2.25}$$

donde $d_{hj} = y_{hj} - Rx_{hj}$, $S_{dh}^2 = \dfrac{1}{N_h - 1}\sum_{j=1}^{N_h}\left(d_{hj} - \overline{D}\right)^2$ y $\overline{D} = \dfrac{1}{N_h}\sum_{j=1}^{N_h} d_{hj}$. Notar que (2.25) se sigue de aplicar el resultado (2.15). Luego, un estimador para ésta varianza esta dada por:

$$v(r_c) = \frac{1}{x_e^2}\left[\sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{s_{dh}^2}{n_h}\right] \tag{2.26}$$

con

$$s_{dh}^2 = \frac{1}{n_h - 1}\sum_{j\in\mathcal{S}_h}\left(\hat{d}_{hj} - \bar{d}_h\right)^2 \quad \text{con} \quad \bar{d}_h = \frac{1}{n_h}\sum_{j\in\mathcal{S}_h}\hat{d}_{hj} \quad \text{y} \quad \hat{d}_{hj} = y_{hj} - r_c x_{hj}.$$

Por las mismas razones que para la razón separada, aquí $\mathbb{E}[s_{dh}^2] = S_{dh}^2$, por tanto (2.26)

es insesgado para la expresión (2.25). A pesar que el estimador combinado tiene menos sesgo cuando el tamaño de la muestra en algunos estratos es pequeño, si las proporciones varían mucho de un estrato a otro, entonces se aprovecha la eficiencia adicional que brinda la estratificación, como lo hace el estimador de razón separado [Lohr].

2.2.3. TAMAÑO DE MUESTRA

Determinar el tamaño de muestra necesario para alcanzar una precisión (ϵ) para cualquiera de los estimadores descritos anteriormente resulta complicado analíticamente, sin embargo, se puede calcular la precisión por considerar diferentes tamaños de muestra n. Por ejemplo para $r_{e.c}$ se tiene que

$$\begin{aligned}\epsilon &= z_{\alpha/2}\sqrt{V\left(r_c\right)} \\ &= \frac{z_{\alpha/2}}{X}\left(\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{dh}^2}{n_h}\right)^{1/2}.\end{aligned}$$

Para calcular estas precisiones se requiere primero, asignar tamaños de muestra n_h a cada estrato, donde estas dependen directamente del tamaño de muestra n. A continuación se presentan las asignaciones más comunes:

Proporcional	Óptima
$n_h = n\dfrac{N_h}{N}$	$n_h = n\dfrac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$.

Para elegir entre estas dos asignaciones se debe tener en cuenta que si las varianzas S_h^2 son aproximadamente iguales en todos los estratos, la asignación proporcional es probablemente la mejor para aumentar la precisión, mientras que si estas varían mucho la asignación optima puede resultar mejor [Lohr].

2.3. MUESTREO POR CONGLOMERADOS

En este tipo de muestreo, la población es dividida en grupos separados, llamados conglomerados, también conocidos como unidades primarias de selección, mientras que los elementos que los componen, unidades secundarias. Entonces, el proceso de muestreo consiste en tomar un muestreo aleatorio simple de clusters. Por lo que, los elementos permitidos en la muestra

de la población serán aquellos que pertenezcan a los conglomerados elegidos.

Al igual que el muestreo aleatorio simple y el muestreo estratificado, el muestreo por conglomerados tiene ventajas y desventajas. Por ejemplo, con el mismo tamaño de muestra, el muestreo por conglomerados generalmente proporciona menos precisión que el muestreo aleatorio simple o el muestreo estratificado. Por otro lado, si los costos de viaje entre grupos son altos, el muestreo de grupos puede ser más rentable que los otros métodos.

NOTACIÓN

El universo $\mathcal{U}$ es la población de N clusters; $\mathcal{S}$ denota la muestra de clusters elegidos de la población, y $\mathcal{S}_h$ la muestra de elementos elegidos del h-ésimo cluster.

$$y_{hj} = \text{medida para el } j\text{-ésimo elemento del } h\text{-ésimo cluster.}$$

Ahora, las cantidades poblacionales se dividen en dos niveles.

NIVEL CONGLOMERADO

$N \equiv$ Número de clusters en la población

$M_h \equiv$ Número de elementos en el conglomerado h, $h = 1, ..., N$

$M_0 = \sum_{h=1}^{N} M_h \equiv$ Número de elementos en la población

$t_h = \sum_{j=1}^{M_h} y_{hj} \equiv$ Total en el cluster h

$t = \sum_{h=1}^{N} t_h = N\bar{t} = \sum_{h=1}^{N}\sum_{j=1}^{M_h} y_{hj} \equiv$ Total poblacional

$S_t^2 = \frac{1}{N-1}\sum_{h=1}^{N}\left(t_h - \frac{t}{N}\right)^2 \equiv$ Varianza poblacional entre totales de conglomerados

NIVEL ELEMENTOS

$$\bar{y}_U = \sum_{h=1}^{N}\sum_{j=1}^{M_h}\frac{y_{hj}}{M_0} = \frac{t}{M_0} \equiv \text{Media poblacional}$$

$$\bar{y}_{hU} = \sum_{j=1}^{M_h}\frac{y_{hj}}{M_h} = \frac{t_h}{M_h} \equiv \text{Media poblacional del cluster } h$$

$$S^2 = \frac{1}{M_0 - 1}\sum_{h=1}^{N}\sum_{j=1}^{M_h}(y_{hj} - \bar{y}_U)^2 \equiv \text{Varianza poblacional de elementos}$$

$$S_h^2 = \frac{1}{M_h - 1}\sum_{j=1}^{M_h}(y_{hj} - \bar{y}_{hU})^2 \equiv \text{Varianza poblacional de elementos del conglomerado } h$$

CANTIDADES MUESTRALES

$n \equiv$ Número de conglomerados en la muestra

$m_h \equiv$ Número de elementos en la muestra del conglomerado h

$$\bar{y}_h = \sum_{j\in\mathcal{S}_h}\frac{y_{hj}}{m_h} \equiv \text{Media muestral del conglomerado } h$$

$$\hat{t}_h = \sum_{j\in\mathcal{S}_h}\frac{M_h}{m_h}y_{hj} = M_h\bar{y}_h \equiv \text{Total estimado para el conglomerado } h$$

$$\hat{t} = \sum_{h\in\mathcal{S}}\frac{N}{n}\hat{t}_h \equiv \text{Estimador insesgado del total poblacional}$$

$$s_t^2 = \frac{1}{n-1}\sum_{h\in\mathcal{S}}\left(\hat{t}_h - \frac{\hat{t}}{N}\right)^2 \equiv \text{Estimador de } S_t^2$$

$$s_h^2 = \frac{1}{m_h - 1}\sum_{j\in\mathcal{S}_h}(y_{hj} - \bar{y}_h)^2 \equiv \text{Varianza muestral en el conglomerado } h.$$

Una vez definida la notación, es importante mencionar que existen diferentes estrategias para seleccionar una muestra de la población. Lo más sencillo es tomar un MAS de los clusters y considerar todos los elementos dentro de estos, a esto se le conoce como muestreo en una etapa. Sin embargo, muchas veces los elementos dentro de un cluster particular pueden ser muy parecidos, por lo que no aportarán mucho a las estimaciones que se deseen realizar. Por tal razón se prefiere trabajar con un muestreo en dos etapas.

2.3.1. MUESTREO EN DOS ETAPAS

Consideremos un muestreo de tipo MAS-MAS. Esta estrategia consiste en seleccionar primero una muestra de conglomerados (primera etapa) y posteriormente submuestrear los elementos (segunda etapa) en cada conglomerado seleccionado, con igual probabilidad. El proceso es el siguiente:

1. Seleccionar un MAS $\mathcal{S}$ de n conglomerados de la población de N.
2. Seleccionar un MAS de m_h elementos de cada conglomerado seleccionado. $\mathcal{S}_h$ denota los elementos seleccionados del h-ésimo conglomerado.

2.3.2. PROPIEDADES DE LOS ESTIMADORES

Insesgamiento. Como en cada cluster se ha seleccionado un MAS, entonces los estimadores son insesgados para las cantidades reales en los cluster. Ademas, por las mismas propiedades se puede probar que los siguientes estimadores son insesgados:

	Real	Estimado
Total:	$t = N\bar{t}$	$\hat{t} = \frac{N}{n}\sum_{h\in\mathcal{S}} \hat{t}_h = N\hat{\bar{t}}_{\mathcal{S}}$
Media :	$\bar{y}_u = \frac{t}{M_0}$	$\hat{\bar{y}}_u = \frac{\hat{t}}{M_0}$

Donde $\hat{\bar{t}}_{\mathcal{S}} = \frac{1}{n}\sum_{h\in\mathcal{S}} \hat{t}_h$ estima el promedio de los totales de los conglomerado. Mas aún, por propiedades del MAS, $\hat{\bar{t}}_{\mathcal{S}}$, es insesgado de $\bar{t}$. Por lo tanto, $\hat{t}$ es insesgado del total poblacional t, consecuentemente $\hat{\bar{y}}_u$ es insesgado de la media poblacional $\bar{y}_u$.

Varianza. En muestreo de dos etapas, los $\hat{t}_h$'s son variables aleatorias. En consecuencia, la varianza de $\hat{t}$ tiene dos componentes:

1. La variabilidad entre los conglomerados
2. La variabilidad de los elementos en los conglomerados

Así, ambas deben ser contempladas al calcular la varianza de estos estimadores. El estimador $\hat{t}$ dado en la tabla anterior se conoce como el estimador de **Horvitz-Thompson** (propuesta en 1952) [Lohr]. Para el cálculo de su varianza se dispone de un teorema general, el cual será

introducido una vez definidos algunos requerimientos previos.

Se requiere definir una variable aleatoria que denotará la inclusión o no de un cluster en la muestra. Sea

$$Z_h = \begin{cases} 1 & \text{si} \quad \text{el conglomerado } h \text{ está en la muestra} \\ \\ 0 & \text{si} \quad \text{el conglomerado } h \text{ no está en la muestra} \end{cases} \tag{2.27}$$

La probabilidad de que el conglomerado h sea incluido en la muestra es

$$\pi_h = P(Z_h = 1) = E(Z_h) = E(Z_h^2) \tag{2.28}$$

y la probabilidad de que el conglomerado h y el conglomerado k ($h \neq k$) sean incluidos en la muestra es

$$\pi_{jk} = P(Z_h = 1|Z_k = 1)P(Z_k = 1) = \mathbb{E}(Z_h Z_k). \tag{2.29}$$

Notar que bajo esta nueva variable, una expresión general para $\hat{t}$ es

$$\hat{t} = \sum_{h=1}^{N} Z_h \frac{\hat{t}_h}{\pi_h}.$$

Teorema 2.1. *(Horvitz–Thompson)[Lohr] Supongamos que el muestreo se realiza en dos etapas de modo que el muestreo en cualquier conglomerado sea independiente del muestreo en cualquier otro conglomerado, y que $\hat{t}$ es independiente de $(Z_1, ..., Z_N)$ con $\mathbb{E}\left[\hat{t}_h\right] = \mathbb{E}\left[\hat{t}_h|Z_1, ..., Z_N\right] = t_h$. Entonces*

$$\mathbb{E}\left[\hat{t}\right] = \mathbb{E}\left[\sum_{h=1}^{N} Z_h \frac{\hat{t}_h}{\pi_h}\right] = \sum_{h=1}^{N} \pi_h \frac{t_h}{\pi_h} = t$$

$$V\left[\hat{t}\right] = V\left[\sum_{h=1}^{N} Z_h \frac{\hat{t}_h}{\pi_h}\right] = V_1 + V_2$$

Donde V_1 representa la variabilidad entre los conglomerado y V_2 la variabilidad de los elementos en los conglomerado. Con

$$V_1 = V\left[\sum_{h=1}^{N} Z_h \frac{t_h}{\pi_h}\right] = \sum_{h=1}^{N}(1-\pi_h)\frac{t_h^2}{\pi_h} + \sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N}(\pi_{hk}-\pi_h\pi_k)\frac{t_h}{\pi_h}\frac{t_k}{\pi_k}. \tag{2.30}$$

$$V_2 = \sum_{h=1}^{N}\frac{V\left(\hat{t}_h\right)}{\pi_h}. \tag{2.31}$$

Ahora, se calcula la varianza del estimador empleando este teorema. Como los conglomerados se han tomado vía un MAS con igual probabilidad de selección, entonces

$$\pi_h = P(Z_h = 1) = \frac{n}{N} \qquad \pi_{hk} = P(Z_h = 1|Z_k = 1)P(Z_k = 1) = \frac{n}{N}\frac{n-1}{N-1}$$

Por lo que,

$$\hat{t} = \sum_{h\in\mathcal{S}}\frac{N}{n}\hat{t}_h = \sum_{h=1}^{N} Z_h\frac{N}{n}\hat{t}_h = \sum_{h=1}^{N} Z_h\frac{\hat{t}_h}{\pi_h}.$$

Así, aplicando el teorema anterior se tiene que

$$\mathbb{E}\left[\hat{t}\right] = \sum_{h=1}^{N}\pi_h\frac{t_h}{\pi_h} = t.$$

Para calcular V_1 se sigue un procedimiento similar al desarrollado para la varianza de $\bar{y}$ en un MAS.

$$\begin{aligned}
V_1 &= \sum_{h=1}^{N}(1-\pi_h)\frac{t_h^2}{\pi_h} + \sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N}(\pi_{hk}-\pi_h\pi_k)\frac{t_h}{\pi_h}\frac{t_k}{\pi_k}\\
&= \sum_{h=1}^{N}\left(1-\frac{n}{N}\right)\left(\frac{N}{n}\right)t^2 + \sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N}\left[\frac{n}{N}\frac{n-1}{N-1} - \left(\frac{n}{N}\right)^2\right]\left(\frac{N}{n}\right)^2 t_h t_k\\
&= \left(\frac{N}{n}\right)\left(1-\frac{n}{N}\right)\left[\sum_{h=1}^{N} t_h^2 - \frac{1}{N-1}\sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N} t_h t_k\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{N}{n(N-1)}\left(1-\frac{n}{N}\right)\left[(N-1)\sum_{h=1}^{N} t_h^2 - \sum_{h=1}^{N}\sum_{k=1}^{N} t_h t_k + \sum_{h=1}^{N} t_h^2\right] \\
&= \frac{N}{n(N-1)}\left(1-\frac{n}{N}\right)\left[N\sum_{h=1}^{N} t_h^2 - \left(\sum_{h=1}^{N} t_h\right)\left(\sum_{k=1}^{N} t_k\right)\right] \\
&= \frac{N}{n(N-1)}\left(1-\frac{n}{N}\right)\left[N\sum_{h=1}^{N} t_h^2 - t^2\right] \\
&= \frac{N}{n(N-1)}\left(1-\frac{n}{N}\right)\left[N\sum_{h=1}^{N}\left(t_h - \frac{t}{N}\right)^2\right] \\
&= N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}. \qquad (2.32)
\end{aligned}
$$

Del muestreo aleatorio simple se tiene que

$$V\left(\hat{t}_h\right) = M_h^2\left(1-\frac{m_h}{M_h}\right)\frac{S_h^2}{m_h}.$$

Así, al aplicar (2.31) con $\pi_h = \dfrac{n}{N}$

$$V_2 = \frac{N}{n}\sum_{h=1}^{N} M_h^2\left(1-\frac{m_h}{M_h}\right)\frac{S_h^2}{m_h}.$$

Por lo tanto, la varianza de $\hat{t}$ es

$$V\left(\hat{t}\right) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{h=1}^{N} M_h^2\left(1-\frac{m_h}{M_h}\right)\frac{S_h^2}{m_h}. \qquad (2.33)$$

Donde S_t^2 es la varianza poblacional del total de conglomerados, y S_h^2 es la varianza entre los elementos dentro del conglomerado h.

Un estimador insesgado de (2.33) está dada por

$$\hat{V}\left(\hat{t}\right) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{h\in\mathcal{S}} M_h^2\left(1-\frac{m_h}{M_h}\right)\frac{s_h^2}{m_h}. \qquad (2.34)$$

Teorema 2.2. *[Lohr] Supongamos que las condiciones del teorema* (2.1) *se mantienen, y*

que $\hat{V}(\hat{t}_h)$ *es un estimador insesgado de* $V(\hat{t})$ *que es independiente de* Z_i*. Entonces,*

$$\mathbb{E}\left[\sum_{h=1}^{N} Z_h \frac{\hat{V}(\hat{t}_h)}{\pi_h^2}\right] = V_2 \tag{2.35}$$

$$\mathbb{E}\left[\sum_{h=1}^{N} Z_h(1-\pi_h)\frac{\hat{t}_h^2}{\pi_h^2} + \sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N} Z_h Z_k \frac{(\pi_{hk}-\pi_i\pi_k)}{\pi_{hk}}\frac{\hat{t}_h}{\pi_h}\frac{\hat{t}_k}{\pi_k}\right] = V_1 + \sum_{h=1}^{N}(1-\pi_h)\frac{V(\hat{t}_h)}{\pi_h}. \tag{2.36}$$

Para mostrar que (2.34) es insesgado hacemos uso de este teorema. Sustituyendo $\frac{N}{n}$ por π_h y $\frac{n(n-1)}{N(N-1)}$ por π_{hk},

$$\begin{aligned}
&\sum_{h=1}^{N} Z_h(1-\pi_h)\frac{\hat{t}_h^2}{\pi_h^2} + \sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N} Z_h Z_k \frac{(\pi_{hk}-\pi_h\pi_k)}{\pi_{hk}}\frac{\hat{t}_h}{\pi_h}\frac{\hat{t}_k}{\pi_k}\\
&\qquad = \left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right)\sum_{h=1}^{N} Z_h\hat{t}_h^2 + \left(\frac{N}{n}\right)^2\left[1-\frac{n(N-1)}{N(n-1)}\right]\sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N} Z_h Z_k\hat{t}_h\hat{t}_k\\
&\qquad = \left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right)\sum_{h=1}^{N} Z_h\hat{t}_h^2 + \left(\frac{N}{n}\right)^2\frac{1}{n-1}\left(1-\frac{n}{N}\right)\sum_{h=1}^{N}\sum_{\substack{k=1\\k\neq h}}^{N} Z_h Z_k\hat{t}_h\hat{t}_k\\
&\qquad = \left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right)\left(\sum_{h=1}^{N} Z_h\hat{t}_h^2 - \frac{1}{n-1}\sum_{h=1}^{N}\sum_{k=1}^{N} Z_h Z_k\hat{t}_h\hat{t}_k + \frac{1}{n-1}\sum_{h=1}^{N} Z_h\hat{t}_h^2\right)\\
&\qquad = \left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left[n\sum_{h=1}^{N} Z_h\hat{t}_h^2 - \left(\sum_{h=1}^{N} Z_h\hat{t}_h\right)^2\right]\\
&\qquad = \left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left[n\sum_{h=1}^{N} Z_h\hat{t}_h^2 - \frac{n^2}{N^2}\left(\frac{N}{n}\sum_{h=1}^{N} Z_h\hat{t}_h\right)^2\right]\\
&\qquad = \left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right)\frac{1}{n-1}\left[n\sum_{h=1}^{N} Z_h\hat{t}_h^2 - n^2\left(\frac{\hat{t}}{N}\right)^2\right]\\
&\qquad = \left(\frac{N}{n}\right)^2\left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{h=1}^{N} Z_h\left(\hat{t}_h - \frac{\hat{t}}{N}\right)^2\\
&\qquad = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}.
\end{aligned}$$

Así, aplicando (2.36) se tiene que

$$\mathbb{E}\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}\right] = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{N} V\left(\hat{t}_h\right). \tag{2.37}$$

Un estimador insesgado de $V\left(\hat{t}_h\right)$ es

$$\hat{V}\left(\hat{t}_h\right) = M_h^2(1-\frac{m_h}{M_h})\frac{s_h^2}{m_h}.$$

Luego por (2.35),

$$\mathbb{E}\left[\sum_{h=1}^{N} Z_h\left(\frac{N}{n}\right)^2 \hat{V}\left(\hat{t}_h\right)\right] = V_2 = \frac{N}{n}\sum_{h=1}^{N} V\left(\hat{t}_h\right).$$

Empleando estos resultados se puede probar que (2.34) es un estimador insesgado de (2.33),

$$\begin{aligned}
\mathbb{E}\left[\hat{V}\left(\hat{t}\right)\right] &= \mathbb{E}\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{h\in\mathcal{S}}\hat{V}\left(\hat{t}_h\right)\right] \\
&= \mathbb{E}\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}\right] + \mathbb{E}\left[\frac{N}{n}\sum_{h\in\mathcal{S}}\hat{V}\left(\hat{t}_h\right)\right] \qquad (2.38) \\
&= N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{N} V\left(\hat{t}_h\right) + \sum_{h=1}^{N} V\left(\hat{t}_h\right) \qquad (2.39) \\
&= N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{h=1}^{N} V\left(\hat{t}_h\right) \\
&= N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{h=1}^{N} M_h^2(1-\frac{m_h}{M_h})\frac{S_h^2}{m_h}.
\end{aligned}$$

Para el cálculo de la segunda expresión en (2.38) se hace uso de la propiedad de condicionamiento sucesivo del valor esperado condicional,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}(Y|X)].$$

Entonces,

$$\begin{aligned}\mathbb{E}\left[\frac{N}{n}\sum_{h\in\mathcal{S}}\hat{V}\left(\hat{t}_i\right)\right] &= \mathbb{E}\left[\frac{N}{n}\sum_{h=1}^{N}Z_h\hat{V}\left(\hat{t}_h\right)\right]\\ &= \mathbb{E}\left\{\mathbb{E}\left[\frac{N}{n}\sum_{h=1}^{N}Z_h\hat{V}\left(\hat{t}_h\right)\Big|Z_h,...,Z_N\right]\right\}\\ &= \mathbb{E}\left[\frac{N}{n}\sum_{h=1}^{N}Z_hV\left(\hat{t}_h\right)\right] = \frac{N}{n}\sum_{h=1}^{N}\mathbb{E}[Z_h]V\left(\hat{t}_h\right)\\ &= \sum_{h=1}^{N}V\left(\hat{t}_h\right).\end{aligned}$$

Esto completa la demostración de el insesgamiento del estimador (2.34). Por otra parte, si se conoce el número total de elementos en la población (M_0), un estimador insesgado para la media poblacional es $\hat{\bar{y}} = \dfrac{\hat{t}}{M_0}$, con varianza estimada

$$\hat{V}\left(\hat{\bar{y}}\right) = \frac{1}{M_0^2}\hat{V}\left(\hat{t}\right) = \frac{1}{M_0^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{h=1}^{N}M_h^2\left(1-\frac{m_h}{M_h}\right)\frac{S_h^2}{m_h}\right]. \tag{2.40}$$

La insesgamiento de ésta última expresión se sigue de la insesgamiento de la varianza del total [Lohr].

2.3.3. ESTIMADOR DE RAZÓN

El estimador del total de la variable y y una variable auxiliar x son las siguientes

$$\hat{t}_y = \frac{N}{n}\sum_{h\in\mathcal{S}}\hat{t}_{yh} \qquad \text{y} \qquad \hat{t}_x = \frac{N}{n}\sum_{h\in\mathcal{S}}\hat{t}_{xh}.$$

Por lo tanto, el estimador de razón es

$$r = \frac{\hat{t}_y}{\hat{t}_x} = \frac{\sum_{h\in\mathcal{S}}\hat{t}_{yh}}{\sum_{h\in\mathcal{S}}\hat{t}_{xh}} = \frac{\sum_{h\in\mathcal{S}}M_h\bar{y}_h}{\sum_{h\in\mathcal{S}}M_h\bar{x}_h}, \tag{2.41}$$

con $\bar{y}_h$ y $\bar{x}_h$ medias muestrales en el conglomerado h. La varianza de este estimador debe contemplar la varianza entre los conglomerados y dentro de cada conglomerado. Entonces,

usando los argumentos de la sección 2.1.4 tenemos que

$$\begin{aligned}
V(r) &= \frac{1}{t_x^2} V\left[\hat{t}_y - R\hat{t}_x\right] \\
&= \frac{1}{t_x^2} V\left[\frac{N}{n}\sum_{h\in\mathcal{S}} \hat{t}_{hy} - R\frac{N}{n}\sum_{h\in\mathcal{S}} \hat{t}_{hx}\right] \\
&= \frac{1}{t_x^2} V\left[\frac{N}{n}\sum_{h\in\mathcal{S}} \left(\hat{t}_{hy} - R\hat{t}_{hx}\right)\right] \\
&= \frac{1}{t_x^2} V\left[\frac{N}{n}\sum_{h\in\mathcal{S}} \frac{M_i}{m_i}\sum_{j\in\mathcal{S}_h} \left(y_{hj} - Rx_{hj}\right)\right] \\
&= \frac{1}{t_x^2} V\left[\frac{N}{n}\sum_{h\in\mathcal{S}} \frac{M_i}{m_i}\sum_{j\in\mathcal{S}_h} d_{hj}\right] \\
&= \frac{1}{t_x^2} V\left[\frac{N}{n}\sum_{h\in\mathcal{S}} \hat{t}_{hd}\right] \\
&= \frac{1}{t_x^2} V\left[\hat{t}_d\right]
\end{aligned}$$

donde $d_{hj} = y_{hj} - Rx_{hj}$ y t_x es el total real de la variable x y R la razón real. Además, la varianza de $\hat{t}_d$ está dada por la expresión (2.33), por lo que

$$V(r) = \frac{1}{t_x^2}\left[N^2\left(1 - \frac{n}{N}\right)\frac{S_{td}^2}{n} + \frac{N}{n}\sum_{h=1}^{N} M_h^2\left(1 - \frac{m_h}{M_h}\right)\frac{S_{hd}^2}{m_h}\right], \tag{2.42}$$

donde

$$S_{td}^2 = \frac{1}{N-1}\sum_{h=1}^{N}\left(t_{hd} - \frac{t_d}{N}\right)^2 \qquad \text{con} \qquad t_d = \sum_{h=1}^{N} t_{hd},$$

$$S_{hd}^2 = \frac{1}{M_h - 1}\sum_{j=1}^{M_h}\left(d_{hj} - \frac{t_{hd}}{M_h}\right)^2 \qquad \text{y} \qquad t_{hd} = \sum_{j=1}^{M_h} d_{hj}.$$

En este caso, la varianza estimada del estimador de razón es

$$v(r) = \frac{1}{\hat{t}_x^2}\left[N^2\left(1 - \frac{n}{N}\right)\frac{s_{td}^2}{n} + \frac{N}{n}\sum_{h=1}^{N} M_h^2\left(1 - \frac{m_h}{M_h}\right)\frac{s_{hd}^2}{m_h}\right]. \tag{2.43}$$

Aquí

$$s_{td}^2 = \frac{1}{n-1}\sum_{h\in\mathcal{S}}\left(\hat{t}_{hd} - \frac{\hat{t}_d}{N}\right)^2 \qquad \text{y} \qquad s_{hd}^2 = \frac{1}{m_h - 1}\sum_{j\in\mathcal{S}_h}\left(\hat{d}_{hj} - \hat{\bar{d}}_h\right)^2.$$

En estas expresiones, $\hat{d}_{hj} = y_{hj} - rx_{hj}$ y

$$\hat{t}_d = \frac{N}{n}\sum_{h\in\mathcal{S}}\hat{t}_{hd} \qquad \hat{t}_{hd} = M_h\hat{\bar{d}}_h \qquad \hat{\bar{d}}_h = \frac{1}{m_h}\sum_{j\in\mathcal{S}_h}\hat{d}_{hj}$$

2.4. MUESTREO ESTRATIFICADO CON CONGLOMERADOS EN DOS ETAPAS

En esta estrategia, la población se divide en H estratos no superpuestos U_h; cada estrato es subdividido en N_h unidades primarias (PSUs o conglomerados) separadas, de las cuales se seleccionan un MAS de n_h elementos. Las dos etapas puede describirse de la siguiente forma:

1. Un MAS de unidades primarias se toma en cada uno de los H estratos.
2. En cada unidad primaria en muestra se toma un MAS de unidades secundarias (SSUs o elementos).

La g-ésima unidad primaria del h-ésimo estrato consta de M_{hg} elementos, que son los elementos de la población. M_h indica el total de elementos en el estrato h.

2.4.1. ESTIMACIÓN

La población total $Y = \sum_{h=1}^{H} Y_h$, con Y_h el total en el estrato h, se estima por

$$\widehat{Y} = \sum_{h=1}^{H}\widehat{Y}_h = \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{g=1}^{n_h}\frac{M_{hg}}{m_{hg}}\sum_{j=1}^{m_{hg}} y_{hgj}, \tag{2.44}$$

donde

$\widehat{Y}_h \equiv$ es el estimador del total Y_h en el estrato h

$m_{hg} \equiv$ es el tamaño de muestra de SSUs en el g-ésimo PSU del estrato h.

$m_h \equiv$ total de SSUs en muestra en el estrato h.

$y_{hgj} \equiv$ es el j-ésimo valor de la variable y en el g-ésimo PSU del estrato h.

$\widehat{Y}_h$ corresponde a un muestreo por conglomerados en dos etapas en el estrato h. Además, como los estratos son independientes entre si, la varianza del estimador (2.44) esta dado por

$$V(\widehat{Y}) = \sum_{h=1}^{H} \frac{N_h}{n_h} \left[(N_h - n_h) S_{1h}^2 + \sum_{g=1}^{Nh} \frac{M_{hg}}{m_{hg}} (M_{hg} - m_{hg}) S_{2hg}^2 \right], \tag{2.45}$$

donde

$$S_{1h}^2 = \frac{1}{N_h - 1} \sum_{g=1}^{N_h} \left(Y_{hg} - \frac{Y_h}{N_h} \right)^2, \qquad Y_h = \sum_{g=1}^{N_h} Y_{hg}, \qquad Y_{hg} = \sum_{j=1}^{M_{hg}} y_{hgj},$$

$$S_{2hg}^2 = \frac{1}{M_{hg} - 1} \sum_{j=1}^{M_{hg}} \left(y_{hgj} - \overline{Y}_{hg} \right)^2 \qquad \text{y} \qquad \overline{Y}_{hg} = \frac{Y_{hg}}{M_{hg}}.$$

Los sumandos de la expresión (2.45) constan de dos componentes que corresponden a un muestreo en dos etapas. Esto es, en un muestreo en dos etapas se debe considerar la varianza entre las unidades primarias de selección y la varianza entre las unidades secundarias. La varianza estimada de (2.45) se tiene al sustituir los valores muestrales apropiados. Así, el estimador correspondiente sera insesgado, pues ésta se hereda de el insesgamiento del muestreo por conglomerados en cada estrato.

2.4.2. ESTIMADOR DE RAZÓN

El estimador de razón se obtiene del cociente entre la estimación del total para la variable y y el total de la variable auxiliar x.

$$r = \frac{\widehat{Y}}{\widehat{X}} = \frac{\displaystyle\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g=1}^{n_h} \frac{M_{hg}}{m_{hg}} \sum_{j=1}^{m_{hg}} y_{hgj}}{\displaystyle\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g=1}^{n_h} \frac{M_{hg}}{m_{hg}} \sum_{j=1}^{m_{hg}} x_{hgj}}. \tag{2.46}$$

La varianza aproximada de este estimador sigue un proceso similar al descrito en la sección (2.3.3), es decir:

$$\begin{aligned} V(r) &= \frac{1}{X^2} V\left[\widehat{Y} - R\widehat{X}\right] \\ &= \frac{1}{X^2} V\left[\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g=1}^{n_h} \frac{M_{hg}}{m_{hg}} \sum_{j=1}^{m_{hg}} (y_{hgj} - Rx_{hgj})\right] \\ &= \frac{1}{X^2} V\left[\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g=1}^{n_h} \frac{M_{hg}}{m_{hg}} \sum_{j=1}^{m_{hg}} d_{hgj}\right]. \end{aligned}$$

Note que la expresión entre corchetes es similar a (2.44), y representa el total estimado para la variable $d_{hgj} = y_{hgj} - Rx_{hgj}$, por lo tanto su varianza está dada por la expresión (2.45). Así

$$V(r) = \frac{1}{X^2} \sum_{h=1}^{H} \frac{N_h}{n_h} \left[(N_h - n_h) S_{1h}^2 + \sum_{g=1}^{Nh} \frac{M_{hg}}{m_{hg}} (M_{hg} - m_{hg}) S_{2hg}^2\right] \tag{2.47}$$

donde

$$S_{1h}^2 = \frac{1}{N_h - 1} \sum_{g=1}^{N_h} \left(D_{hg} - \frac{D_h}{N_h}\right)^2, \qquad D_h = \sum_{g=1}^{N_h} D_{hg}, \qquad D_{hg} = \sum_{j=1}^{M_{hg}} d_{hgj}.$$

$$S_{2hg}^2 = \frac{1}{M_{hg} - 1} \sum_{j=1}^{M_{hg}} \left(d_{hgj} - \overline{D}_{hg}\right)^2 \qquad \text{y} \qquad \overline{D}_{hg} = \frac{D_{hg}}{M_{hg}}.$$

La varianza estimada se obtiene al sustituir en la expresión (2.47) las cantidades muestrales correspondientes, en cuyo caso $\hat{d}_{hgj} = y_{hgj} - rx_{hgj}$,

$$s_{1h}^2 = \frac{1}{n_h - 1} \sum_{g \in \mathcal{S}_h} \left(d_{hg} - \frac{d_h}{N_h} \right)^2, \qquad d_h = \frac{N_h}{n_h} \sum_{g \in \mathcal{S}_h} d_{hg}, \qquad d_{hg} = \sum_{j \in \mathcal{S}_{hg}} M_{hg} \bar{d}_{hg}.$$

$$s_{2hg}^2 = \frac{1}{m_{hg} - 1} \sum_{j \in \mathcal{S}_{hg}} \left(d_{hgj} - \bar{d}_{hg} \right)^2, \quad \text{y} \qquad \bar{d}_{hg} = \frac{1}{m_{hg}} \sum_{j \in \mathcal{S}_{hg}} d_{hgj}.$$

3. MUESTREO PROBABILÍSTICO APLICADO AL CR

La teoría del muestreo probabilístico es la que generalmente se usa para realizar las estimaciones para el Conteo Rápido, esto es, de un total de N casillas que se instalan en la jornada electoral, se selecciona una muestra aleatoria de n casillas, empleando alguna de las estrategias de selección descritas en el capítulo anterior. Una vez que se dispone de esta información, se hacen estimaciones para el porcentaje de votos para cada candidato.

El objetivo del Conteo Rápido es producir estimaciones del porcentaje de votos, en favor de cada candidato que participa en la elección. Este porcentaje, se calcula como el cociente entre el número de votos a favor de un candidato y el número total de votos emitidos para todos los candidatos, incluyendo los votos nulos y los candidatos no registrados. Es importante mencionar que en este cociente se desconoce tanto el numerador como el denominador, por lo que será obligado utilizar estimadores de razón.

NOTACIÓN

En esta sección se introduce la variable candidato a los estimadores del porcentaje de votos para cada estrategia de selección descrita en el capítulo anterior. La variable candidato se denotará con el subíndice k, con $k = 1, 2, ..., B$, donde B es el número total de candidatos incluyendo los votos nulos y los candidatos no registrados. Así,

$$p_k \equiv \text{denota al porcentaje real de votos para el candidato } k.$$

3.1. MUESTREO ALEATORIO SIMPLE SIN REMPLAZO (MASSR)

Suponiendo una población de N casillas para alguna elección particular, el estimador de la proporción de votos para un candidato k vía un MASSR, tomando un tamaño de muestra n, es

$$\widehat{p}_k = \frac{\bar{y}_k}{\bar{x}}. \tag{3.1}$$

Donde

y_i^k Número de votos a favor del candidato k en la casilla i

$\bar{y}_k = \dfrac{1}{n}\sum_{i\in\mathcal{S}} y_i^k$ Estimador de la media de votos para el candidato k.

$\bar{x}$ Estimador del promedio de votos totales.

x_i Total de votos en la casilla i.

Definiendo $x_i = \sum_{k=1}^{B} y_i^k$, se tiene que

$$\bar{x} = \frac{1}{n}\sum_{i\in\mathcal{S}} x_i = \frac{1}{n}\sum_{i\in\mathcal{S}}\sum_{k=1}^{B} y_i^k = \sum_{k=1}^{B} \bar{y}_k \tag{3.2}$$

Así, la varianza real de la expresión (3.1) se obtiene aplicando directamente la expresión (2.9) dada en la sección (2.1.4). Es decir,

$$V\left(\widehat{p}_k\right) = \left(1 - \frac{n}{N}\right)\frac{S_{dk}^2}{n\overline{X}^2} \qquad \text{con} \qquad S_{dk}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(d_i - \overline{D}\right)^2.$$

$\overline{X}$ es la media de votos totales en la elección, $d_i = y_i^k - p_k x_i$, $\overline{D} = \dfrac{1}{N}\sum_{i=1}^{N} d_i$. Cuando estos valores son desconocidos, se estiman por

$$v\left(\widehat{p}_k\right) = \left(1 - \frac{n}{N}\right)\frac{s_{dk}^2}{n\overline{x}^2} \qquad \text{con} \qquad s_{dk}^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\hat{d}_i - \bar{d}\right)^2.$$

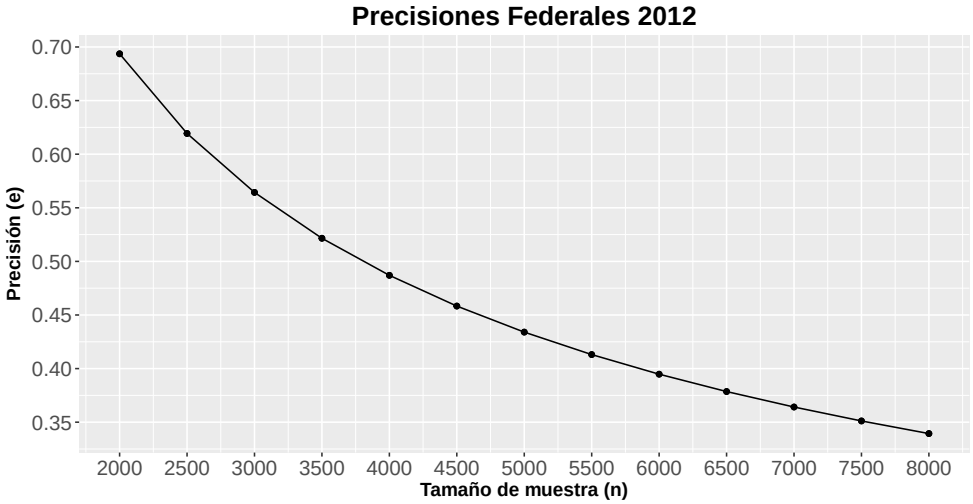
En este caso, $\hat{d}_i = y_i^k - \widehat{p}_k x_i$ y $\bar{d} = \dfrac{1}{n}\sum_{i=1}^{n}\hat{d}_i$. Otro valor de interés es el margen de error para la estimación del porcentaje de votos de un candidato,

$$\epsilon = z_{\alpha/2}\sqrt{V\left(\widehat{p}_k\right)}. \tag{3.3}$$

EJEMPLO

Consideremos la base de datos de la Elección Presidencial de 2012, en esta elección se instalaron $N = 143,437$ casillas (esta sería nuestra población total). Se calculará un margen de error, ϵ, para diferentes tamaños de muestra, $n = 2000, 2500, \ldots, 8000$, vía un MASSR. Esto se realizará para el partido de mayor varianza, que en este caso fue la coalición PRD-PT-MC.

Si consideramos un nivel de confianza del 95 % y la expresión (3.3) obtenemos la siguiente gráfica de precisiones.



Gráfica 3-1.: Muestreo Aleatorio Simple

En la gráfica se observa que con un tamaño de muestra de $n = 7,500$ se alcanza un margen de error menor o igual a $\epsilon = 0.35$. Si se desea mejorar el margen de error para el porcentaje de votos, el tamaño de muestra aumentara drásticamente reflejando la simplicidad de un MASSR. Hay que mencionar que en este caso, obviamente se conocen todos los valores poblacionales, sin embargo esta información podría darnos una muy buena idea de los tamaños de muestra que serían necesarios en el Conteo Rápido de 2018 para alcanzar una determinada precisión mediante un MASSR.

3.2. MUESTREO ALEATORIO ESTRATIFICADO (MAE)

Al igual que en la sección anterior, k denota el candidato de interés y n_h el número de casillas seleccionadas en el estrato h. Entonces, los valores muestrales a nivel estrato son

Media Razón

$$\overline{y}_{hk} = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}^k \qquad \widehat{p}_{hk} = \frac{\bar{y}_{hk}}{\bar{x}_h}.$$

donde y_{hj}^k denota el número de votos a favor del candidato k en la casilla j del estrato h y $\bar{x}_h$ es el promedio de votos totales en cada estrato, similar al descrita en (3.2).

$$\bar{x}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} x_{hj}.$$

donde $x_{hj} = \sum_{k=1}^{B} y_{hj}^k$ es el total de votos en la casilla j del estrato h.

PROPUESTA ESTIMADOR DE RAZÓN SEPARADO

Denotaremos a la lista nominal de la casilla j en el estrato h por l_{hj}. Así, el total de votos a favor de un candidato k se estima por:

$$\hat{t}_k = \sum_{h=1}^{H} \hat{t}_{hk} = \sum_{h=1}^{H} L_h \frac{\bar{y}_{hk}}{\bar{l}_h}. \tag{3.4}$$

donde L_h denota el total de la lista nominal en el estrato h, es decir, $L_h = \sum_{j=1}^{Nh} l_{hj}$, mientras que $\bar{l}_h$ es el promedio muestral de la lista nominal en el mismo estrato. Por otra parte, la votación total se calcula como

$$\begin{aligned} \hat{t} = \sum_{k=1}^{P} \hat{t}_k &= \sum_{h=1}^{H} \sum_{k=1}^{P} L_h \frac{\bar{y}_{hk}}{\bar{l}_h} \\ &= \sum_{h=1}^{H} L_h \frac{\sum_{k=1}^{P} \bar{y}_{hk}}{\bar{l}_h} = \sum_{h=1}^{H} L_h \frac{\bar{x}_h}{\bar{l}_h}. \end{aligned} \tag{3.5}$$

con $\bar{x}_h = \sum_{k=1}^{B} \bar{y}_{hk}$, esto se puede ver en la expresión (3.2). Por lo tanto, el estimador para el porcentaje de votos del candidato k es

$$\hat{p}_k = \frac{\hat{t}_k}{\hat{t}}. \tag{3.6}$$

Para el cálculo de la varianza se sigue un procedimiento similar al descrito en la sección (2.1.4).

$$\begin{aligned} V(\hat{p}_k) &= \frac{1}{t^2} V\left[\hat{t}_k - p_k \hat{t}\right] \\ &= \frac{1}{t^2} V\left[\sum_{h=1}^{H} L_h \frac{\bar{y}_{hk}}{\bar{l}_h} - p_k \sum_{h=1}^{H} L_h \frac{\bar{x}_h}{\bar{l}_h}\right] \\ &= \frac{1}{t^2} V\left[\sum_{h=1}^{H} \frac{L_h}{\bar{l}_h} \left(\bar{y}_{hk} - p_k \bar{x}_h\right)\right] \\ &= \frac{1}{t^2} V\left[\sum_{h=1}^{H} L_h \left(\frac{\bar{d}_{hk}}{\bar{l}_h}\right)\right] \\ &= \frac{1}{t^2} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{dhk}^2}{n_h^2}. \end{aligned} \tag{3.7}$$

La última expresión se sigue de (2.20). En este desarrollo, $\bar{d}_{hk}$ es la media muestral de la variable, $d_{hkj} = y_{hkj} - p_k x_{hj}$, t representa el total de votos y

$$S_{dhk}^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} \left(d_{2hkj} - \overline{D}_{2hk}\right)^2, \qquad \begin{aligned} d_{2hkj} &= d_{hkj} - p_{hk} l_{hj} \\ &= y_{hkj} - p_k x_{hj} - p_{hk} l_{hj}. \end{aligned}$$

con $\overline{D}_{2hk} = \dfrac{1}{N_h} \sum_{j=1}^{N_h} d_{2hkj}$, $p_{hk} = \dfrac{\overline{D}_{hk}}{\overline{L}_h}$, $\overline{D}_{hk} = \dfrac{1}{N_h} \sum_{j=1}^{N_h} d_{hkj}$ y $\overline{L}_h$ es la media real de la lista nominal en el estrato h. Con esto, (3.7) se estima por

$$v(\hat{p}_k) = \frac{1}{\hat{t}^2} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{dhk}^2}{n_h}, \tag{3.8}$$

donde

$$s^2_{dhk} = \frac{1}{n_h - 1} \sum_{j \in \mathcal{S}_h} \left(\hat{d}_{2hkj} - \bar{d}_{2hk} \right)^2, \quad \bar{d}_{2hk} = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} \hat{d}_{2hkj} \quad \text{y} \quad \hat{d}_{2hkj} = y_{hkj} - \hat{p}_k x_{hj} - \hat{p}_{hk} l_{hj}.$$

Mientras que, $\hat{p}_{hk} = \dfrac{\hat{\bar{d}}_{hk}}{\bar{l}_h}$, $\hat{\bar{d}}_{hk} = \dfrac{1}{n_h} \sum_{j \in \mathcal{S}_h} \hat{d}_{hkj}$ con $\hat{d}_{hkj} = y_{hkj} - \hat{p}_k x_{hj}$.

ESTIMADOR COMBINADO

Los estimadores para el total de votos a favor del candidato k y el total de votos en la elección son:

$$y_k = \sum_{h=1}^{H} N_h \bar{y}_{hk} \qquad\qquad x_e = \sum_{h=1}^{H} N_h \bar{x}_h.$$

Así el estimador de razón combinado se escribe como

$$\widehat{p}_k = \frac{y_k}{x_e}. \tag{3.9}$$

Cuya varianza esta dada directamente por la expresión (2.25) definida en la sección (2.2.2):

$$V\left(\hat{p}_k\right) = \frac{1}{X^2} \left[\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S^2_{dhk}}{n_h} \right], \tag{3.10}$$

X representa el total de la variable x. La varianza estimada de $\hat{p}_k$ esta descrita por

$$v\left(\hat{p}_k\right) = \frac{1}{x_e^2} \left[\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s^2_{dhk}}{n_h} \right]. \tag{3.11}$$

Donde

$$S^2_{dhk} = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} \left(d_{hkj} - \overline{D}_{hk} \right)^2 \qquad \text{y} \qquad s^2_{dhk} = \frac{1}{n_h - 1} \sum_{j \in \mathcal{S}_h} \left(\hat{d}_{hkj} - \bar{d}_{hk} \right)^2.$$

Aquí, $\overline{D}_{hk} = \dfrac{1}{N_h} \sum_{j=1}^{N_h} d_{hkj}$, $\bar{d}_{hk} = \dfrac{1}{n_h} \sum_{j \in \mathcal{S}_h} \hat{d}_{hkj}$ con $d_{hkj} = y_{hkj} - p_k x_{hj}$ y $\hat{d}_{hkj} = y_{hkj} - \hat{p}_k x_{hj}$.

MARGEN DE ERROR:

Tanto para el estimador separado como para el combinado, el margen de error se escribe como en la expresión (3.3). Es decir,

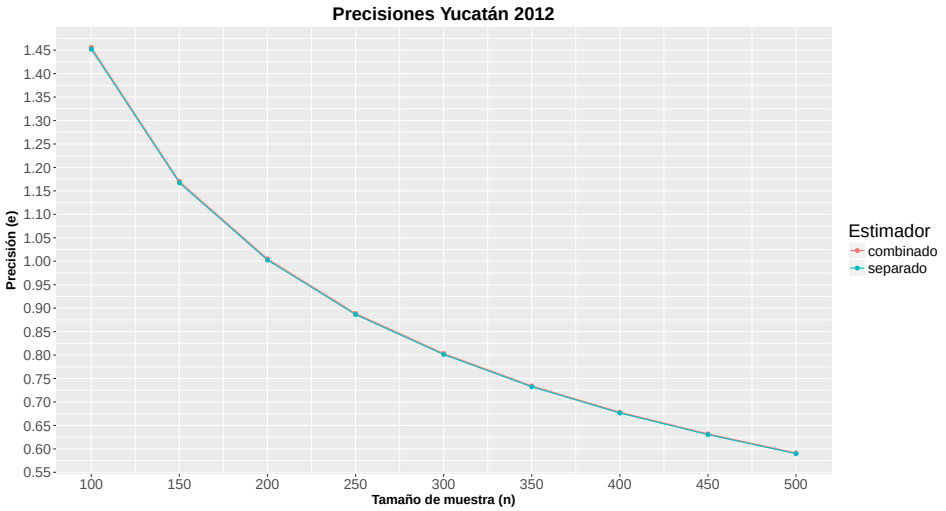
$$\epsilon = Z_{\alpha/2}\sqrt{V\left(\widehat{p}_k\right)}. \tag{3.12}$$

EJEMPLOS

En los siguientes ejemplos se calcula el margen de error para el estimador separado y el combinado, tomando una distribución de muestra proporcional al número de casillas en el estrato y considerando una estratificación por distritos federales. El margen de error se calcula para el candidato que obtuvo mayor varianza en porcentaje de votos.

ELECCIONES YUCATÁN 2012

En esta elección se contó con una población de $N = 2,921$ casillas, donde el candidato de la coalición PRI-PVEM-PSDYUC obtuvo mayor varianza. Así, el margen de error para diferentes tamaños de muestra para este candidato es:

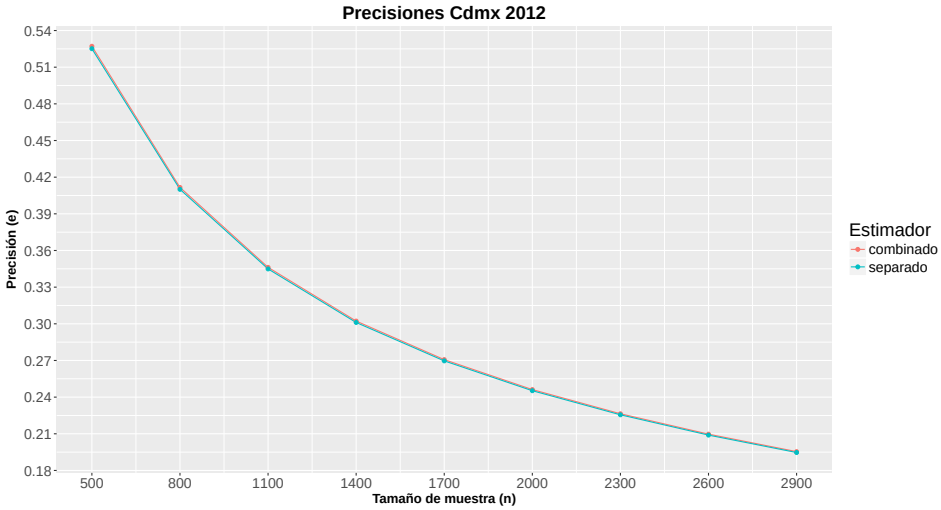


Gráfica 3-2.: Estratificación por distrito federal

En esta gráfica se observa que el error para ambos estimadores es prácticamente el mismo, es decir, que no existe diferencia entre emplear un estimador u otro.

ELECCIONES PARA JEFE DE GOBIERNO CDMX, 2012

En la elección para Jefe de Gobierno de 2012, se instalaron $N = 12,383$ casillas, la mayor varianza se obtuvo para la coalición PRD-PT-MC. La correspondiente gráfica para el margen de error es

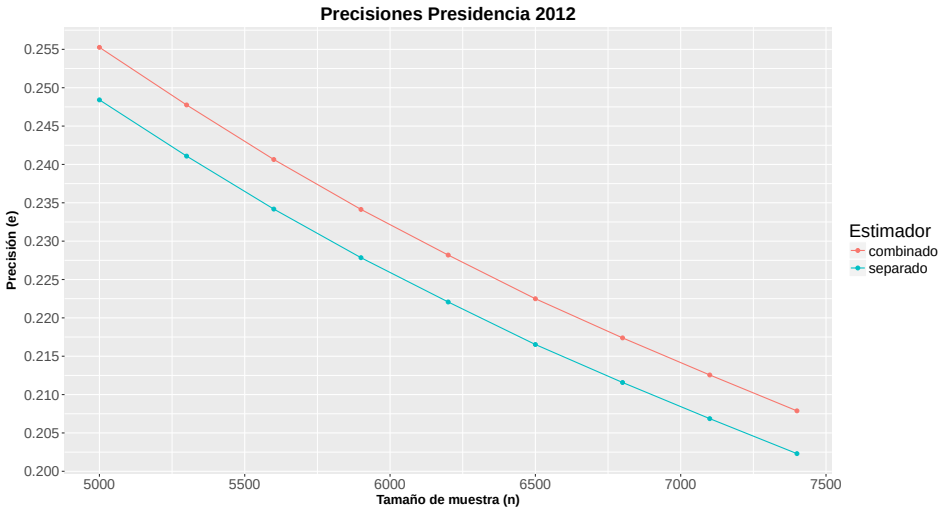


Gráfica 3-3.: Estratificación por distrito federal

Nuevamente se aprecia la similitud entre ambos estimadores.

ELECCIÓN PRESIDENCIAL 2012

Para la base de datos de la Elección Presidencial de 2012, se instalaron $N = 143,437$ casillas, en este caso la mayor varianza se registró para la coalición PRD-PT-MC, la gráfica con los márgenes de error se muestra a continuación.



Gráfica 3-4.: Estratificación por distrito federal

Esta gráfica difiere de las anteriores pues la base de datos es considerablemente mas grande a las dos primeras. En este caso, se observan un mejor comportamiento para el estimador separado con una diferencia media de 0.0062 % entre los márgenes de error.

COMPARACIÓN ENTRE AMBOS ESTIMADORES VÍA SIMULACIÓN

El proceso consiste en seleccionar muestras vía un MAE con estratificación por distritos federales, se construye un intervalo de confianza al 95 %, vía las fórmulas vistas anteriormente, y se cuenta el número de veces que el verdadero valor del porcentaje de votos, para el candidato de mayor varianza, cayó en dicho intervalo. Este proceso se repitió $10,000$ veces para diferentes tamaños de muestra.

Las primeras dos tablas corresponden a elecciones locales, donde se tiene que para los tamaños de muestra considerados, el 95 % de los intervalos captura al verdadero valor del porcentaje de votos.

YUCATÁN

n	Separado (%)	Combinado (%)
125	95	95
175	95	95
225	95	95
275	95	95
325	95	95
375	95	95

CDMX

n	Separado (%)	Combinado (%)
500	95	95
800	95	95
900	95	95
1000	95	95
1100	95	95
1200	95	95

Por otra parte, para la elección presidencial también se tiene un comportamiento similar. Esto puede explicarse debido a la similitud de las estimaciones del porcentaje de votos entre ambos estimadores y a que la diferencia entre las precisiones es de apenas 0.0062.

PRESIDENCIAL

n	Separado (%)	Combinado (%)
5000	95	95
5500	95	95
6000	95	95
6500	95	95
7000	95	95
7500	95	95

De las tablas anteriores podemos concluir que prácticamente no existe diferencia entre el estimador de razón separado y combinado.

3.3. MUESTREO POR CONGLOMERADOS EN DOS ETAPAS (MPCDE)

Estimadores del total de votos para un candidato k y el total de votos en la elección son los siguientes, suponiendo MASSR en cada etapa.

$$\hat{t}_k = \frac{N}{n}\sum_{h\in\mathcal{S}} \hat{t}_{hk} = \frac{N}{n}\sum_{h\in\mathcal{S}} M_h \bar{y}_{hk} \qquad \text{y} \qquad \hat{t} = \frac{N}{n}\sum_{h\in\mathcal{S}} \hat{t}_h = \frac{N}{n}\sum_{h\in\mathcal{S}} M_h \bar{x}_h.$$

Así, el estimador del porcentaje de votos para el candidato k vía muestreo por conglomerados es

$$\hat{p}_k = \frac{\hat{t}_k}{\hat{t}} = \frac{\sum_{h\in\mathcal{S}} \hat{t}_{hk}}{\sum_{h\in\mathcal{S}} \hat{t}_h} = \frac{\sum_{h\in\mathcal{S}} M_h \bar{y}_{hk}}{\sum_{h\in\mathcal{S}} M_h \bar{x}_h}. \tag{3.13}$$

Donde

$N \equiv$ Número conglomerados (PSU) en la población

$M_h \equiv$ Número de elementos (SUS's) en el conglomerado h

$m_h \equiv$ Tamaño de muestra en el conglomerado h

$\hat{t}_k \equiv$ Estimador del total de votos para el candidato k

$\bar{y}_{hk} \equiv$ Estimador de la media de votos en el conglomerado h para el candidato k

$y_{hjk} \equiv$ Votos en la casilla j del conglomerado i para el candidato k

$\hat{t} \equiv$ Estimador del total de votos de la elección

$\bar{x}_h \equiv$ Estimador de la media de votos en el conglomerado h

$x_{hj} = \sum_{k=1}^{B} y_{hkj} \equiv$ Total de votos en casilla j del conglomerado$h$

La varianza del estimador (3.16) es la siguiente

$$V(\hat{p}_k) = \frac{1}{t^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{S^2_{tkd}}{n} + \frac{N}{n}\sum_{h=1}^{N} M_h^2\left(1-\frac{m_h}{M_h}\right)\frac{S^2_{hkd}}{m_h}\right] \tag{3.14}$$

$$S^2_{tkd} = \frac{1}{N-1}\sum_{h=1}^{N}\left(d_{hk}-\frac{dk}{N}\right)^2 \qquad \text{y} \qquad S^2_{hkd} = \frac{1}{M_h-1}\sum_{j=1}^{M_h}\left(d_{hkj}-\frac{d_{hk}}{M_h}\right)^2,$$

con t el total de votos en la elección. $d_{hkj} = y_{hkj} - p_k x_{hj}$, $d_{hk} = \sum_{j=1}^{N} d_{hkj}$ y $d_k = \sum_{h=1}^{N} d_{hk}$. Luego la estimación de la expresión (3.15) se obtiene sustituyendo S^2_{tkd} y S^2_{hkd} por sus respectivas valores muestrales.

$$v(\hat{p}_k) = \frac{1}{\hat{t}^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{s^2_{tkd}}{n} + \frac{N}{n}\sum_{h=1}^{N} M_h^2\left(1-\frac{m_h}{M_h}\right)\frac{s^2_{hkd}}{m_h}\right] \tag{3.15}$$

$$s^2_{tkd} = \frac{1}{n-1}\sum_{h\in\mathcal{S}}\left(\hat{d}_{hk} - \frac{\hat{d}_k}{N}\right)^2 \qquad \text{y} \qquad s^2_{hkd} = \frac{1}{m_h - 1}\sum_{j\in\mathcal{S}_h}\left(\hat{d}_{hkj} - \frac{\hat{d}_{hk}}{m_h}\right)^2,$$

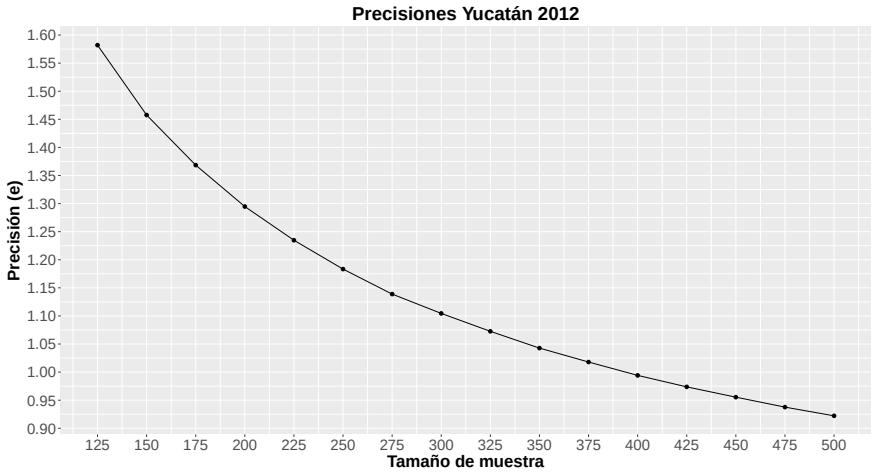
donde $\hat{d}_{hkj} = y_{hkj} - \hat{p}_k x_{hj}$, $\hat{d}_{hk} = M_h \hat{\bar{d}}_{hk}$, $\hat{d}_k = \frac{N}{n}\sum_{h\in\mathcal{S}} \hat{d}_{hk}$ y $\hat{\bar{d}}_{hk} = \frac{1}{m_h}\sum_{j\in\mathcal{S}_h} \hat{d}_{hkj}$.

EJEMPLOS

En las siguientes gráficas se muestran márgenes de error, considerando diferentes tamaños de muestra, para los estados de YUCATÁN y CDMX. En ambos casos se trabaja con las coaliciones de mayor varianza. Estratificando a la población de casillas por los distritos federales que comprenden cada estado, 5 y 24 respectivamente.

YUCATÁN

Para el cálculo de estos márgenes de error se seleccionó un muestreo aleatorio simple de 4 distritos federales (PSU), de un total de 5. Los tamaños de muestra de casillas (USS) empleados son de $125, 175, ..., 475$, tomadas de manera proporcional al número de casillas en cada conglomerado (distrito). El comportamiento del margen de error con estos tamaños de muestra se presenta en la siguiente gráfica:



Gráfica 3-5.: Coalición PRI_PVEM_PSDYUC.

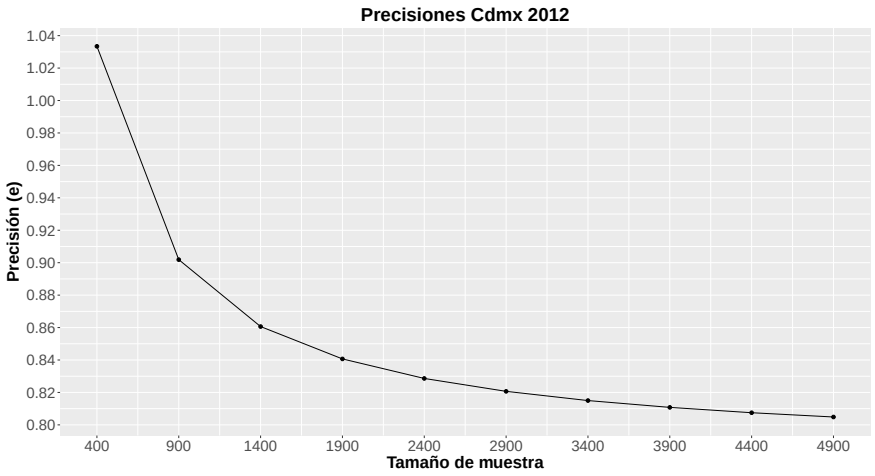
Aquí se puede observar que esta estrategia de muestreo requiere mayor número de casillas para alcanzar un buen margen de error comparado con el muestreo estratificado, Por ejemplo,

n	Error por MPCDE	Error por MAE
200	1.30 %	0.96 %

Tabla 3-1.: MPCDE vs MAE, Yucatán 2012

CDMX

En este caso, de un total de 24 distritos federales, se toma por muestreo aleatorio simple 19 distritos y se consideran tamaños de muestra de $400, 900, ..., 4900$ casillas.



Gráfica 3-6.: Coalición PRD_PT_MC.

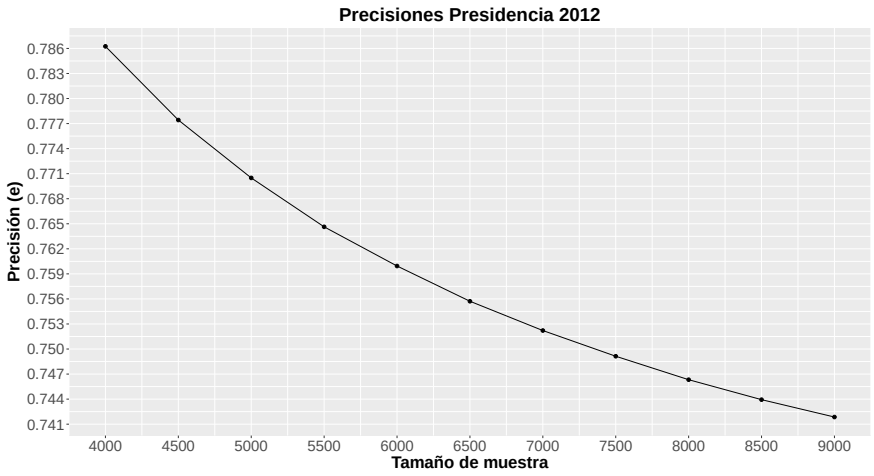
Se observa que el margen de error es muy volátil, ya que además de depender de las casillas seleccionadas, éstas también dependen de los distritos que se seleccionaron. Por lo que, comparado con el muestreo estratificado estos errores son mayores. Por ejemplo,

n	Error por MPCDE	Error por MAE
200	0.83 %	0.30 %

Tabla 3-2.: MPCDE vs MAE, CDMX 2012

PRESIDENCIAL

En total existen 300 distritos federales, de los cuales se ha tomando de forma aleatoria $n = 240$. Se ha calculado el margen de error para distintos tamaños de muestra: $4000, 4500, ..., 9000$ casillas distribuidas de forma proporcional en estos distritos.



Gráfica 3-7.: Coalición PRD_PT_MC

Con esta estrategia de selección se tiene una precisión de 0.742 % con 9,000 casillas en muestra, mientras que con un muestreo estratificado se alcanza una precisión de aproximadamente 0.20 % con el mismo tamaño. Por lo tanto, el muestreo por conglomerado ofrece una eficiencia muy por debajo del muestreo estratificado.

3.4. ESTRATIFICADO CON CONGLOMERADOS EN DOS ETAPAS (ECCDE)

Dada una población de casillas de una elección dividida en H estratos independientes, el estimador del porcentaje de votos para un candidato k es

$$\hat{p}^k = \frac{\hat{y}^k}{\hat{x}} = \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g \in \mathcal{S}_h} \frac{M_{hg}}{m_{hg}} \sum_{j \in \mathcal{S}_{hg}} y_{hgkj}}{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g \in \mathcal{S}_h} \frac{M_{hg}}{m_{hg}} \sum_{j \in \mathcal{S}_{hg}} x_{hgj}}. \tag{3.16}$$

donde $x_{hgj} = \sum_{k=1}^{B} y_{hgkj}$ es el total de votos en la casilla j del PSU g perteneciente al estrato h. La varianza de la proporción de votos está descrita por

$$V\left[\hat{p}_k\right] = \frac{1}{X^2} \sum_{h=1}^{H} \frac{N_h}{n_h} \left[(N_h - n_h) S_{1h}^2 + \sum_{g=1}^{Nh} \frac{M_{hg}}{m_{hg}} (M_{hg} - m_{hg}) S_{2hg}^2 \right], \tag{3.17}$$

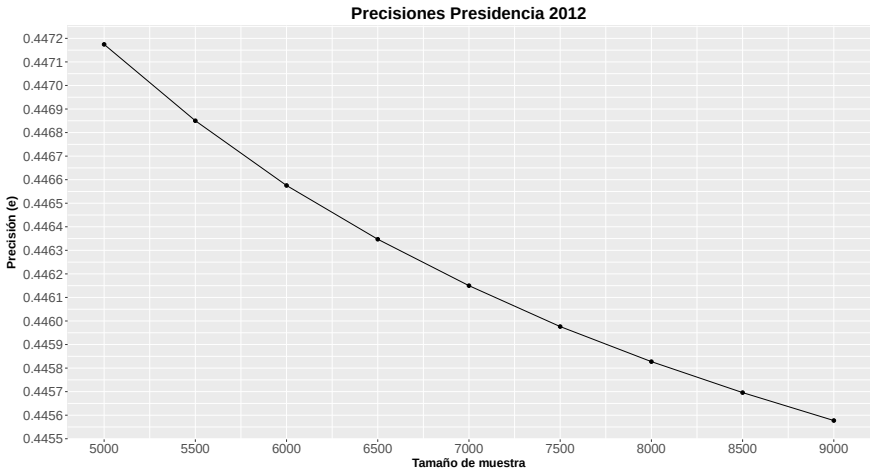
con varianza estimada

$$v\left[\hat{p}_k\right] = \frac{1}{\hat{x}^2}\sum_{h=1}^{H}\frac{N_h}{n_h}\left[(N_h - n_h)s_{1h}^2 + \sum_{g=1}^{Nh}\frac{M_{hg}}{m_{hg}}(M_{hg} - m_{hg})s_{2hg}^2\right]. \tag{3.18}$$

S_{1h}^2, S_{2hg}^2 así como sus estimaciones se describen en la sección (2.4.2), únicamente se remplaza d_{hgj} por $d_{hgkj} = y_{hgkj} - p_k x_{hgj}$ y $\hat{d}_{hgj}$ por $\hat{d}_{hgkj} = y_{hgkj} - \hat{p}_k x_{hgj}$.

EJEMPLO

En este ejemplo, se trabaja con la base de datos de la elección federal de 2012. Considerando a los estados como variable de estratificación. En cada estrato, las unidades primarias de selección son los distritos federales, y las unidades secundarias las casillas. Se tomó un MAS del 80 % de unidades primarias en cada estrato. Entonces, para diferentes tamaños de muestra se obtuvieron los siguientes márgenes de error.



Gráfica 3-8.: Coalición PRD_PT_MC

Claramente este método ofrece mejor margen de error que el método anterior. No obstante, es menos eficiente que un muestreo aleatorio simple. Por ejemplo,

n	Error por ECCDE	Error por MPCDE	Error por MASSR
7500	0.45 %	0.75 %	0.36 %

Tabla 3-3.: ECCDE vs MPCDE, Presidencia 2012

Finalmente, de todas las gráficas anteriores, se puede observar que el muestreo aleatorio estratificado es el que ofrece mejor margen de error en cada caso. Por tal razón, se suele usar con mayor frecuencia en el conteo rápido.

4. RETOS Y POSIBLES SOLUCIONES

Del capítulo anterior, podemos observar que el muestreo aleatorio estratificado, en comparación con las otras estrategias de selección, ofrece un menor margen de error con menor tamaño de muestra. Por esta razón, los Comités Técnicos para el Conteo Rápido han decidido consistentemente usar esta estrategia de selección. Bajo esta estrategia, lo que se busca es la estratificación "óptima", i.e. la que arroja el menor margen de error con el menor tamaño de muestra.

Sin embargo, a pesar de los buenos resultados, existen ciertas restricciones operativas que no han sido atendidas en su totalidad y que representan un reto para el Conteo Rápido. Una de estas restricciones, involucra directamente al personal que colabora en el proceso del Conteo Rápido, los llamados Capacitadores y Asistentes Electorales (mejor conocidos como CAEs). El problema radica en que por limitaciones de tiempo, distancia y debido a las múltiples actividades que los CAEs deben realizar el día de la jornada electoral, es sumamente difícil que un mismo CAE reporte información, para CR, de más de una casilla. Como se describió en la introducción, cada CAE tiene a su cargo, en promedio, 4 casillas. A estas casillas, se les conoce como el área de responsabilidad del CAE o ARE.

Posiblemente debido a lo anterior, además de otras razones, hasta antes de 2018 se observaron porcentajes de no respuesta muy altos (no respuesta se entiende como la información que no se reportó a tiempo para la estimación). Por ejemplo, en la elección de 2016 para gobernador del estado de Oaxaca se observó una no respuesta del 36.74 %. Mientras que en la elección de Nayarit 2017, la no respuesta fue del 50 %. Otro caso extremo, se presentó en la elección de Colima, en el mismo año, con un porcentaje de no respuesta del 53.39 %. Sin duda, estos porcentajes de no respuesta son muy altos, y el riesgo que se corre es que las estimaciones resultantes pueden llegar a distar mucho de las que se obtendrían con la muestra completa.

Para resolver este problema, en los Conteos Rápidos de 2018, el COTECORA impuso una restricción al tamaño de la muestra con la intención de disminuir la presión sobre los CAEs y en teoría obtener tasas menores de no respuesta. La restricción fue que al menos 80 % de los CAEs, que tienen bajo su responsabilidad alguna casilla de la muestra para el conteo

rápido, reporten información de sólo una casilla. Esto no es otra cosa que una penalización a tamaños de muestra "grandes". A pesar de que con esto, los porcentajes de no respuesta fueron menores a los observados en 2017, aun fueron bastante altos. Por citar algunos casos, en la elección para gobernador del estado de Tabasco, la no respuesta fue del 41.6 %. Cabe mencionar que el intervalo de confianza reportado para el segundo lugar no contuvo al porcentaje definitivo para dicho candidato (ver tabla **1-8**). Adicionalmente, para la elección de Chiapas, con un corte en la recepción de información de las 00 : 30 horas del día siguiente a la jornada electoral, la no respuesta era del 41.2 %.

En este capítulo, se propone una estrategia de selección y forma del estimador de la proporción de votos en favor de cada candidato, que aborda la restricción operativa sobre los CAEs, sin penalizar el tamaño de muestra y asignando como máximo una sola casilla por CAE. Lo que se hace es cambiar el diseño de muestreo.

4.1. NOTACIÓN

Usaremos la siguiente notación.

CANTIDADES POBLACIONALES

$N = \sum_{h=1}^{H} N_h \equiv$ Total casillas en la elección

$X = \sum_{h=1}^{H} X_h \equiv$ Total de votos en la elección

$K = \sum_{h=1}^{H} K_h \equiv$ Total de $AREs$ en la elección

$N_h \equiv$ Número de casillas en el estrato h

$K_h \equiv$ Número de $AREs$ en el estrato h

$C_h \equiv$ Número de casillas en los $AREs$ del estrato h

$\overline{Y}_{kh} = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{khi} \equiv$ Media real de votos a favor del k-ésimo candidato en el estrato h

$X_h = \sum_{i=1}^{N_h} x_{hi} \equiv$ Total de votos en el estrato h

$\overline{Y}_{khr} = \dfrac{1}{C_h}\sum_{i=1}^{C_h} y_{khi} \equiv$ Media real de votos para el candidato k en el ARE_r del estrato h

$x_{hi} = \sum_{k=1}^{B} y_{khi} \equiv$ Total de votos en la i-ésima casilla del estrato h

$y_{khi} \equiv$ Votos en la i-ésima casilla del estrato h a favor del candidato k

CANTIDADES NIVEL ESTRATO

$n_h \equiv$ Número de casillas en muestra/Número de $AREs$ en muestra en el estrato h

$\bar{x}_h = \dfrac{1}{n_h}\sum_{i=1}^{n_h} x_{hi} \equiv$ Estimador de la media del total de votos en el estrato h

$s^2_{1h} \equiv$ Varianza estimada entre los votos del estrato h

$s^2_{2h} \equiv$ Varianza estimada de votos entre los $AREs$ del estrato h

$\mathcal{S}_h \equiv$ Conjunto de casillas seleccionadas en el estrato h

$\mathcal{S}_{hr} \equiv$ Conjunto de casillas seleccionadas en el ARE_r del estrato h

Para candidato k

$\bar{y}_{kh} \equiv$ Estimador de la media de votos para el k-ésimo candidato en el estrato h

$t_{khr} = C_h \sum_{i\in\mathcal{S}_{hr}} y_{hki} = C_h y_{hk} \equiv$ Estimador del total en el ARE_r en el estrato h

$\bar{y}_{khr} = \dfrac{t_{khr}}{C_h} = y_{hki} \equiv$ Estimador de la media en el ARE_r en el estrato h

$\bar{t}_{kh} = \dfrac{1}{n_h}\sum_{r=1}^{n_h} t_{khr} \equiv$ Media estimada de los totales entre $AREs$ en el estrato h

CANTIDADES NIVEL POBLACIONAL

$n = \sum_{h=1}^{H} n_h \equiv$ Tamaño de muestra

$y_k \equiv$ Estimador del total de votos a favor del k-ésimo candidato

$x \equiv$ Estimador de la votación total

$r_k \equiv$ Proporción de votos estimado a favor del k-ésimo candidato

$\mathcal{S} \equiv$ Conjunto de casillas seleccionadas en muestra

4.2. CONSTRUCCIÓN DE LOS ESTRATOS

Para dividir la población de casillas en H estratos, se toma en cuenta el número de casillas que componen a los $ARESs$. Por ejemplo, para una elección local o estatal, los estratos se forman de las siguiente manera:

1. Se forman grupos con los $ARESs$ que tengan el mismo número de casillas.
2. Cada grupo representará un estrato, por lo tanto, se tienen tantos estratos como grupos formados en el punto 1.

Con esta construcción de los estratos, cada estrato constará de $K_h \geq 1$ $ARESs$ que contienen el mismo número de casillas ($C_h \geq 1$) y en consecuencia, el número de casillas en el estrato h, será de $N_h = K_h C_h$. Para una elección federal, estos pasos se repiten en cada uno de los estados de la república. Por tanto, en cada estado se construye un número diferente de estratos, que en su totalidad definen a la estratificación final.
La razón para construir estratos con $ARESs$ que tengan el mismo número de casillas, es para obtener un estimador insesgado, esto se verá a detalle en las secciones (4.4) y (4.5).

4.3. ESTRATEGIA DE SELECCIÓN

La estrategia de selección consistirá en hacer una combinación de un muestreo estratificado y muestreo por conglomerados, con dos etapas en cada estrato, donde las unidades primarias de selección serán los $ARESs$ y las casillas como las unidades secundarias.

Primero, se fija un tamaño de muestra, de n casillas a seleccionarse de las N casillas de la población, y posteriormente se realizan los siguientes pasos, en cada estrato:

1. Se toma un muestreo aleatorio simple de n_h $AREs$. Con selección proporcional al número de casillas en el estrato, es decir

$$n_h = \text{mín}\left\{K_h, n\frac{N_h}{N}\right\}. \tag{4.1}$$

   Esta restricción implica que no podremos seleccionar más casillas que AREs, y es una restricción natural ya que buscamos que máximo se asigne una casilla por cada CAE.

2. Se selecciona por muestreo aleatorio simple una única casilla de cada ARE seleccionado en el primer paso.

Esta manera de hacer la selección implica que el número de casillas seleccionadas en cada estrato será igual al número de $AREs$ seleccionados en la primera etapa, es decir, el número de casillas en muestra, en el estrato h, será igual a n_h.

4.4. ESTIMADOR DE LA MEDIA DE VOTOS EN EL ESTRATO h

El objetivo del Conteo Rápido es estimar la proporción de votos para el k-ésimo candidato (ver sección 4.5), con la estrategia anterior utilizaremos el siguiente estimador.

$$\widehat{p}_k = \frac{\sum_{h=1}^{H} N_h \bar{y}_{kh}}{\sum_{h=1}^{H} N_h \bar{x}_h}.$$

Sin embargo, la varianza de esta expresión está fuertemente ligada a la varianza de la media de votos en el estrato h, por lo tanto, el cálculo de esta es esencial. Como puede verse en la sección 2.3, del capítulo 2, el estimador usual de la media de votos en el estrato h, para un muestreo en dos etapas, es:

$$\bar{y}_{kh} = \frac{1}{n}\sum_{r\in\mathcal{S}_{ha}}\sum_{i\in\mathcal{S}_{hr}} y_{khri}, \tag{4.2}$$

en donde $\mathcal{S}_{ha}$ tendría que ser el conjunto de $AREs$ en muestra, del estrato h, y $\mathcal{S}_{hr}$ el conjunto de casillas seleccionadas en el r-ésimo ARE de $\mathcal{S}_{ha}$. En nuestro caso, $\mathcal{S}_{hr}$ consta de una sola casilla, entonces el número de casillas en muestra es igual al número de $AREs$ seleccionados permitiendo reescribir a $\bar{y}_{kh}$ como el estimador tradicional para un muestreo estratificado con tamaño de muestra de n_h casillas. Esto es,

$$\bar{y}_{kh} = \frac{1}{n_h} \sum_{i \in \mathcal{S}_h} y_{khi}. \tag{4.3}$$

A partir de este momento, cuando se mencione al estimador de la media, nos estaremos refiriendo a la expresión (4.3). Esta manera de expresar a la media, además de ser mas sencilla, tiene la propiedad de ser insesgada. Para probar esto, definamos una variable auxiliar Z_i tal que tome valor 1 si la i-ésima casilla está en muestra y 0 en caso contrario. Por lo tanto,

$$\bar{y}_{kh} = \frac{1}{n_h} \sum_{i \in \mathcal{S}_h} y_{khi} = \frac{1}{n_h} \sum_{i=1}^{N_h} Z_i y_{khi}.$$

Además, del **Apéndice** B, se tiene que $\mathbb{E}[Z_i] = \dfrac{n_h}{K_h C_h}$, entonces

$$\begin{aligned} \mathbb{E}[\bar{y}_{kh}] &= \frac{1}{n_h} \sum_{i=1}^{N_h} \mathbb{E}[Z_i] y_{khi}, \\ &= \frac{1}{n_h} \sum_{i=1}^{N_h} \left(\frac{n_h}{K_h C_h} \right) y_{khi}, \\ &= \frac{1}{K_h C_h} \sum_{i=1}^{N_h} y_{khi}, \\ &= \frac{1}{N_h} \sum_{i=1}^{N_h} y_{khi} = \overline{Y}_{kh}. \end{aligned} \tag{4.4}$$

Es importante notar que la última igualdad es posible gracias a que $N_h = K_h C_h$, es decir, que todos los $AREs$ del estrato h tengan el mismo número de casillas, de lo contrario $\mathbb{E}[\bar{y}_{kh}] \neq \overline{Y}_{kh}$. Esta fue la razón por la que se construyeron estratos formados por AREs con el mismo número de casillas, ver Sección (4.2).

4.4.1. VARIANZA ANALÍTICA

Si se emplean los resultados habituales, ver Lohr [Lohr] o Cochran [Cochran], para un muestreo de dos etapas con tamaño de muestra m_{hr} en cada ARE, la varianza analítica estaría dada por la expresión (2.40) del Capítulo 2.

$$v(\bar{y}_{kh}) = \frac{K_h(K_h - n_h)}{n_h N_h^2} s_{kht}^2 + \frac{K_h}{n_h N_h^2} \sum_{r=1}^{K_h} C_h(C_h - m_r) \frac{s_{khr}^2}{m_r}. \tag{4.5}$$

en donde, s_{kht}^2 estima la varianza poblacional de las votaciones totales entre los $AREs$ y s_{hr}^2 la varianza de los votos en el ARE_r.

$$s_{kht}^2 = \frac{1}{n_h - 1} \sum_{r=1}^{n_h} (t_{khr} - \bar{t}_{kh})^2 \qquad s_{khr}^2 = \frac{1}{m_{hr} - 1} \sum_{i=1}^{m_r} (y_{kri} - \bar{y}_{khr})^2 .$$

Como en nuestro caso se selecciona un único elemento en cada unidad primaria (ARE) entonces $m_{hr} = 1$ y en consecuencia s_{khr}^2 no está definido. Por lo tanto, los resultados habituales no pueden emplearse con este tamaño de muestra. El reto de esta tesis es de hecho estimar esta cantidad muestral.

La deducción de la varianza poblacional, cuando $m_{hr} = 1$, se presenta en el **Apéndice** B en la expresión ($B.22$), que re-escrita con las cantidades correspondientes al estrato h está dada por

$$V(\bar{y}_{kh}) = \frac{N_h - 1}{n_h N_h} S_{1kh}^2 - \frac{(n_h - 1)}{n_h K_h} S_{2kh}^2. \tag{4.6}$$

En esta nueva expresión,

$$S_{1kh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(y_{khi} - \overline{Y}_{kh}\right)^2 \qquad \text{y} \qquad S_{2kh}^2 = \frac{1}{K_h - 1} \sum_{r=1}^{K_h} \left(\overline{Y}_{khr} - \overline{Y}_{kh}\right)^2 .$$

S_{1kh}^2 representa la varianza real entre los votos del estrato y S_{2kh}^2 la varianza de las votaciones entre los $AREs$. Ahora, estas cantidades ya no dependen de la varianza de votos en el interior de los $AREs$, por lo que debe ser posible estimarlos. Es importante resaltar que, para garantizar que (4.6) sea positivo, se debe satisfacer la siguiente condición sobre el tamaño de muestra en el estrato h:

$$\frac{K_h(N_h - 1)}{N_h} \frac{S_{1kh}^2}{S_{2kh}^2} > n - 1 \tag{4.7}$$

Esta condición, sumada a la restricción $K_h \geq n_h$ garantizan la no negatividad de la varianza analítica (4.6).

4.4.2. VARIANZA ESTIMADA

Construir un estimador insesgado para (4.6), puede ser visto como encontrar estimadores insesgados para S^2_{1kh} y S^2_{2kh}. Sin embargo, como se ha discutido en el **Apéndice** B, esto resulta difícil para el tipo de selección que se está considerando, para aplicar las técnicas clásicas se necesitan al menos 2 casillas por ARE, lo cual difiere con el objetivo de la tesis (garantizar una casilla por CAE). No obstante, tras hacer varios intentos por definir un buen estimador, se implementó la siguiente idea.

$$v(\bar{y}_{kh}) = \frac{N_h - 1}{n_h N_h} s^2_{1kh} - \frac{(n_h - 1)}{n_h K_h} s^2_{2kh}, \tag{4.8}$$

con

$$s^2_{1kh} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{khi} - \bar{y}_{kh})^2 \qquad \text{y} \qquad s^2_{2kh} = \frac{1}{n_h - 1} \sum_{r=1}^{n_h} (\bar{y}_{khr} - \bar{y}_{kh})^2 \tag{4.9}$$

Como es de esperarse este estimador es sesgado ya que s^2_{1kh} y s^2_{2kh} son sesgados para los valores poblacionales, la prueba de esto se ve en la expresión $(B.25)$ del mismo Apéndice. A pesar de ello, este estimador sirvió como punto de partida para proponer un nuevo estimador que corrige de manera significativa el sesgo. La deducción se basa en hacer algunos cambios a las expresiones (4.9), consiguiendo nuevos valores s^2_{1kh*} y s^2_{2kh*}. Para conseguir s^2_{1kh*}, el proceso consiste en tomar m muestras con remplazo de las casillas seleccionadas con la estrategia descrita en la Sección 4.3, calculando s^2_{1kh} en cada selección, esto es, aplicar un proceso de remuestreo. Entonces, el valor de s^2_{1kh*} será la media de estos m nuevos valores. Para s^2_{2kh*} se sigue un proceso distinto, primero se asigna un orden a las casillas, luego se calcula la media de cada dos casillas consecutivas. Así, se tiene $n_h - 1$ valores para la media de las casillas en los AREs, finalmente se toma la media de estos nuevos valores como s^2_{2kh*}. Este procedimiento se puede consultar con mayor detalle en el **Apéndice** B. Por tanto, el estimador que se emplea en los cálculos sucesivos es de la forma

$$v(\bar{y}_{kh}) = \frac{N_h - 1}{n_h N_h} s^2_{1kh*} - \frac{(n_h - 1)}{n_h K_h} s^2_{2kh*}. \tag{4.10}$$

4.5. PROPORCIÓN DE VOTOS Y SU VARIANZA

Para estimar la proporción de votos es necesario definir al estimador de las votaciones totales para los candidatos (y_k) y el estimador de las votaciones totales en la elección (x). En este punto, se hace uso de los estimadores habituales dados por muestreo estratificado:

$$y_k = \sum_{h=1}^{H} N_h \bar{y}_{hk} \qquad \text{y} \qquad x = \sum_{h=1}^{H} N_h \bar{x}_h.$$

Luego, el estimador de la proporción de votos (p) a favor de un candidato k se consigue de la razón entre estas cantidades,

$$\hat{p}_k = \frac{y_k}{x}. \tag{4.11}$$

Este estimador se conoce como el estimador combinado para un muestreo estratificado. Para calcular su varianza se emplea una aproximación de Taylor de primer orden, es decir:

$$\hat{p}_k \approx p_k + \frac{1}{X}\left(y_k - p_k x\right),$$

donde p_k es la proporción real de votos para el candidato k. De esta forma, la varianza aproximada de $\hat{p}_k$ se obtiene por:

$$\begin{aligned} V(\hat{p}_k) &= \frac{1}{X^2} V\left(y_k - p_k x\right) \\ &= \frac{1}{X^2} V\left(\sum_{h=1}^{H} N_h(\bar{y}_{kh} - p_k \bar{x}_h)\right) \\ &= \frac{1}{X^2} V\left(\sum_{h=1}^{H} N_h \bar{d}_{kh}\right) \\ &= \frac{1}{X^2} \sum_{h=1}^{H} N_h^2 V\left(\bar{d}_{kh}\right). \end{aligned}$$

con $\bar{d}_{kh} = \bar{y}_{kh} - p_k \bar{x}_{kh} = \dfrac{1}{n_h} \displaystyle\sum_{i=1}^{n_h} d_{khi}$, $d_{khi} = y_{khi} - p_k x_{hi}$. Esta forma de expresar a $\bar{d}_{kh}$, con la variable auxiliar d_{khi}, permite emplear la expresión (4.6) para calcular su varianza. Así,

$$V(\bar{d}_{kh}) = \frac{N_h - 1}{n_h N_h} S^2_{1dh} - \frac{(n_h - 1)}{n_h K_h} S^2_{2dh}, \tag{4.12}$$

donde S^2_{1dh} y S^2_{2dh} corresponden a valores poblacionales para la variable auxiliar d_{khi},

$$S_{1dh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(d_{khi} - \overline{D}_{kh} \right)^2 \qquad \text{y} \qquad S_{2dh}^2 = \frac{1}{K_h - 1} \sum_{r=1}^{K_h} \left(\overline{D}_{khr} - \overline{D}_{kh} \right)^2 .$$

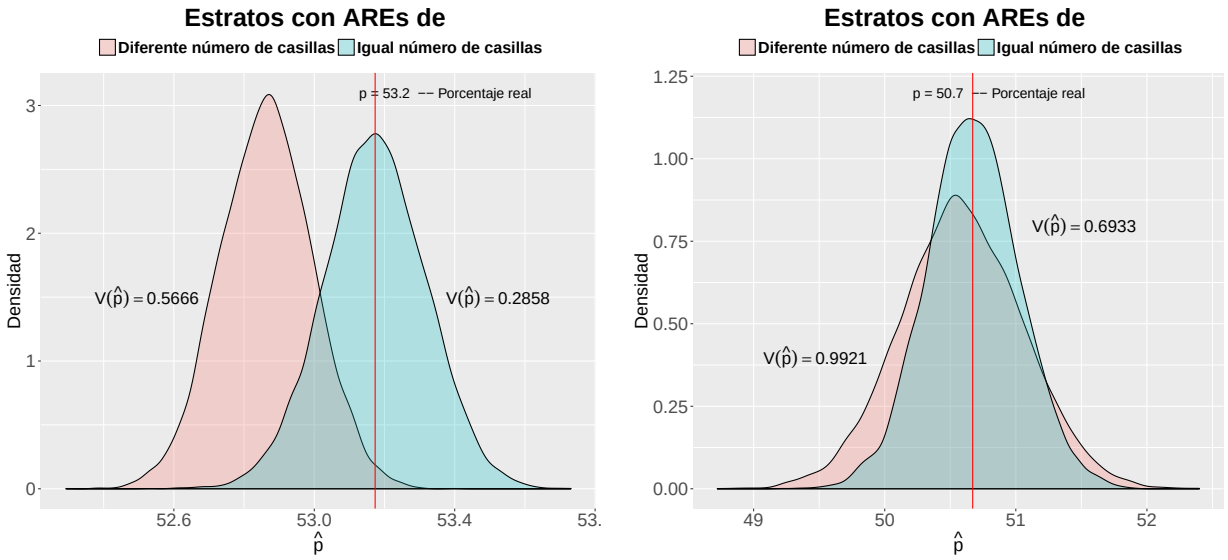
$$\overline{D}_{khr} = \frac{1}{C_h} \sum_{i=1}^{C_h} d_{khi} \equiv \text{la media de la variable auxiliar en el } ARE_r \text{ del estrato } h$$

$$\overline{D}_{kh} = \frac{1}{N_h} \sum_{i=1}^{N_h} d_{khi} \equiv \text{la media de estas variables en el estrato } h$$

Por lo tanto, la expresión final de la varianza de $\hat{p}_k$ es

$$V(\hat{p}_k) = \frac{1}{X^2} \sum_{h=1}^{H} N_h^2 \left(\frac{N_h - 1}{n_h N_h} S_{1dh}^2 - \frac{(n_h - 1)}{n_h K_h} S_{2dh}^2 \right). \tag{4.13}$$

Hay que notar que la no negatividad de esta última expresión es heredada de garantizar la no negatividad de la varianza en cada estrato h, ver expresión (4.7). Por otra parte, cabe resaltar que ésta expresión esta diseñada para hacer el cálculo cuando se tienen estratos formados por $AREs$ con el mismo número de casillas, de lo contrario, la expresión sería mucho más complicada. Pero, en este punto, queremos mostrar lo que sucede cuando no se usa esta restricción. Como ejemplo, se muestra la siguiente gráfica: distribución estimada con $10,000$ muestras para la proporción de votos estimada por la expresión (4.11). Estas estimaciones corresponden al porcentaje de votos para Andrés Manuel López Obrador, empleando la base de datos de la elección presidencial de 2018 y al porcentaje de votos para la coalición PRI_PVEM_PSDYUC, usando la base de datos de la elección a gobernador YUCATÁN 2012. Los resultados de seleccionar casillas, en dos etapas, de estratos con $AREs$ de distinto número de casillas se presenta en color rojo y de color azul los resultados al estratificar con $AREs$ con igual número de casillas, tal como se define en la sección 4.2.



(a) PRESIDENCIA 2018

(b) YUCATÁN 2012

Gráfica 4-1.: Distribución estimada del porcentaje de votos

Como se puede observar en la Gráfica **4-1**, considerar estratos con $AREs$ de distinto número de casillas resulta en una distribución con sesgo a la izquierda, esto es, que las estimaciones obtenidas de cada muestra subestiman de forma significativa al porcentaje real de votos (linea de color rojo). No obstante, la misma gráfica muestra que el sesgo en la elección de YUCATÁN es menor comparado a la elección presidencial donde el sesgo es muy marcado. Mientras que, construir estratos de $AREs$ con el mismo número de casillas nos da una distribución estimada que centra al porcentaje real de votos. Esto significa, por supuesto, que la estratificación propuesta es adecuada para realizar las estimaciones correspondientes a la elección. Esto nos lleva de nuevo a la expresión (4.4), en donde se comentó este punto: ahora es posible observar el efecto de no forzar a que el estimador fuera insesgado.

4.5.1. VARIANZA ESTIMADA

Dadas las observaciones realizadas en la sección 4.4.2, un estimador sesgado para la varianza de $\hat{p}_k$ está dada por:

$$v(\hat{p}) = \frac{1}{X^2}\sum_{h=1}^{H} N_h^2 \left(\frac{N_h - 1}{n_h N_h} s_{1dh}^2 - \frac{(n_h - 1)}{n_h K_h} s_{2dh}^2 \right), \tag{4.14}$$

con s_{1dh}^2 y s_{2dh}^2 son estimadores muestrales para S_{1dh}^2 y S_{2dh}^2 respectivamente. La variable auxiliar que se forma con los elementos de la muestra es $\hat{d}_{khi} = y_{khi} - \hat{p}_k x_{hi}$, por tanto

$$s^2_{1dh} = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(\hat{d}_{khi} - \hat{\bar{d}}_{kh}\right)^2 \qquad \text{y} \qquad s^2_{2dh} = \frac{1}{n_h - 1}\sum_{r=1}^{n_h}\left(\hat{\bar{d}}_{khr} - \hat{\bar{d}}_{kh}\right)^2.$$

Las siguientes cantidades refieren a la variable auxiliar $\hat{d}_{khi}$, en el estrato h:

$$t_{khr} = C_h \sum_{i\in\mathcal{S}_{hr}} \hat{d}_{hki} = C_h \hat{d}_{hki} \equiv \text{ Total estimado en el } ARE_r$$

$$\hat{\bar{d}}_{khr} = \frac{t_{khr}}{C_h} = \hat{d}_{hki} \equiv \text{Media estimada en el } ARE_r$$

$$\hat{\bar{d}}_{kh} = \frac{1}{n_h}\sum_{i=1}^{n_h} \hat{d}_{khi} \equiv \text{Media estimada en el estrato.}$$

Por las mismas razones mencionadas anteriormente, la expresión (4.14) es un estimador sesgado para la varianza (4.13). Entonces, para corregir el sesgo, se toma a s^2_{1dh} y s^2_{2dh}, y se realizan las modificaciones necesarias, tal como se hace para s^2_{1h} y s^2_{2h},ver la Sección 4.4.2. Es decir, en este caso, se deben usar los pasos descritos en el Apéndice B empleando la variable auxiliar $\hat{d}_{khi} = y_{khi} - \hat{p}_k x_{hi}$ en lugar de y_{khi}. Con esto, el estimador que se empleará para los ejemplos sucesivos, y como estimador final para la varianza estimada de $\hat{p}_k$, es de la forma:

$$\dot{v}(\hat{p}) = \frac{1}{X^2}\sum_{h=1}^{H} N_h^2 \left(\frac{N_h - 1}{n_h N_h} s^2_{1dh*} - \frac{(n_h - 1)}{n_h K_h} s^2_{2dh*}\right). \tag{4.15}$$

4.5.2. TAMAÑOS DE MUESTRA, MARGEN DE ERROR E INTERVALO DE CONFIANZA

El tamaño de muestra (n) se elije vía simulación, esto es, se toman diferentes tamaños de muestra y se calcula el margen de error real para el partido con mayor varianza en los votos a favor. Para este proceso, n está sujeto al número de $AREs$ (K) en la población, es decir, $n \leq K$. Además, si suponemos que las votaciones siguen una distribución normal, podemos calcular el margen de error por:

$$\epsilon = Z_{\alpha/2}\sqrt{V(\hat{p}_k)}, \tag{4.16}$$

donde $Z_{\alpha/2}$ es el cuantil inferior de la normal estándar. Así, si se considera una confianza del 95 %, $Z_{\alpha/2}$ se remplaza por su correspondiente valor $Z_{0.025} = 1.96$. Por otra parte, dada la muestra de n casillas, un intervalo de confianza del 95 % para el porcentaje de votos estimado

del candidato k esta dado por:

$$\left(\hat{p}_k - Z_{0.025}\sqrt{\dot{v}(\hat{p}_k)}, \hat{p}_k + Z_{0.025}\sqrt{\dot{v}(\hat{p}_k)}\right). \tag{4.17}$$

4.6. APLICACIÓN AL CONTEO RÁPIDO

En esta Sección, se aplica la estrategia de selección y estimación descrita anteriormente a diferentes bases de datos, comparando los resultados obtenidos con los del método estratificado tradicional.

- Elecciones locales 2012: CDMX, GUANAJUATO, JALISCO, MORELOS, TABASCO, CHIAPAS y YUCATÁN
- Elecciones locales 2016: PUEBLA y VERACRUZ
- ELECCIÓN FEDERAL 2018

Los detalles de estas bases de datos, así como la adecuación necesaria para su implementación se describen en el Apéndice D.

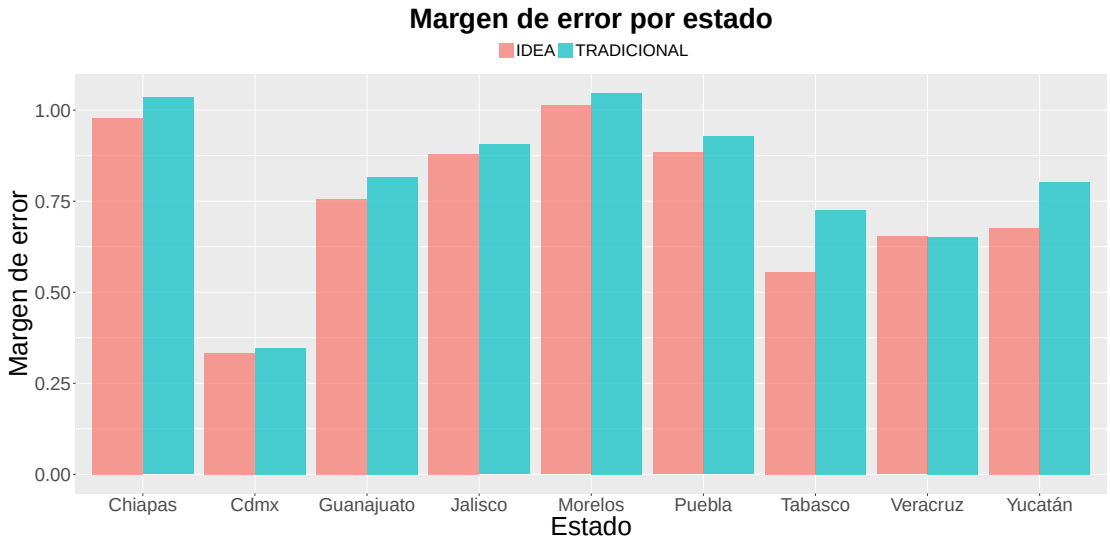
4.6.1. ELECCIONES LOCALES

Para realizar las comparaciones, se usan los tamaños de muestra empleados para las estimaciones de los Conteos Rápidos 2018.

	Estado								
	CHIAPAS	CDMX	GUANAJUATO	JALISCO	MORELOS	PUEBLA	TABASCO	VERACRUZ	YUCATÁN
n	500	1,100	500	467	200	509	450	1,100	300
K	1,434	1,668	1,350	1,724	439	1,464	575	2,186	475

Tabla 4-1.: Tamaños de muestra 2018

Como primera observación, vemos que los tamaños de muestra (n) son menores al número de $CAEs$ (K) en cada elección, es decir, $n \leq K$. Con estos tamaños de muestra, se tiene la siguiente gráfica con el margen de error esperado en cada estado para el partido o coalición de mayor varianza.



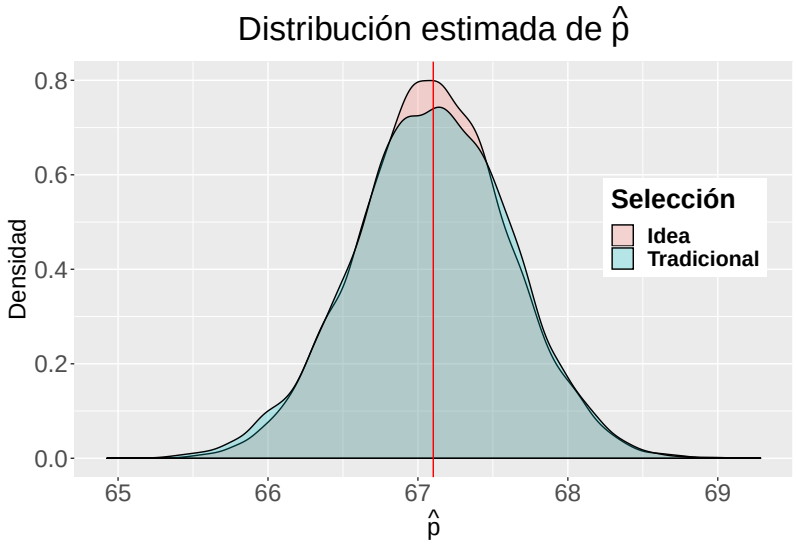
Gráfica 4-2.: Margen de error poblacional, con nivel de confianza del 95 %

En esta gráfica se aprecia que en casi todos los estados se tiene menor margen de error con el método propuesto. Únicamente el estado de VERACRUZ el margen de error es igual al margen dado por el método tradicional. Por lo tanto, esto prueba que la estrategia propuesta mejora el margen de error al mismo tiempo que resuelve el problema operativo referente a los $CAEs$, asignándoles un única casilla.

DISTRIBUCIÓN ESTIMADA Y COBERTURA

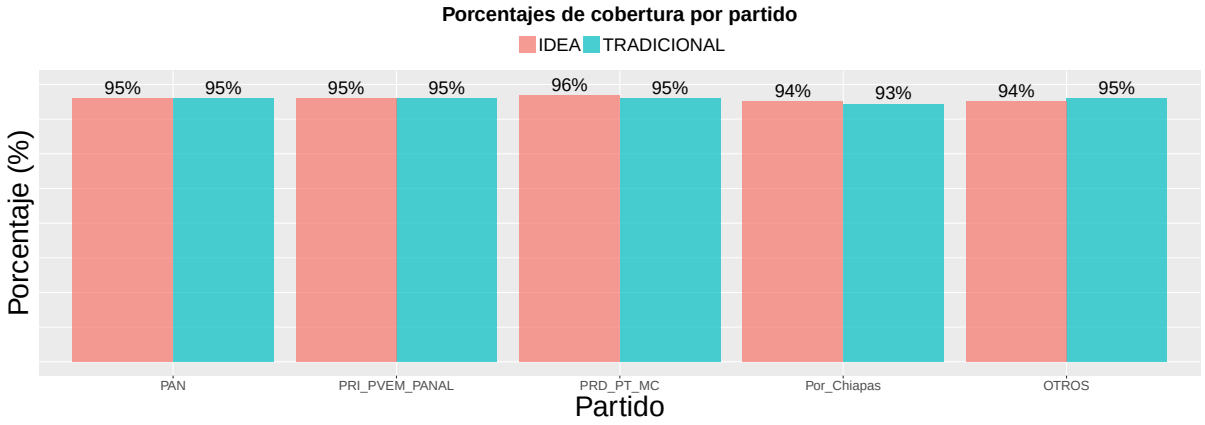
En esta sección se describen los resultados obtenidos de tomar $10,000$ muestras distintas de tamaño n en cada base de datos. Estos resultados corresponden por una parte a la distribución estimada para la proporción de votos de la coalición de mayor varianza en cada elección, empleando $\hat{p}_k$ como estimador. Por otra parte, se presenta la cobertura de los intervalos con 95 % de confianza (cobertura refiere al % de los intervalos, que se generan por cada muestra, que contienen el valor poblacional que se está estimando y que debe ser semejante al nivel de confianza especificado).

Por ejemplo, para CHIAPAS la coalición con mayor varianza fue PRI_PVEM_PANAL. Entonces, la distribución muestral del estimador del porcentaje real de votos para este partido es:



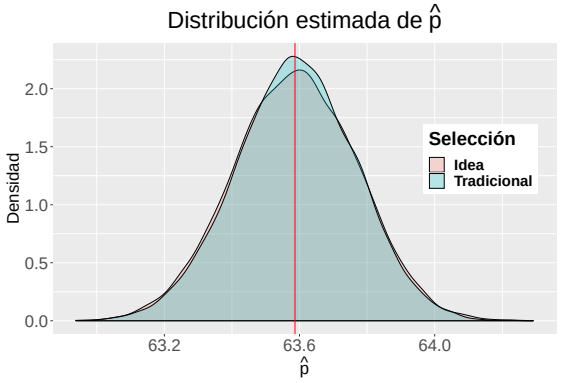
Gráfica 4-3.: Coalición PRI_PVEM_PANAL, Chiapas 2012

En esta gráfica se puede ver cómo ambas estrategias proporcionan una distribución muy parecida, pero que la estrategia de selección propuesta centra mejor las estimaciones. También se observa que el porcentaje real de votos, linea de color rojo, esta centrado en esta distribución. Por otro lado, la cobertura para cada partido, empleando la expresión (4.17) para calcular los intervalos de confianza, es:

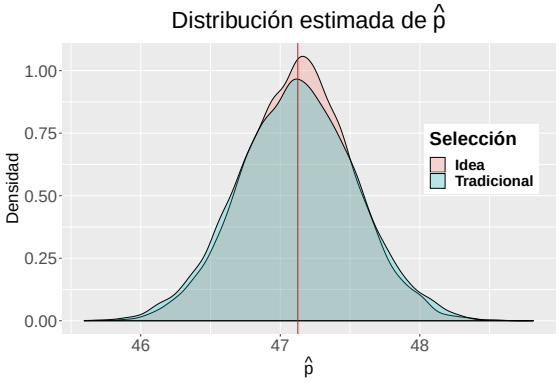


Gráfica 4-4.: Elección Chiapas 2012

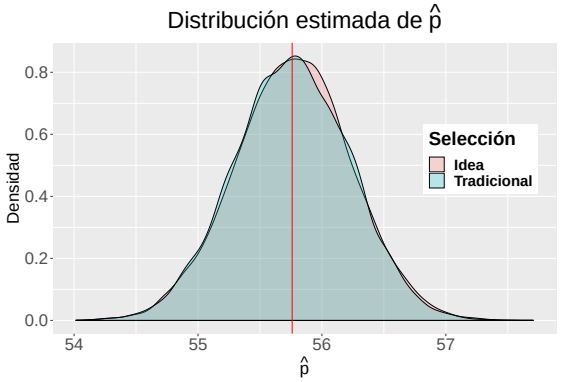
Para esta elección en particular, ambos métodos proporcionan buena cobertura del porcentaje real de votos por partido. Esto muestra que la propuesta es adecuada, pues es equiparable al método tradicional. Ahora, en las siguientes gráficas se muestran las distribuciones estimadas del estimador del porcentaje de votos de los partidos de mayor varianza en la elección de cada estado.



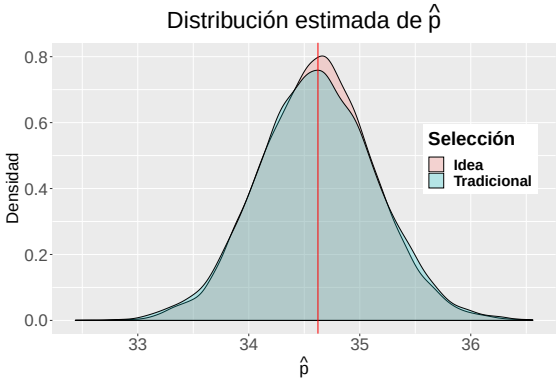
(a) Cdmx: PRD_PT_MC



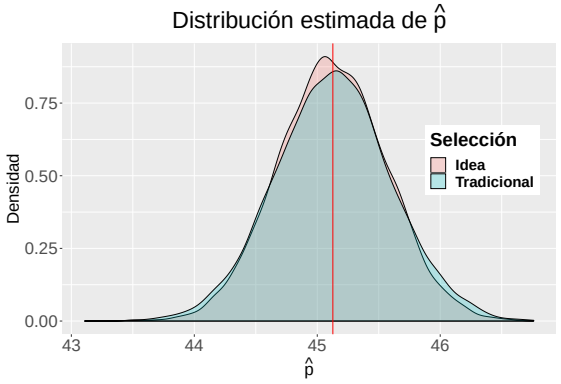
(b) Guanajuato: PAN_PANAL



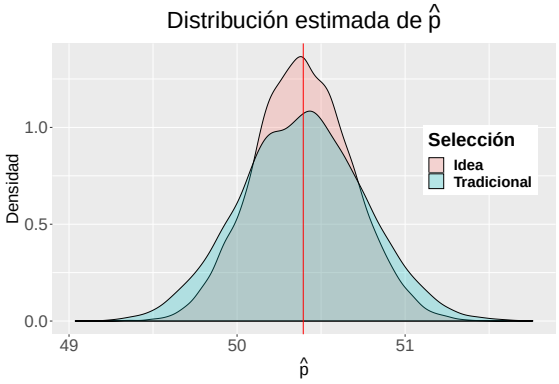
(c) Jalisco: PRI_PVEM



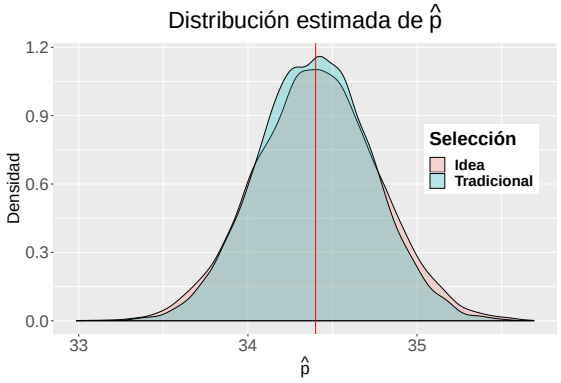
(d) Morelos: PRI_PVEM_PANAL



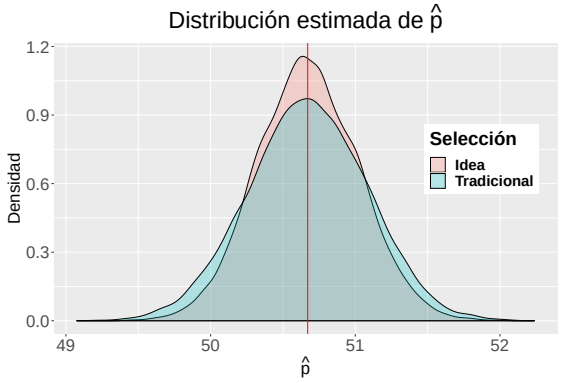
(e) Puebla: PAN_PT_NVA_ALIANZA_PCPP_PSI



(f) Tabasco: PRD_PT_MC



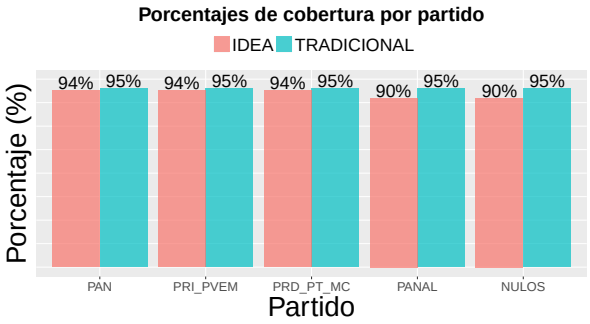
(g) Veracruz: PAN_PRD



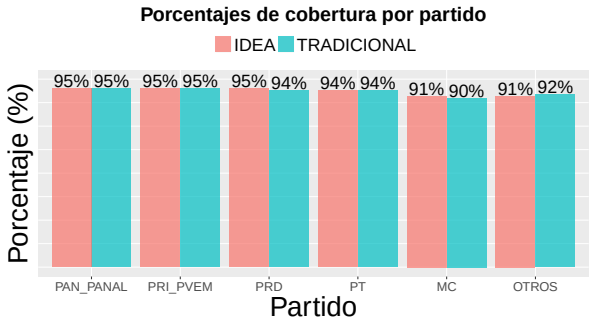
(h) Yucatán: PRI_PVEM_PSDYUC

En todas estas gráficas, el comportamiento de los porcentajes estimados por ambas estrategias es muy parecido, incluso en algunos casos (TABASCO y YUCATÁN) la estrategia de selección propuesta es mucho mejor que la tradicional. Por lo tanto, esto muestra que el método propuesto es tan eficiente como el procedimiento tradicional, con la ventaja que esta resuelve el problema operativo sobre los $CAEs$.

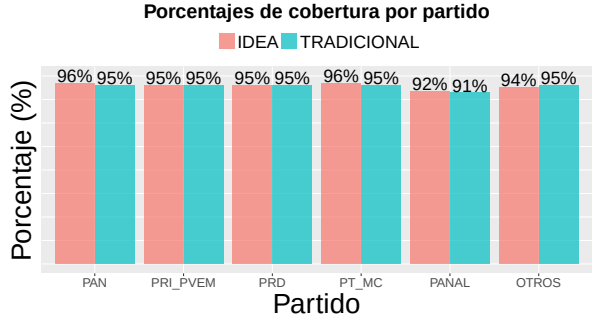
Para ver qué tan bien funciona el estimador de la varianza (4.15), se presentan las coberturas obtenidas por cada partido en cada elección. Recordar que para estas coberturas los intervalos se han calculado con un nivel de confianza del 95 %, empleando la expresión 4.17.



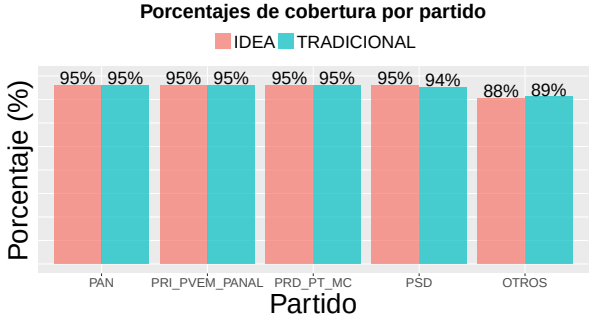
(i) Cdmx 2012



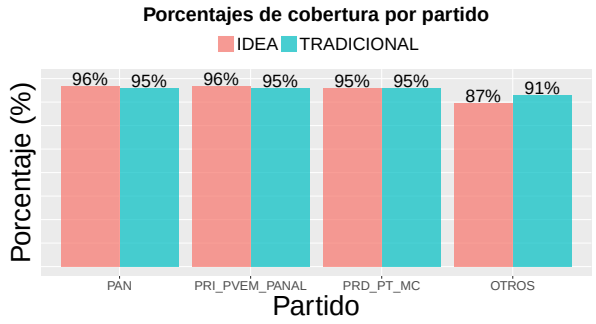
(j) Guanajuato 2012



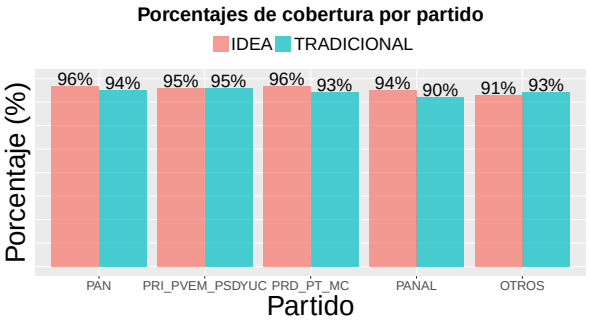
(k) Jalisco 2012



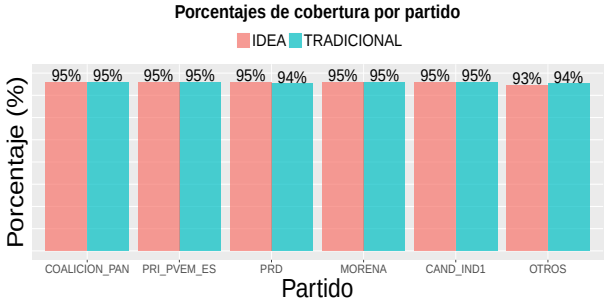
(l) Morelos 2012



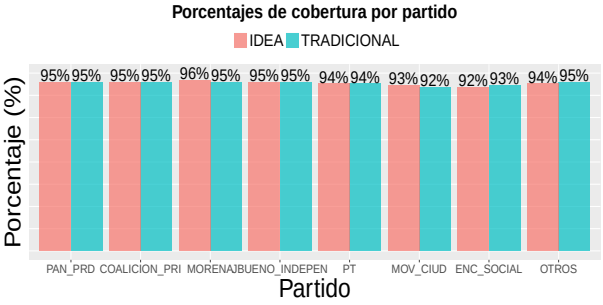
(m) Tabasco 2012



(n) Yucatán 2012



(ñ) Puebla 2016



(o) Veracruz 2016

En general, los resultados proporcionan coberturas del 95 %, salvo en algunos casos, que corresponden a los partidos con porcentajes de votos muy pequeños, por decir, menores del 2 %. Por lo tanto, podemos concluir que con la estrategia propuesta se obtienen buenos resultados, esto es, un margen de error menor al obtenido con el método tradicional y que los porcentajes de cobertura satisfacen, en su mayoría, el nivel de confianza esperado. Ademas, como ya se ha mencionado antes, a diferencia del método tradicional, ésta propuesta resuelve la restricción operativa sobre los $CAEs$, que es el objetivo principal de la tesis.

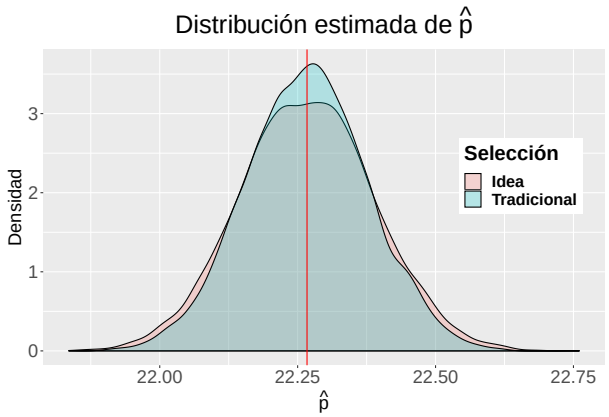
4.6.2. ELECCIÓN FEDERAL 2018

Para las estimaciones de la elección federal, el COTECORA consideró un tamaño de muestra de $n = 7,500$ casillas. Con este tamaño de muestra se obtiene el margen de error mostrado en la tabla **4-2**, tomando siempre como referencia al partido o candidato con mayor varianza y buscando una confianza del 95 %. En este caso la mayor varianza fue para Andrés Manuel López Obrador (AMLO), por lo que el margen de error, para el método propuesto y el tradicional es:

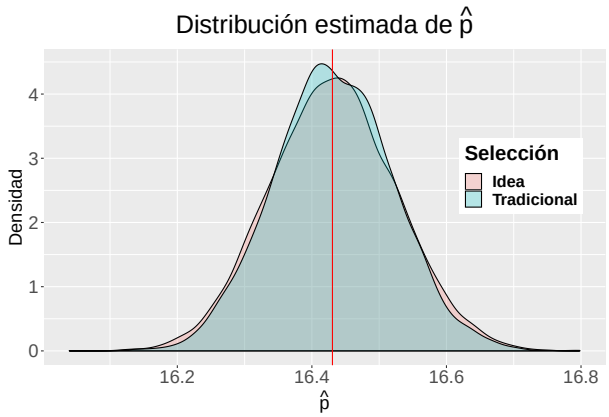
	Tradicional	Propuesta
ϵ	0.267580	0.284725

Tabla 4-2.: Margen de error Federal

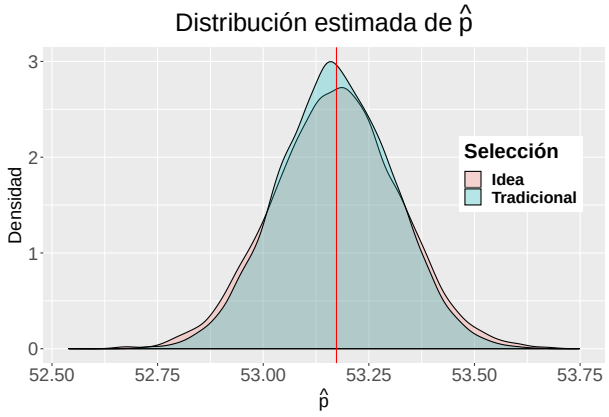
Esta tabla muestra que el margen calculado por el método tradicional es relativamente mejor que el calculado con la propuesta, sin embargo considerando el objetivo principal de nuestra propuesta este margen de error es aceptable. Por otra parte, también se calcula la distribución estimada del porcentaje de votos para los principales candidatos en la elección, después de AMLO, estos fueron: Ricardo Anaya Cortés (RAC) y José Antonio Meade Kuribreña (JAMK).



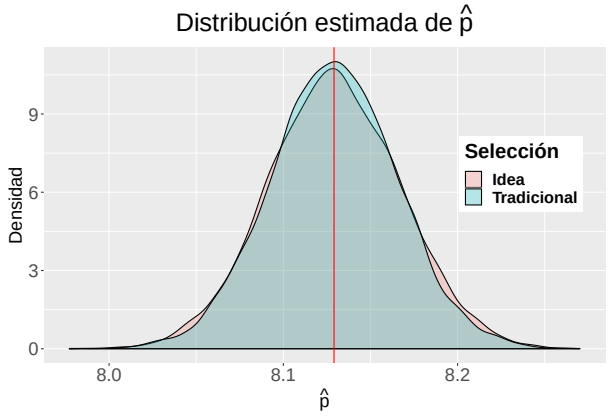
(p) RAC



(q) JAMK

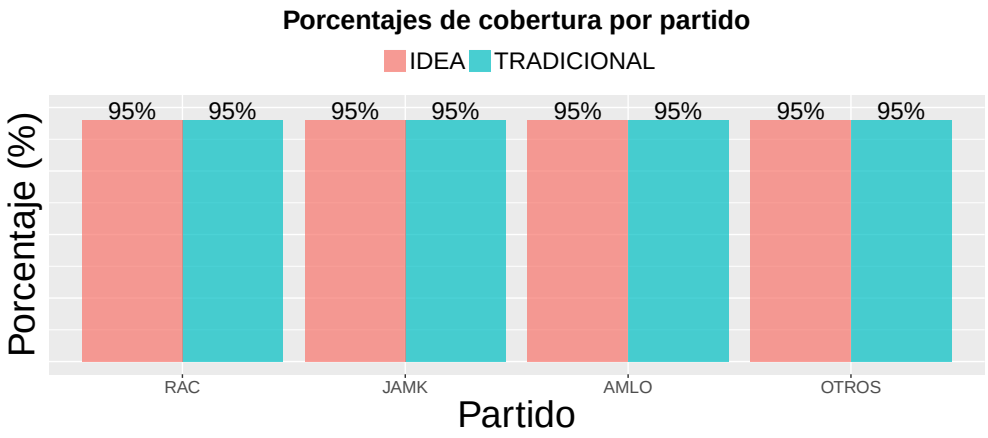


(r) AMLO



(s) OTROS

En estas gráficas se observa que la estrategia de selección propuesta es equiparable a la tradicional ya que el comportamiento que describen es similar y además ambas centran el porcentaje real de votos. Con cada una de las estimaciones, que describen las distribuciones anteriores, también se calculan intervalos de confianza para el verdadero valor del porcentaje de votos, en particular se toma un nivel de confianza del 95 %. Las siguiente gráfica muestra el porcentaje de estos intervalos que capturan al porcentaje real de votos por cada candidato a la elección presidencial.



Gráfica 4-5.: Elección Federal 2018

Esta gráfica muestra que para poblaciones con tamaño de muestra grande, ambos métodos proporcionan coberturas excelentes para el porcentaje real de votos, es decir, se satisface el nivel de confianza esperado. Por lo tanto, podemos concluir que los estimadores empleados para el método propuesto, en particular $\dot{v}(\hat{p}_k)$, son apropiados para realizar las estimaciones. Por último, es importante resaltar que aún cuando los resultados en la elección presidencial, donde el número de casillas es mucho mayor que en las elecciones locales, son parecidos, el método aquí propuesto tiene el extra de resolver la restricción operativa sobre los CAEs.

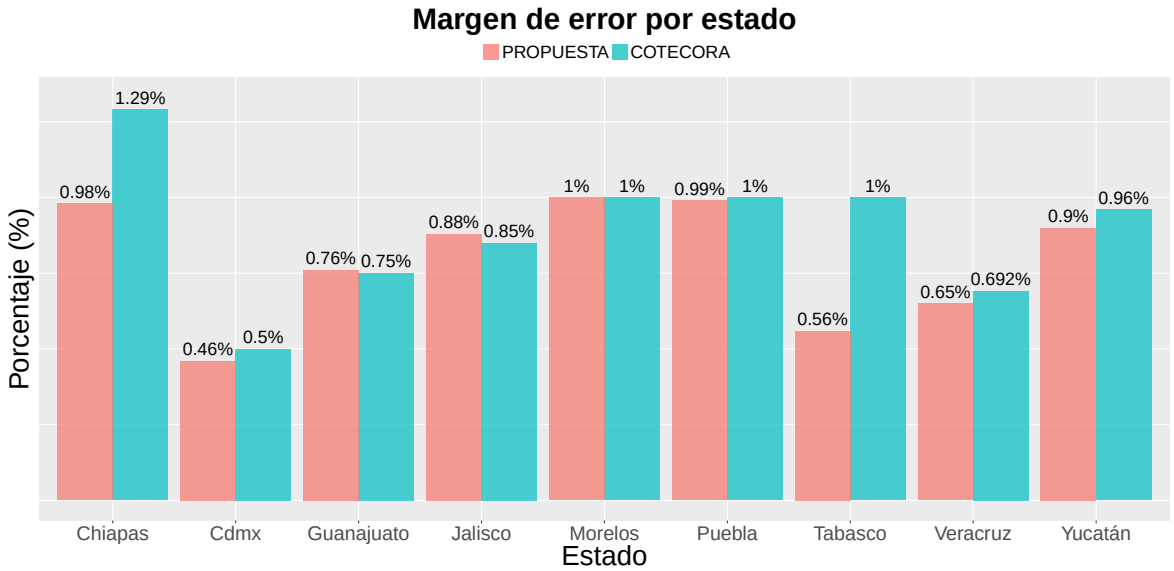
5. CONCLUSIONES

Si bien los resultados para el método propuesto son equiparables al método tradicional, este último aumenta el margen de error esperado al considerar la restricción operativa sobre los CAEs. Por ejemplo, para calcular el margen de error esperado en las elecciones del 2018, el Comité Técnico del Conteo Rápido (COTECORA) propuso los tamaños de muestra de la tabla **5-1**, forzando a que al menos el 80 % de los CAEs que colaboraron en el proceso del conteo rápido reportaran información de una sola casilla.

	CHIAPAS	CDMX	GUANAJUATO	JALISCO	MORALES	PUEBLA	TABASCO	VERACRUZ	YUCATÁN
n	500	700	500	467	200	424	450	1,100	205

Tabla 5-1.: Tamaños de muestra 2018

Con esto, el margen de error esperado, con una confianza del 95 %, para los partidos con mayor varianza en cada estado se muestran en la gráfica (**5-1**), donde las barras de color rojo corresponden al error que se hubiera obtenido con el método propuesto y de azul el error calculado por los integrantes del COTECORA para cada estado.



Gráfica 5-1.: Margen de error para los partidos de mayor varianza

De ésta gráfica se aprecia que el margen de error esperado, calculado por el método propuesto, que además de resolver la restricción operativa sobre los $CAEs$, es mejor que el empleado por los miembros del COTECORA para el Conteo Rápido del 2018. Incluso en algunos estados, la diferencia entre estos márgenes son significativos. Por ejemplo, en CHIAPAS y TABASCO se observan errores del 0.98 % y 0.56 % contra 1.26 % y 1 %, respectivamente. Además, todas las bases empleadas para los ejemplos del capítulo anterior, muestran que los estimadores empleados, $\hat{p}$ y $\dot{v}(\hat{p})$, estiman de forma adecuada a los valores poblacionales. Esto se puede ver con las distribuciones estimadas del porcentaje real y las coberturas, que, entre otras cosas, son equiparables al método usual.

Por otra parte, es importante notar que las coberturas observadas para algunos partidos en las elecciones locales 2012 y 2016, son menores a la confianza dada, tanto para el método propuesto como para el tradicional. Sin embargo, este es un error común y no se debe a la estrategia empleada, mas bien, este error se debe a que el número de votos para tales partidos es muy chico, en otras palabras el porcentaje de votos para estos es casi despreciable (menor del 2 %). Por lo tanto, calcular intervalos de confianza empleando la expresión 4.17, mejor conocida como intervalo de Wald, no es apropiado. Para corregir este error es necesario emplear otras herramientas estadísticas que contemplen proporciones pequeñas, tales como intervalos descritos por proporciones binomiales [Brown]. No obstante, abordar este tema no era objetivo de ésta tesis.

También es importante mencionar que, con el método propuesto, el número de estratos en que se divide a la población es mucho menor y que a pesar de esto las estimaciones obtenidas son mejores o iguales a las que se obtiene con el método tradicional . En particular, para la Elección Presidencial de 2018 el COTECORA estratificó a la población de casillas en un total de 350 estratos, mientras que con la estratificación definida en la sección 4.2, se obtienen 184 estratos.

En suma, a pesar de los buenos resultados que se pueden obtener con este método, aún existen cosas a mejorar. Por ejemplo, mejorar el estimador de la varianza, $\dot{v}(\hat{p})$, que a pesar de arrojar buenos resultados, el proceso de simulación para obtener los porcentajes de cobertura es relativamente lento. Por otra parte, también sería importante buscar la forma de estratificar con $AREs$ de distinto número de casillas sin que las estimaciones subestimen o sobrestimen el porcentaje real de votos, como se observa en la gráfica **4-1**.

A. CONCEPTOS IMPORTANTES

DISTRIBUCIÓN MUESTRAL DE UN ESTIMADOR

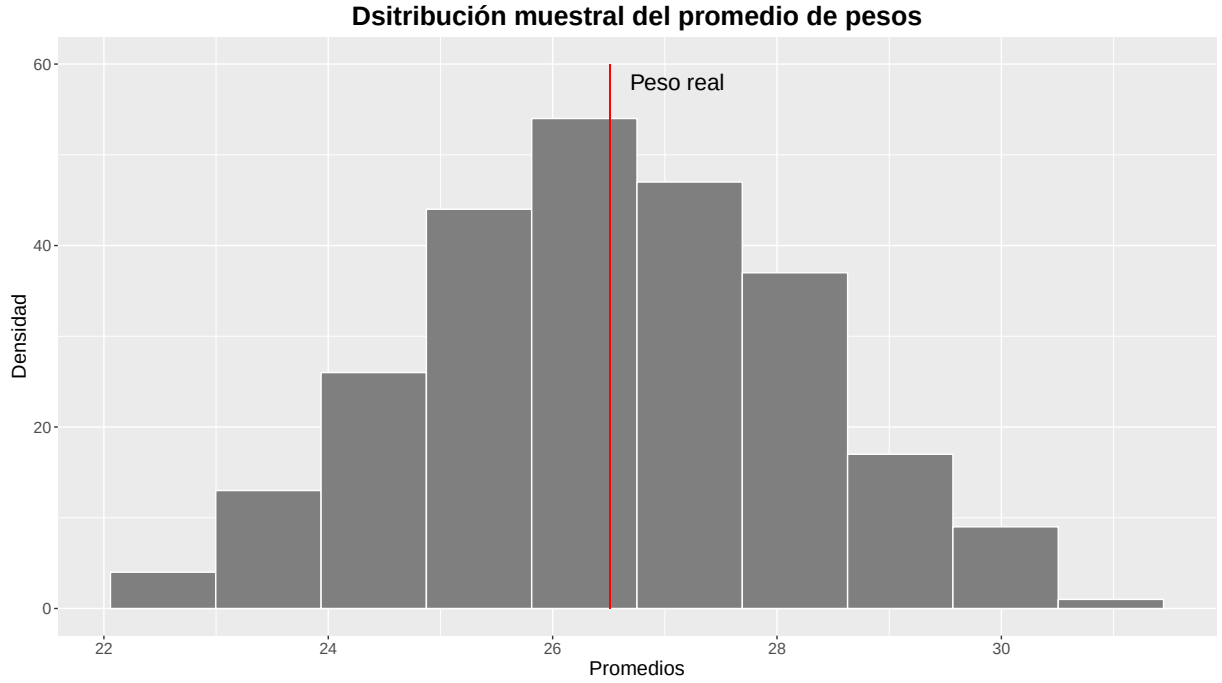
La distribución muestral de un estimador es la distribución de los valores que tomará el estimador considerando todas las muestras posibles. Esta distribución permite conocer el error esperado entre el estimador y el valor real del parámetro estudiado, construir intervalos de confianza y determinar el tamaño de muestra necesario para cometer un error especifico.

Teóricamente, para conocer esta distribución, deberíamos considerar todas las posibles muestras de un tamaño n, de la población. Sin embargo, tomar todas las muestras de la población es prácticamente imposible. Por ejemplo, tomando como referencia la base de datos de COLIMA con una población de 900 casillas. Si se selecciona un MASSR de 30 casillas, entonces habrían $9.803348 \text{x} 10^{55}$ muestras, esto es, las combinaciones de 900 en 30. Para conocer la distribución exacta del estimador de interés se tendría que calcular el valor del estimador para cada una de las muestras.

Para comprender mejor la idea de la distribución muestral consideremos un ejemplo genérico. Supongamos una población de 10 estudiantes de educación primaria y que se desea conocer el peso promedio de ellos tomando una muestra aleatoria sin remplazo de tamaño 5. Los pesos de cada estudiante se presentan en la siguiente tabla.

Estudiante	1	2	3	4	5	6	7	8	9	10
Peso	22.34	24.53	20.89	30.34	35.10	26.5	23.56	34.45	27.29	20.12

En este caso existe un total de 252 muestras posibles, esto es las combinaciones de 10 en 5. Si consideramos como estimador del promedio de pesos a la media entonces obtenemos la siguiente gráfica de distribución para el promedio del peso grupal.



Gráfica A-1.: Distribución muestral real

Aquí se aprecian todos los posibles valores que puede tomar el promedio de pesos del grupo al seleccionar una muestra aleatoria de 5 estudiantes. La linea de color rojo representa el valor real del peso grupal y se observa que está en el centro de la distribución muestral.

Para tomar todas las muestras posibles se empleó el paquete *comb*() del lenguaje de programación *R*. La instrucción básica para su ejecución es la siguiente:

$$combn(x, m, FUN = NULL, simplify = TRUE).$$

Este paquete genera todas las combinaciones de los elementos del vector x tomando m de estos. Si x es un entero positivo, devuelve todas las combinaciones de los elementos de la secuencia $1, 2, ..., x$ tomando m de estos. Con el argumento FUN se puede definir una función que será aplicada a cada combinación posible de m elementos. Si $simplify$ es $FALSE$, devuelve una lista; de lo contrario, devuelve un vector. Finalmente si $FUN = NULL$, entonces devuelve una matriz.

Así para obtener la media de cada una de las muestras posibles de tamaño 5 se ejecutó la siguiente instrucción

$$combn(x = pesos, m = 5, FUN = mean, simplify = TRUE).$$

MUESTRA REPRESENTATIVA

Cuando se habla de encuestas, el concepto de muestra representativa es mencionado a menudo en medios de comunicación, autoridades y por la población en general. A continuación se enlistan algunas definiciones tomadas de diferentes fuentes.

1. Una buena muestra será representativa en el sentido de que las características de interés en la población se pueden estimar a partir de la muestra con un grado de precisión conocido [Lohr].

2. Muestra elegida de tal manera que todas las partes de la población tienen la misma oportunidad de ser incluidas en la muestra [Ross].

3. Una muestra representativa es una pequeña cantidad de algo que refleja con precisión la entidad más grande. Un ejemplo es cuando un pequeño número de personas refleja con precisión a los miembros de toda una población. En una clase de 30 estudiantes, en la que la mitad de los estudiantes son varones y la mitad son mujeres, una muestra representativa podría incluir seis estudiantes: tres hombres y tres mujeres [Link].

4. Una muestra representativa es un grupo que se ajusta estrechamente a las características de su población en su conjunto. En otras palabras, la muestra es un reflejo bastante preciso de la población de la que se extrae la muestra [Link].

B. FORMA ANALÍTICA DE LA VARIANZA

Los siguientes resultados se obtienen al fijarse en un estrato.

NOTACIÓN

$N \equiv$ Número de casillas en la elección

$n \equiv$ Número de casillas en muestra, $n \leq K$

$K \equiv$ Número de AREs en la elección

$n' \equiv$ Número de AREs en muestra

$K_r = C \equiv$ Número de casillas en el ARE_r

$m_r \equiv$ Número casillas en muestra en el ARE_r

$\mathcal{S} \equiv$ Conjunto de casillas seleccionadas mediante las dos etapas

$\mathcal{S}' \equiv$ Conjunto de AREs seleccionados en la primera etapa

$\mathcal{S}_r \equiv$ Conjunto de AREs seleccionados en el ARE_r

$y_{kri} \equiv$ Votos en la i-ésima casilla del r-ésimo ARE para el candidato k

$y_{ki} \equiv$ Votos en la i-ésima casilla para el candidato k

$\overline{Y}_k = \frac{1}{N}\sum_{i=1}^{N}\overline{Y}_{ki} = \frac{1}{K}\sum_{r=1}^{K}\overline{Y}_{kr} \equiv$ Valor real de la media de votos para el k-ésimo candidato

$\overline{Y}_{kr} \equiv$ Valor real de la media de votos para el k-ésimo candidato en el ARE_r

$t_{kr} = \sum_{i=1}^{C} y_{kri} \equiv$ Total de votos para el candidato k en el ARE_r.

Teorema B.1. *Considere una población de N casillas distribuidas en K AREs con igual número de casillas ($K_r = C$ para $r = 1, 2, ..., K$). Suponga una selección de casillas en dos etapas de la siguiente forma:*

E1: Un MASSR de n' AREs.

E2: Un MASSR de una única casilla en cada ARE seleccionado en la primera etapa.

De esta forma $n = n'$, $m_r = 1$ y $N = KC$. Así, bajo esta estrategia de selección, un estimador insesgado para la media a favor del k-ésimo candidato es:

$$\bar{y}_k = \frac{1}{n} \sum_{i \in \mathbb{S}} y_{ki} \tag{B.1}$$

con varianza

$$V(\bar{y}_k) = \frac{N-1}{nN} S_1^2 - \frac{n-1}{nK} S_2^2 \tag{B.2}$$

donde

$$S_1^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_{ki} - \overline{Y}_k)^2 \qquad S_2^2 = \frac{1}{K-1} \sum_{r=1}^{K} (\overline{Y}_{kr} - \overline{Y}_k)^2$$

$$\overline{Y}_k = \frac{1}{N} \sum_{i=1}^{N} y_{ki} \qquad \overline{Y}_{kr} = \frac{1}{C} \sum_{i=1}^{C} y_{ki}$$

Demostración:

Considérese la siguiente variable auxiliar

$$Z_i = \begin{cases} 1 & \text{si la casilla } i \text{ pertenece a la muestra} \\ 0 & \text{en otro caso} \end{cases}$$

Entonces

$$P(Z_i = 1) = \sum_{r=1}^{K} P(Z_i = 1 | ARE_r) P(ARE_r).$$

$P(ARE_r)$ es la probabilidad de que el r-ésimo ARE sea seleccionado, esto es, $P(ARE_r) = \frac{n}{K}$ por ser una MASSR de AREs. Mientras que,

$$P(Z_i = 1|ARE_r) = \begin{cases} \frac{1}{C} & \text{si } y_{ki} \in ARE_r \\ \\ 0 & \text{en otro caso} \end{cases}$$

Por lo tanto,

$$\begin{aligned} P(Z_i = 1) &= \sum_{r=1}^{K} P(Z_i = 1|ARE_r)P(ARE_r) \\ &= \frac{n}{K} \sum_{r=1}^{K} P(Z_i = 1|ARE_r) \\ &= \frac{n}{K}\left(\frac{1}{C}\right) = \frac{n}{N}. \end{aligned} \tag{B.3}$$

Consecuentemente,

$$\mathbb{E}[Z_i] = 0P(Z_i = 0) + 1P(Z_i = 1) = P(Z_i = 1) = \frac{n}{N}. \tag{B.4}$$

Además, la probabilidad de que dos casillas, Z_i y Z_j pertenezcan a la muestra siendo de distintos $ARE s$, es

$$\begin{aligned} P(Z_i Z_j = 1) &= P(Z_j = 1|Z_i = 1)P(Z_i = 1) \\ &= \left(\frac{n-1}{K-1}\right)\left(\frac{1}{C}\right)\left(\frac{n}{N}\right) \\ &= \frac{n}{N}\left(\frac{n-1}{N-C}\right) \end{aligned} \tag{B.5}$$

En cambio, si ambas casillas pertenecen al mismo ARE

$$\begin{aligned} P(Z_i Z_j = 1) &= P(Z_j = 1|Z_i = 1)P(Z_i = 1) \\ &= 0\left(\frac{n}{N}\right) \\ &= 0 \end{aligned} \tag{B.6}$$

Por lo tanto, si las casillas i y j pertenecen a AREs diferentes, obtenemos que

$$\mathbb{E}[Z_i Z_j] = 0P(Z_i Z_j = 0) + 1P(Z_i Z_j = 1) \tag{B.7}$$
$$= P(Z_i Z_j = 1) \tag{B.8}$$
$$= \frac{n}{N}\left(\frac{n-1}{N-C}\right) \tag{B.9}$$

Mientras que si ambas casillas pertenecen al mismo ARE se tiene

$$\mathbb{E}[Z_i Z_j] = 0P(Z_i Z_j = 0) + 1P(Z_i Z_j = 1) \tag{B.10}$$
$$= P(Z_i Z_j = 1) \tag{B.11}$$
$$= 0 \tag{B.12}$$

Así,

$$\mathbb{E}[Z_i Z_j] = \begin{cases} \dfrac{n}{N}\left(\dfrac{n-1}{N-C}\right) & \text{si } i \text{ y } j \text{ pertenecen a distintos } AREs \\ \\ 0 & \text{si } i \text{ y } j \text{ pertenecen al mismo } ARE \end{cases}$$

Se prueba la expresión B.1

Demostración. Bajo un muestreo en dos etapas, la media de votos para el k-ésimo candidato se puede estimaría por

$$\bar{y}_k = \frac{1}{n}\sum_{r\in\mathcal{S}'}\sum_{i\in\mathcal{S}_r} y_{kri}.$$

Sin embargo, con la estrategia de selección propuesta, en la segunda etapa solo se considera un único elemento en cada ARE de la primera etapa por lo que podemos omitir a $\mathcal{S}_r$ y reescribir a $\bar{y}_k$ de tal forma que solo dependa de las las casillas seleccionadas,

$$\bar{y}_k = \frac{1}{n}\sum_{i\in\mathcal{S}} y_{ki}. \tag{B.13}$$

Para probar el insesgamiento de $\bar{y}_k$ se hace uso de la variable auxiliar Z_i, es decir, se reescribe a $\bar{y}_k$ como

$$\bar{y}_k = \frac{1}{n}\sum_{i\in\mathcal{S}} y_{ki} = \frac{1}{n}\sum_{i=1}^{N} Z_i y_{ki} \tag{B.14}$$

Ahora es fácil ver que

$$\mathbb{E}[\bar{y}_k] = \frac{1}{n}\sum_{i=1}^{N} y_{ki}\mathbb{E}[Z_i] = \frac{1}{n}\sum_{i=1}^{N} y_{ki}(\frac{n}{N}) = \overline{Y}_k.$$

□

Se prueba la expresión B.2

Demostración. Para obtener la varianza de $\bar{y}_k$ se hace uso de la forma $(B.14)$,

$$\begin{aligned} V(\bar{y}_k) &= \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_{ki}\right) \\ &= \frac{1}{n^2}\sum_{i,j=1}^{N} y_{ki}y_{kj}Cov(Z_i, Z_j) \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_{ki}^2 Var(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_{ki}y_{kj}Cov(Z_i, Z_j)\right] \end{aligned} \tag{B.15}$$

Luego,

$$Var(Z_i) = \mathbb{E}[Z_i^2] - \mathbb{E}[Z_i]^2 = \mathbb{E}[Z_i] - \mathbb{E}[Z_i]^2 = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right). \tag{B.16}$$

En cambio, la covarianza se debe calcular en dos partes, para casillas en diferentes AREs y para casillas en el mismo ARE. Para esto se emplea el siguiente resultado

$$Cov(Z_i, Z_j) = \mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j].$$

Entonces,

para casillas i y j en AREs diferentes

$$\begin{aligned} Cov(Z_i, Z_j) &= \mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j] \\ &= \frac{n}{N}\left(\frac{n-1}{N-C}\right) - \left(\frac{n}{N}\right)^2 \qquad \text{(por ser } \mathbb{E}[Z_i] = \mathbb{E}[Z_j]\text{)} \\ &= -\frac{n}{N^2}\left(\frac{N-nC}{N-C}\right) \end{aligned} \tag{B.17}$$

para casillas i y j en el mismo ARE

$$\begin{aligned} Cov(Z_i, Z_j) &= \mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j] \\ &= 0 - \left(\frac{n}{N}\right)^2 \qquad \text{(por ser } \mathbb{E}[Z_i] = \mathbb{E}[Z_j]\text{)} \\ &= -\left(\frac{n}{N}\right)^2 \end{aligned} \tag{B.18}$$

Luego, para poder sustituir estas expresiones en $(B.15)$ es necesario separar a la suma

$$\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_{ki}y_{kj} \tag{B.19}$$

en dos partes: **Casillas en el mismo ARE** + **Casillas en AREs distintos**. Por tanto, se tienen que expresar a $(B.19)$ en término de las AREs.

$$\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_{ki}y_{kj} = \sum_{r=1}^{K}\sum_{i=1}^{C}\sum_{r'=1}^{K}\sum_{\substack{j\neq i\\(\text{si } r'=r)}}^{C} y_{kri}y_{kr'j}$$

$$= \sum_{r=1}^{K}\sum_{i=1}^{C}\left(\sum_{\substack{j=1\\j\neq i}}^{C} y_{kri}y_{krj} + \sum_{\substack{r'=1\\r'\neq r}}^{K}\sum_{j=1}^{C} y_{kri}y_{kr'j}\right)$$

$$= \sum_{r=1}^{K}\sum_{i=1}^{C}\sum_{\substack{j=1\\j\neq i}}^{C} y_{kri}y_{krj} + \sum_{r=1}^{K}\sum_{i=1}^{C}\sum_{\substack{r'=1\\r'\neq r}}^{K}\sum_{j=1}^{C} y_{kri}y_{kr'j}$$

$$= \sum_{r=1}^{K}\left(\sum_{i=1}^{C}\sum_{j=1}^{C} y_{kri}y_{krj} - \sum_{i=1}^{C} y_{kri}^2\right) + \sum_{r=1}^{K}\sum_{i=1}^{C} y_{kri}\left(\sum_{r'=1}^{K}\sum_{j=1}^{C} y_{kr'j} - \sum_{j=1}^{C} y_{krj}\right)$$

$$= \sum_{r=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{C} y_{kri}y_{krj} - \sum_{r=1}^{K}\sum_{i=1}^{C} y_{kri}^2 + \sum_{r=1}^{K}\sum_{i=1}^{C} y_{kri}\sum_{r'=1}^{K}\sum_{j=1}^{C} y_{kr'j} - \sum_{r=1}^{K}\left(\sum_{i=1}^{C} y_{kri}\right)^2$$

$$= \sum_{r=1}^{K}\left(\sum_{i=1}^{C} y_{kri}\right)^2 - \sum_{r=1}^{K}\sum_{i=1}^{C} y_{kri}^2 + \sum_{r=1}^{K}\sum_{i=1}^{C} y_{kri}\sum_{r'=1}^{K}\sum_{j=1}^{C} y_{kr'j} - \sum_{r=1}^{K}\left(\sum_{i=1}^{C} y_{kri}\right)^2$$

$$= \sum_{r=1}^{K} t_{kr}^2 - \sum_{i=1}^{N} y_{ki}^2 + \sum_{r=1}^{K} t_{kr}\sum_{r'=1}^{K} t_{kr'} - \sum_{r=1}^{K} t_{kr}^2$$

$$= \sum_{r=1}^{K} t_{kr}^2 - \sum_{i=1}^{N} y_{ki}^2 + \left(\sum_{r=1}^{K} t_{kr}\right)^2 - \sum_{r=1}^{K} t_{kr}^2, \qquad \text{por ser } (\sum_{r=1}^{K} t_{kr} = \sum_{r'=1}^{K} t_{kr'})$$

$$= \sum_{r=1}^{K} t_{kr}^2 - \sum_{i=1}^{N} y_{ki}^2 + \left(\sum_{r=1}^{K} t_{kr}\right)^2 - \sum_{r=1}^{K} t_{kr}^2$$

$$= \underbrace{C^2\sum_{r=1}^{K}\bar{y}_{kr}^2 - \sum_{i=1}^{N} y_{ki}^2}_{\textbf{Casillas en el mismo ARE}} + \underbrace{N^2\overline{Y}_k^2 - C^2\sum_{r=1}^{K}\bar{y}_{kr}^2}_{\textbf{Casillas en AREs distintos}} \quad . \tag{B.20}$$

Ahora podemos sustituir ($B.17$), ($B.18$) y ($B.20$) en ($B.15$) y obtener que

$$
\begin{aligned}
V(\bar{y}_k) &= \frac{1}{n^2}\sum_{i=1}^{N} y_{ki}^2 Var(Z_i) + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_{ki}y_{kj}Cov(Z_i,Z_j) \\
&= \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N} y_{ki}^2 - \underbrace{\frac{1}{n^2}\left(\frac{n}{N}\right)^2\left(C^2\sum_{r=1}^{K}\bar{y}_r^2 - \sum_{i=1}^{N} y_{ki}^2\right)}_{\textbf{Parte mismo ARE}} - \\
&\quad \underbrace{\frac{1}{n^2}\frac{n}{N^2}\left(\frac{N-nC}{N-C}\right)\left(N^2\overline{Y}_k^2 - C^2\sum_{r=1}^{K}\bar{y}_r^2\right)}_{\textbf{Parte AREs distintos}} \\
&= \frac{1}{nN}\left[\left(1-\frac{n}{N}\right)+\frac{n}{N}\right]\sum_{i=1}^{N} y_{ki}^2 + \left[\frac{1}{nN^2}\left(\frac{N-nC}{N-C}\right) - \frac{1}{N^2}\right]C^2\sum_{r=1}^{K}\bar{y}_r^2 \\
&\quad - \frac{1}{n}\left(\frac{N-nC}{N-C}\right)\overline{Y}_k^2 \\
&= \frac{1}{nN}\sum_{i=1}^{N} y_{ki}^2 + \frac{C^2}{nN^2}\left(\frac{N-nN}{N-C}\right)\sum_{r=1}^{K}\bar{y}_r^2 - \frac{1}{n}\left(\frac{N-nC}{N-C}\right)\overline{Y}_k^2 \\
&= \frac{1}{nN}\left[\sum_{i=1}^{N} y_{ki}^2 - N\overline{Y}_k^2 + N\overline{Y}_k^2\right] + \frac{C^2}{nN^2}\left(\frac{N-nN}{N-C}\right)\left[\sum_{r=1}^{K}\bar{y}_r^2 - K\overline{\overline{Y}}_k^2 + K\overline{\overline{Y}}_k^2\right] \\
&\quad - \frac{1}{n}\left(\frac{N-nC}{N-C}\right)\overline{Y}_k^2 \\
&= \frac{1}{nN}\left[(N-1)S_1^2 + N\overline{Y}_k^2\right] + \frac{C^2}{nN}\left(\frac{1-n}{N-C}\right)\left[(K-1)S_2^2 + K\overline{\overline{Y}}_k^2\right] \\
&\quad - \frac{1}{n}\left(\frac{N-nC}{N-C}\right)\overline{Y}_k^2. \qquad \text{(B.21)}
\end{aligned}
$$

En esta expresión,

$$
\overline{Y}_k = \frac{1}{N}\sum_{i=1}^{N} y_{ki} \qquad \text{y} \qquad \overline{\overline{Y}}_k = \frac{1}{K}\sum_{r=1}^{K}\bar{y}_r = \frac{1}{KC}\sum_{r=1}^{K}\sum_{i=1}^{C} y_{kri} = \overline{Y}_k.
$$

La segunda igualdad es posible gracias a que los $AREs$ tienen el mismo número de elementos. Así, sustituyendo $\overline{\overline{Y}}_k$ en $(B.21)$ y agrupando para $\overline{Y}_k^2$ obtenemos que

$$
V(\bar{y}_k) = \frac{N-1}{nN}S_1^2 + \frac{C^2}{nN}\left(\frac{1-n}{N-C}\right)(K-1)S_2^2 + \frac{1}{n}\left[1 + \frac{C^2}{N^2}\left(\frac{N-nN}{N-C}\right)K - \left(\frac{N-nC}{N-C}\right)\right]\overline{Y}_k^2.
$$

Sin embargo, el coeficiente de $\overline{Y}_k^2$ es igual a cero, ya que

$$\begin{aligned}
\frac{1}{n}\left[1+\frac{C^2}{N^2}\left(\frac{N-nN}{N-C}\right)K-\left(\frac{N-nC}{N-C}\right)\right] &= \frac{1}{n}\left[1+\frac{KC^2}{N^2}\left(\frac{N-nN}{N-C}\right)-\left(\frac{N-nC}{N-C}\right)\right] \\
&= \frac{1}{n}\left[1+\frac{1}{K}\left(\frac{N-nN}{N-C}\right)-\left(\frac{N-nC}{N-C}\right)\right] \\
&= \frac{1}{n}\left[\frac{K(N-C)+N-nN-K(N-nC)}{K(N-C)}\right] \\
&= \frac{1}{n}\left[\frac{KN-KC+N-nN-KN+nKC)}{K(N-C)}\right] \\
&= \frac{1}{n}\left[\frac{KN-N+N-nN-KN+nN)}{K(N-C)}\right] \\
&= 0.
\end{aligned}$$

Por lo tanto, la expresión final para la varianza de la media de votos es

$$\begin{aligned}
V(\bar{y}_k) &= \frac{N-1}{nN}S_1^2+\frac{C^2(1-n)}{nN(N-C)}(K-1)S_2^2 \\
&= \frac{N-1}{nN}S_1^2+\frac{C^2(1-n)}{nKC^2(K-1)}(K-1)S_2^2 \\
&= \frac{N-1}{nN}S_1^2-\frac{n-1}{nK}S_2^2. \qquad \text{(B.22)}
\end{aligned}$$

Donde

$$S_1^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_{ki}^2 - N\overline{Y}_k^2\right) \qquad \text{y} \qquad S_2^2 = \frac{1}{K-1}\left(\sum_{r=1}^{K}\bar{y}_r^2 - K\overline{\overline{Y}}_k^2\right)$$
$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(y_{ki}-\overline{Y}_k\right)^2 \qquad = \frac{1}{K-1}\sum_{r=1}^{K}\left(\bar{y}_r-\overline{\overline{Y}}_k\right)^2.$$

Luego, como $\overline{\overline{Y}}_k = \overline{Y}_k$ entonces

$$S_2^2 = \frac{1}{K-1}\sum_{r=1}^{K}\left(\bar{y}_r - \overline{Y}_k\right)^2.$$

□

S_1^2 es la varianza entre los votos y S_2^2 representa a la varianza de votos entre $AREs$. Por lo tanto, si la población consta de un solo ARE entonces el segundo componente de $(B.22)$ no debe considerarse ya que se requiere de al menos dos $AREs$ para poder tomar la varianza entre ellos. Por otra parte, es importante resaltar que para garantizar la no negatividad de la varianza $(B.22)$ se debe elegir un tamaño de muestra n que satisfaga la siguiente restricción

$$\frac{K(N-1)}{N}\frac{S_1^2}{S_2^2} > n-1 \tag{B.23}$$

Esto garantiza que la varianza calculada sea siempre positiva.

RESULTADOS POR SIMULACIÓN

En esta sección se aportan algunos ejemplos de resultados vía simulación que se realizaron para comprobar que la varianza $(B.22)$ siempre es positiva dada la condición $(B.23)$ para diferentes tamaños de N, K, C y n. Para ello, se realizaron $10,000$ simulaciones en cada caso, generando números aleatorios y_i en diferentes rangos de valor.

Para valores aleatorios $y_i \in (1, 10000)$ y valores mostrados en la siguiente tabla:

N	100	1000	10000	100000	500000
K	50	250	2000	12500	250000
C	2	4	5	8	2
n	40	200	1000	7000	100000

En todos estos casos se satisface la condición $(B.23)$, consecuentemente la varianza $(B.22)$ siempre fue positiva en cada una de las $10,000$ ejecuciones. Este mismo comportamiento se observó al cambiar el rango de valores por $y_i \in (0, 1)$. Se consideraron muchos más casos, se trató de contemplar cada caso posible, sin embargo por simplicidad y no extender la tabla anterior, solo se reportan algunos de ellos.

ESTIMADOR ANALÍTICO

Nos gustaría estimar a $V(\bar{y}_k)$ con la siguiente expresión

$$v(\bar{y}_k) = \frac{N-1}{nN} s_1^2 - \frac{n-1}{nK} s_2^2, \tag{B.24}$$

con s_1^2 y s_2^2 estimadores muestrales para S_1^2 y S_2^2, respectivamente:

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ki} - \bar{y}_k)^2 \qquad \text{y} \qquad s_2^2 = \frac{1}{n-1} \sum_{r=1}^{n} \left(\hat{\bar{y}}_r - \bar{y}_k\right)^2$$

en estas expresiones $\bar{y}$ es el estimador muestral para $\overline{Y}$. Sin embargo, dado que en cada ARE se selecciona un único elemento entonces $\hat{\bar{y}}_r = y_i$ con i el indicador de la casilla seleccionada en el ARE r. De esta forma s_2^2 puede reescribirse y ser exactamente igual a s_1^2.

$$s_2^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ki} - \bar{y}_k)^2 .$$

Inconvenientemente esto significa que s_2^2 no es un estimador insesgado para S_2^2. Además, se puede probar que $\mathbb{E}[s_1^2] \neq S_1^2$.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} (y_{ki} - \bar{y}_k)^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \{(y_{ki} - \overline{Y}_k) - (\bar{y}_k - \overline{Y}_k)\}^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} (y_{ki} - \overline{Y}_k)^2 - n(\bar{y}_k - \overline{Y}_k)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N} Z_i (y_{ki} - \overline{Y}_k)^2\right] - nV(\bar{y}_k) \\
&= \frac{n}{N} \sum_{i=1}^{N} (y_{ki} - \overline{Y}_k)^2 - \frac{N-1}{N} S_1^2 + \frac{(n-1)C}{N} S_2^2
\end{aligned}$$

$$= \frac{n(N-1)}{N} S_1^2 - \frac{N-1}{N} S_1^2 + \frac{n-1}{K} S_2^2$$

$$= \frac{(N-1)(n-1)}{N} S_1^2 + \frac{n-1}{K} S_2^2$$

Así,

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_{ki}-\bar{y}_k)^2\right] = \mathbb{E}\left[s_1^2\right] = \frac{(N-1)}{N} S_1^2 + \frac{1}{K} S_2^2. \tag{B.25}$$

Para corregir el sesgo en s_1^2 se requiere conocer la varianza real entre los $AREs$ o construir un estimador insesgado para este. Sin embargo, la varianza real no puede conocerse hasta después de las elecciones, y debido al tipo de selección en cada ARE tampoco es posible construir un estimador analítico insesgado para S_2^2. Por lo tanto, $v(\bar{y}_k)$ es un estimador sesgado de $V(\bar{y}_k)$.

ESTIMADOR NUMÉRICO

Un mejor estimador para $V(\bar{y}_k)$, que se obtiene modificando los estimadores s_1^2 y s_2^2, es

$$\tilde{v}(\bar{y}_k) = \frac{N-1}{nN} s_{1*}^2 - \frac{n-1}{nK} s_{2*}^2. \tag{B.26}$$

Para calcular s_{1*}^2 se realiza un proceso de remuestreo sobre la muestra de casillas elegidas mediante las dos etapas, es decir:

1. Se toma una muestra con remplazo de tamaño n de la muestra original.
2. Se calcula la varianza (s_1^2) de esta muestra.
3. Se repite m veces el paso 1 y 2.
4. Finalmente s_{1*}^2 será la media de estas varianzas.

Para s_{2*}^2, se sigue el siguiente procedimiento para reagrupar las n casillas de la muestra:

1. Se ordenan las casillas de la muestra original, $y_1, y_2, y_3, ..., y_n$.
2. Se re-calcula la media estimada de votos para el i-ésimo ARE como:

$$\hat{\bar{y}}_i^* = \frac{y_i + y_{i+1}}{2}, \qquad i = 1, 2, 3, ..., n-1$$

3. Se toma la varianza de las $n-1$ nuevas medias como s^2_{2*}.

4. Si $n = 2$, entonces s^2_{2*} debe igualarse a cero.

OBSERVACIONES

1. Para poder estimar la varianza, el tamaño de muestra debe ser de al menos 2 casillas. Esto es, en la primera etapa se deben seleccionar por lo menos 2 ARE's de los K disponibles.

2. Si la muestra consta de un solo ARE, la varianza debe igualarse a cero u omitirse ya que su contribución a la varianza total no es significativa.

C. BASE DE DATOS

Las bases de datos empleadas corresponden a los cómputos distritales (CD) de las elecciones locales de 2016 y 2017, mismas que fueron empleadas por el Comité Técnico para el Conteo Rápido de 2018 (COTECORA), paras las estimaciones de ese año. Para la elección presidencial se trabajó con la base de datos más reciente, 2018. Todas estas bases fueron proporcionadas por el Dr. Carlos Erwin Rodríguez Hernández-Vela, quien formó parte del COTECORA 2018. Las bases de datos nos dan información de la votación a nivel casilla, en varios casos pueden obtenerse por Internet y son los Cómputos Distritales.

- Elecciones locales 2012: CDMX, GUANAJUATO, JALISCO, MORELOS, TABASCO, CHIAPAS y YUCATÁN
- Elecciones locales 2016: PUEBLA y VERACRUZ
- ELECCIÓN PRESIDENCIAL 2018

La estructura de estas bases de datos es muy similar, las primeras columnas refieren a datos generales sobre la distribución de las casillas. Por ejemplo,

| ID_ESTADO | NOMBRE_ESTADO | DIST_LOC | CABECERA_DISTRITAL | ID_MUNICIPIO | MUNICIPIO | SECCION | CASILLA |

Seguido de estos datos, continuarían todos los partidos que participan en la elección, una columna por cada partido. Si en la elección existen coaliciones, entonces todas las posibles combinaciones de estas aparecen en la base, cada una en columnas diferentes. Estas columnas contienen las votaciones obtenidas en la elección correspondiente, también se disponen de columnas para las las votaciones nulas y candidatos no registrados. Un ejemplo de esto se puede ver en el Apéndice dedicado a los diccionarios para los cómputos distritales.

Particularmente, las bases de datos pasaron por varios filtros antes de poder emplearse, filtros realizados por el COTECORA. Por lo tanto, no se tuvieron que hacer muchas adecuaciones para disponer de una base de datos final. La modificación más relevante fue agrupar en una sola columna los votos para las coaliciones e identificarlas con un solo nombre (partido de mayor antigüedad), también hubo que agrupar las votaciones nulas y los votos para candidatos no registrados en una sola columna etiquetada como OTROS. Por último anexar una nueva columna correspondiente a la nueva estratificación. No obstante, en los inicios

de la tesis solo se disponían de algunas bases de datos, por lo que hubo que corroborar la información obtenida de los Cómputos Distritales, descargados de las páginas oficiales de cada estado, por ejemplo Colima, con los resultados de bases de datos de CAEs para la misma elección. El proceso se describe a continuación.

UNIÓN DE BASES DE DATOS (CD y CAE)

Para obtener el resultado a nivel candidato, se realizaron los siguientes ajustes en ambas bases de datos:

- Los datos de identificación de las casillas permanecen igual.
- Se sumaron las votaciones de la coalición PRI.VERDE.NA.ES como un solo candidato, eliminando así las columnas correspondientes a las votaciones individuales de cada candidato.

El siguiente paso fue construir la base de datos para el análisis desarrollado en este trabajo. Para ello se pegó la información de los CAE's a la del los Cómputos Distritales. En algunos casos no se contaba con una clave única de CAE por lo que esta se construyó concatenando DISTRITO FEDERAL Y ARE. Después fue necesario crear una identificador único de casilla. Esta se consigue concatenando las siguientes variables:

VARIABLE	CARACTERES
ID_ESTADO	2
DISTRITO_LOC	2
SECCION	4
CASILLA	6

Cuando en la base de datos la variable tiene menos caracteres, entonces se agregan ceros a la izquierda para completar hasta obtener el número de caracteres indicado.

Por ejemplo, para la siguiente casilla

ID_ESTADO	NOMBRE_ESTADO	DIST_LOC	CABECERA_DISTRITAL	ID_MUNICIPIO	MUNICIPIO	SECCION	CASILLA
15	MEXICO	11	TULTITLAN DE MARIANO ESCOBEDO	110	TULTITLAN	5491	C02

El identificador único es: 15115491C02000. En este identificador se ha completado con 3 ceros la variable CASILLA para obtener los caracteres descritos en la tabla anterior.

Este identificador se construye en ambas bases de datos anexando una columna con este identificador. Posteriormente se hace la unión de ambas bases, tomando como referencia esta nueva columna. La información combinada será la base de datos final, BDE_FINAL.

D. GLOSARIO ELECTORAL

Para contabilizar los votos de las elecciones para Gobernadores o Diputados el INE emplea tres distintos mecanismos: el Programa de Resultados Electorales Preliminares (PREP), el Conteo Rápido (CR) y los Cómputos Distritales (CD).

PREP

El PREP es el mecanismo de información electoral encargado de proveer los resultados preliminares y no definitivos, de carácter estrictamente informativo a través de la captura, digitalización y publicación de los datos asentados en las Actas de Escrutinio y Cómputo de las casillas que se reciben en los Centros de Acopio y Transmisión de Datos autorizados por el Instituto o por los Organismos Públicos Locales.

CR

El CR es un procedimiento estadístico diseñado con la finalidad de estimar con oportunidad las tendencias de los resultados finales de una elección [Link]. Está basado en las actas de escrutinio y cómputo de casilla a fin de conocer las tendencias de los resultados de la jornada electoral[Constitución Política].

CD

Es la suma que realiza el Consejo Distrital de los resultados anotados en las actas de escrutinio y cómputo de las casillas en un distrito electoral [Link].

El Consejo Distrital deberá realizar nuevamente el escrutinio y cómputo cuando[Constitución Política]:

- Existan errores o inconsistencias evidentes en los distintos elementos de las actas, salvo que puedan corregirse o aclararse con otros elementos a satisfacción plena de quien lo haya solicitado.
- El número de votos nulos sea mayor a la diferencia entre los candidatos ubicados en el primero y segundo lugares en votación
- Todos los votos hayan sido depositados a favor de un mismo partido.

E. DICCIONARIO PARA LOS CÓMPUTOS DISTRITALES

Como se ha mencionado antes, las bases de datos con las que se reporta la información de las elecciones tienen una estructura similar. Por lo tanto, para dar una noción de su estructura, consideremos el diccionario electoral para la elección a gobernador del estado de Colima 2016.

DICCIONARIO COLIMA

ESTADO: Número de la Entidad Federativa.

NOMBRE_ESTADO: Nombre de la Entidad Federativa.

DISTRITO: Número del distrito electoral de la entidad.

CABECERA_DISTRITAL: Nombre de la Cabecera Distrital.

SECCION: Número de sección correspondiente a la casilla.

ID_CASILLA: Número identificador de la casilla.

TIPO_CASILLA: Tipo de casilla.

EXT_CONTIGUA: Número de casilla contigua a una extraordinaria.

UBICACION_CASILLA: Identifica el tipo de casilla Urbana o No urbana.

TIPO_ACTA: Especifica el tipo de Acta, tomando alguno de los siguientes valores:
GOB: Acta de casilla para Gobernador.
GOBS: Acta de casilla para Gobernador especial.

PAN: Número de votos para el Partido Acción Nacional.

PRI: Número de votos para el Partido Revolucionario Institucional.

PRD: Número de votos para el Partido de la Revolución Democrática.

PVEM: Número de votos para el Partido Verde Ecologista de México.

PT: Número de votos para el Partido del Trabajo.

MC: Número de votos para el Movimiento Ciudadano.

NUEVA_ALIANZA: Número de votos para Nueva Alianza.

MORENA: Número de votos para Morena.

ES: Número de votos para Encuentro Social.

PRI-PVEM-PT-PANAL: Número de votos para la coalición PRI-PVEM-PT-PANAL.

PRI-PVEM-PT: Número de votos para la coalición PRI-PVEM-PT.

PRI-PVEM-PANAL: Número de votos para la coalición PRI-PVEM-PANAL.

PRI-PT-PANAL: Número de votos para la coalición PRI-PT-PANAL.

PVEM-PT-PANAL: Número de votos para la coalición PVEM-PT-PANAL.

PRI-PVEM: Número de votos para la coalición PRI-PVEM.

PRI-PT: Número de votos para la coalición PRI-PT.

PRI-PANAL: Número de votos para la coalición PRI-PANAL.

PVEM-PT: Número de votos para la coalición PVEM-PT.

PVEM-PANAL: Número de votos para la coalición PVEM-PANAL.

PT-PANAL: Número de votos para la coalición PT-PANAL.

NO_REGISTRADOS: Número de votos para candidatos no registrados.

NULOS: Número de votos nulos.

TOTAL_VOTOS: Suma total de votos para la casilla (votos por partido, coalición, candidatos independientes, candidatos no registrados y nulos). La suma se realiza automáticamente por el Sistema de Cómputos Distritales con el fin de evitar errores de registro o aritméticos en las casillas.

LISTA_NOMINAL: Número de votantes posibles de acuerdo a la lista nominal, no se incluyen a los representantes de partidos polÌticos/candidatos y/o candidatos independientes, y/o a los ciudadanos que se encuentran en secciones con menos de 100 electores.

ESTATUS_ACTA: Indica el estado del acta, tomando alguno de los siguientes valores: Recuento (SRA). Recuento (Cómputos). Acta del consejo. Paquete no entregado. Casilla no instalada. Casilla con suspensión definitiva de la votación.

BASE DE CAE's

Las variables contenidas en esta base de datos son las siguientes:

ID_ESTADO: Número de la Entidad Federativa.

ID_DISTRITO: Número del distrito electoral de la entidad.

SECCION: Número de sección correspondiente a la casilla.

ID_CASILLA: Número identificador de la casilla.

TIPO_CASILLA: Tipo de casilla.
B.- Básica
C.- Contigua
E.- Extraordinaria
S.- Especial

EXT_CONTIGUA: Número de casilla contigua a una extraordinaria.

ID_MUNICIPIO: Número del distrito o municipio electoral de la entidad.

NUMERO_AREA_RESPONSABILIDAD: Número del área de responsabilidad del CAE.

ID_DISTRITO_LOCAL: Número del distrito electoral de la entidad local.

ESTRATO: Estrato al que pertenece la casilla.

BIBLIOGRAFÍA

[Cochran] William Cochran. *Sampling Techniques.* Copyright 1977, by John Wiley and Sons, Inc. Canada.

[Stanley] Paul S. Levy and Stanley Lemeshow. *Sampling of Populations Methods and Applications.* Fourth Edition 2008.

[Lohr] Sharon L. Lohr. *Sampling: Design and Analysis.* Second Edition 2010.

[Hani] Hani M. Samawi – Mahmoud I. Siam *Ratio estimation using stratified ranked set sample* 2001

[Ross] Sheldon M. Ross. *Introductory Statistics.* Third Edition 2010

[Brown] Lawrence D. Brown, T. Tony Cai and Anirban DasGupta. *Interval Estimation for a Binomial Proportion* Statistical Science 2001, Vol. 16, No. 2, 101-133

[Constitución Política] Constitución Política de los Estados Unidos Mexicanos. Tomo II.

[Código Electoral] Código Electoral del Estado de México. Primera edición 2017.

[Alonso y Coria] Luis Carlos Ugalde y Said Hernández. Fortalezas y debilidades del sistema electoral mexicano. Perspectiva federal y local. Capítulo 23, pag. 488. Por Alberto Alonso y Coria.

[Reglamento de Elecciones 2016] Reglamento de Elecciones 2016 Aprobado en Sesión Extraordinaria del Consejo General, celebrada el 07 de septiembre de 2016, mediante Acuerdo ine/cg661/2016.

[Lawrence D. Brown] Lawrence D. Brown, T. Tony Cai and Anirban DasGupta Interval Estimation for a Binomial Proportion. Statistical Science 2001, Vol. 16, No. 2, 101-133.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00189
Matrícula: 2163803461

ESQUEMAS DE MUESTREO PARA EL CONTEO RAPIDO BAJO RESTRICCIONES OPERATIVAS.

JERONIMO HERNANDEZ MENDOZA
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

En la Ciudad de México, se presentaron a las 15:30 horas del día 18 del mes de octubre del año 2019 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. ALBERTO CASTILLO MORALES
M.C. PATRICIA ROMERO MARES
DR. CARLOS ERWIN RODRIGUEZ HERNANDEZ VELA
DR. GABRIEL NUÑEZ ANTONIO

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (MATEMÁTICAS APLICADAS E INDUSTRIALES)

DE: JERONIMO HERNANDEZ MENDOZA

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

APROBAR

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTE

DR. ALBERTO CASTILLO MORALES

VOCAL

M.C. PATRICIA ROMERO MARES

VOCAL

DR. CARLOS ERWIN RODRIGUEZ HERNANDEZ VELA

SECRETARIO

DR. GABRIEL NUÑEZ ANTONIO