



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y
DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

COMPARACIÓN DE ESTRATEGIAS PARA HACER INFERENCIA BAJO
ESQUEMAS DE MUESTREO PROBABILÍSTICO

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:
JOSÉ ALBERTO FLORENCIO CHÁVEZ

DIRECTOR:
CARLOS ERWIN RODRÍGUEZ HERNÁNDEZ-VELA
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS (IIMAS)

CIUDAD DE MÉXICO, A 8 DE ENERO DEL 2024.



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Con mucho amor a Lee.

Agradecimientos

Agradezco especialmente al Dr. Carlos Rodríguez por el tiempo y la entrega destinados en el desarrollo de la presente tesis. Adicionalmente agradezco a los sinodales que apoyaron con la revisión de la misma, la Dra. Ruth Fuentes, la Dra. Lizbeth Naranjo, el Dr. Raúl Rueda y la Dra. Silvia Ruiz-Velasco.

Investigación realizada gracias al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM TA100623. Agradezco a la DGAPA-UNAM la beca recibida

Resumen

El propósito central de este trabajo de tesis consiste, en primer lugar, en explorar diversas alternativas para llevar a cabo inferencias en poblaciones finitas, y en segundo lugar, en identificar la opción más efectiva. En este contexto, la noción de “mejor” se refiere a una estrategia que demuestre una mayor robustez frente a valores atípicos. Con este objetivo en mente, se presentan inicialmente las estrategias frecuentista y Bayesiana para la inferencia en poblaciones finitas. Posteriormente, se aborda una propuesta reciente fundamentada en un enfoque predictivo que incorpora la imputación reforzada por martingalas. Finalmente, se lleva a cabo una comparación exhaustiva de las estrategias descritas con el fin de determinar cuál de ellas se erige como la más sólida. La motivación subyacente se origina en el marco del Conteo Rápido coordinado por el Instituto Nacional Electoral (INE), con especial atención a los acontecimientos relacionados con el Conteo Rápido de la Consulta Popular de 2021.

La Consulta Popular de 2021 tenía como objetivo que los ciudadanos expresaran su posición respecto a la viabilidad de emprender investigaciones penales contra ex-presidentes mexicanos por presunta corrupción durante sus respectivos mandatos. Esta iniciativa de participación ciudadana fue promovida por el presidente Andrés Manuel López Obrador desde antes de asumir el cargo. Sin embargo, la realización de la consulta enfrentó diversas críticas y desafíos. Entre ellos, destacan: 1) recortes presupuestarios al Instituto Nacional Electoral (INE), lo que limitó la instalación de casillas a la mitad de las que se hubieran dispuesto para una elección nacional; 2) la realización de la consulta en medio de la tercera ola de la pandemia por COVID-19, generando desafíos logísticos y de seguridad considerables; 3) inquietudes relacionadas con la formulación de la pregunta en la boleta, elaborada por la Suprema Corte de Justicia de la Nación, respecto a su claridad e interpretación por parte de los ciudadanos; y 4) la necesidad de una participación ciudadana mínima del 40% (de un total aproximado de 93 millones de ciudadanos en la lista nominal) para que los resultados fueran vinculantes. En un análisis exploratorio de la base de datos de la Consulta Popular 2021 se encontraron valores atípicos de los votos obtenidos para la opción NULOS, correspondiendo a la diferencia de la lista nominal de esa casilla y los votos obtenidos para las otras opciones, implicando valores atípicos del 100% para la participación ciudadana en un ejercicio con apenas un 7.11% de participación ciudadana, además la baja participación maximiza los efectos de la heterogeneidad de los estratos, ya que hay muchos registros con 0 votos para la opción SI y otros tantos más en los que dicha opción obtenía casi la totalidad de la lista nominal en votos.

El Conteo Rápido constituye un ejercicio estadístico que, mediante el uso de información de la votación recolectada a través de una muestra probabilística de casillas tiene como objetivo estimar las tendencias de una elección específica, recientemente aplicado también en consultas populares, y comunicar estos resultados a la población en la misma noche de la elección. Aunque el Conteo Rápido de 2021 logró estimar de manera precisa la participación en la Consulta Popular, aproximadamente un 7%, y, por ende, determinó que no sería vinculante, un análisis subsiguiente, considerando la base de datos con la totalidad de los votos y las estimaciones para cada opción de respuesta, reveló que la presencia de valores atípicos, la baja participación y la heterogeneidad de los estratos generó una disminución en la cobertura de los intervalos de confianza, que originalmente se diseñaron con un nivel del 95%. Esta observación motivó la búsqueda de estrategias de estimación más robustas frente a este tipo de desafíos.

La estructura de este trabajo es la siguiente: En el Capítulo 1, iniciamos con una breve contextualización del Conteo Rápido y presentamos los resultados obtenidos en la Consulta Popular de 2021. Además, exploramos las tres perspectivas de inferencia bajo poblaciones finitas que serán consideradas en el estudio. Esta introducción establece el escenario para la estructura del trabajo. En el Capítulo 2, se expone la teoría básica del muestreo probabilístico para la estimación frecuentista bajo poblaciones finitas, presentando el estimador de razón combinado y su aplicación al Conteo Rápido como nuestra primera alternativa de comparación. En el Capítulo 3, abordamos la perspectiva clásica de la inferencia Bayesiana bajo poblaciones finitas, introduciendo dos modelos paramétricos simples. En el Capítulo 4, presentamos las ideas de Fong et al. (2021) sobre su enfoque de inferencia predictiva Bayesiana, junto con las propuestas de Holmes and Walker (2023) de un algoritmo predictivo. El Capítulo 5 proporciona la teoría necesaria sobre modelos de mezclas de distribuciones y el algoritmo Esperanza-Maximización. Conjuntamente con lo descrito en el Capítulo 4, proponemos dos modelos de mezclas de distribuciones univariadas aplicables al Conteo Rápido, completando así las tres perspectivas de inferencia bajo poblaciones finitas que se compararán con muestras de la base de datos reales de la Consulta Popular de 2021. Finalmente, en el Capítulo 6, llevamos a cabo la comparación detallada de estas perspectivas utilizando datos reales, ofreciendo una evaluación crítica de las estrategias propuestas.

Índice general

Agradecimientos	V
Resumen	VII
Motivación	XV
1. Introducción	1
1.1. El Conteo Rápido	1
1.1.1. Objetivo	2
1.2. La Consulta Popular de 2021	2
1.2.1. El Conteo Rápido para la Consulta Popular de 2021	3
1.2.2. Desempeño del estimador de razón combinado	6
1.3. Estimación bajo poblaciones finitas desde varios enfoques	8
1.4. Organización de la tesis	8
2. Teoría clásica del muestreo probabilístico	11
2.1. Introducción	11
2.2. Muestreo aleatorio simple (MAS)	12
2.2.1. Muestreo aleatorio simple sin reemplazo (MASSR)	13
2.2.2. Estimadores	13
2.2.3. Estimador de razón	16
2.2.4. Tamaño de la muestra	17
2.3. Muestreo aleatorio estratificado (MAE)	18
2.3.1. Estimadores	18
2.3.2. Estimador de razón combinado	19
2.4. Aplicación al Conteo Rápido	20
2.4.1. Notación	20
2.4.2. Estimadores y varianza estimada del estimador	21
2.5. Trabajo futuro: estimación usando información auxiliar	22
2.5.1. Propuestas a comparar	23
3. Inferencia Bayesiana bajo poblaciones finitas	25
3.1. Ideas básicas de la inferencia Bayesiana	25
3.2. Intercambiabilidad y teorema de representación	26
3.3. Consideraciones especiales bajo poblaciones finitas	27

3.3.1.	Generalidades	27
3.3.2.	Verosimilitud-completa y verosimilitud-observada	27
3.3.3.	Superpoblación y población finita	28
3.3.4.	Ignorabilidad	28
3.3.5.	Muestreo aleatorio simple: Diseño ignorable y conocido	28
3.3.6.	Muestreo aleatorio estratificado	29
3.4.	Aplicación al Conteo Rápido	29
3.4.1.	Modelo normal univariado	30
3.4.2.	Modelo Poisson	31
3.4.3.	Muestreo aleatorio estratificado (MAE)	32
4.	Inferencia predictiva	35
4.1.	Motivación y panorama general	35
4.2.	Enfoque predictivo bajo poblaciones finitas	36
4.3.	Metodología	37
4.3.1.	Modelo normal univariado	37
4.4.	Inferencia predictiva bajo MASSR y MAE	39
5.	Inferencia predictiva usando modelos de mezclas	41
5.1.	Modelos de mezclas de distribuciones	41
5.2.	Modelo de mezclas de normales univariadas	42
5.3.	Modelo de mezclas Poisson	44
5.4.	Algoritmo Esperanza-Maximización (EM)	45
5.4.1.	Teoría general	46
5.4.2.	Algoritmo EM para mezclas	47
5.4.3.	Mezclas de normales univariadas	49
5.4.4.	Mezclas de distribuciones Poisson	49
5.4.5.	Consideraciones adicionales	50
6.	Comparaciones	53
6.1.	Estrategias a comparar	53
6.2.	Evaluación de las estrategias de estimación vía simulación	55
6.3.	Consulta Popular 2021	56
6.3.1.	Resultados: MASSR	56
6.3.2.	Resultados: MAE	58
6.4.	Elección Presidencial de 2018	59
6.4.1.	Resultados: MASSR	59
6.4.2.	Resultados: MAE	61
7.	Conclusiones	63
	Bibliografía	65

Índice de figuras

1.1. Intervalos del estimador de Razón Combinado para la opción SÍ. . . .	6
1.2. Intervalos del estimador de Razón Combinado para la opción NO. . .	7
1.3. Intervalos del estimador de Razón Combinado para la opción NULOS.	7

Índice de tablas

1.1. Comparación de resultados entre el Conteo Rápido y los cálculos finales, Consulta Popular 2021.	4
1.2. Resumen descriptivo de la distribución para los votos obtenidos por las opciones y el total de votos por casilla, Consulta Popular 2021 base completa.	4
1.3. Comparación de resultados entre el Conteo Rápido y los cálculos finales, Elección Presidencial de 2018.	5
1.4. Resumen descriptivo de la distribución para los votos obtenidos por los candidatos y el total de votos por casilla, Elección Presidencial de 2018.	5
6.1. PRC para la Consulta Popular calculado con 200 muestras seleccionadas vía MASSR.	57
6.2. EPM para la Consulta Popular calculado con 200 muestras seleccionadas vía MASSR.	57
6.3. PRC para la Consulta Popular calculado con 200 muestras seleccionadas vía MAE.	58
6.4. EPM para la Consulta Popular calculado con 200 muestras seleccionadas vía MAE.	59
6.5. PRC para la Elección Presidencial de 2018 calculado con 200 muestras seleccionadas vía MASSR.	60
6.6. EPM para las Elecciones Federales de 2018 calculado con 200 muestras seleccionadas vía MASSR.	61
6.7. PRC para la Elección Presidencial de 2018 calculado con 200 muestras seleccionadas vía MAE.	61
6.8. EPM para la Elección Presidencial de 2018 calculado con 200 muestras seleccionadas vía MAE.	62

Motivación

La inferencia estadística bajo poblaciones finitas desempeña un papel crucial en una variedad de disciplinas, que abarcan desde la medicina y biología hasta las ciencias sociales y del comportamiento. En este contexto, tanto gobiernos como entidades privadas dependen significativamente de información relacionada con poblaciones finitas, tales como estadísticas geográficas, de salud y económicas, obtenidas a través de censos y encuestas. Estos datos no solo son fundamentales para la toma de decisiones políticas, sino que también desempeñan un papel esencial en procesos de planificación y administración. En la búsqueda de mejorar las herramientas para realizar inferencia en poblaciones finitas, esta tesis se adentrará en el análisis de enfoques clásicos, Bayesianos y explorará nuevas perspectivas basadas en conceptos predictivos basados en martingalas, contribuyendo así al avance y enriquecimiento del campo de la inferencia estadística bajo poblaciones finitas.

Una motivación directa para este trabajo es el Conteo Rápido que organiza el Instituto Nacional Electoral (INE). El objetivo del Conteo Rápido es usar la información de la votación en una muestra probabilística de n casillas electorales, de un total de N , para estimar los porcentajes de votos en favor de cada uno de los candidatos a un puesto de elección popular. Este es el contexto básico de la inferencia sobre poblaciones finitas. En el Conteo Rápido las estimaciones siempre se hacen usando intervalos del 95 % de confianza para los porcentajes poblacionales. En esta caso, la teoría clásica del muestreo probabilístico se usa para 1) diseñar la estrategia de selección de casillas, 2) determinar el número de casillas a seleccionar, n , para alcanzar una cierta precisión con el 95 % de confianza y 3) obtener expresiones analíticas para calcular las estimaciones por intervalo siempre bajo el supuesto de normalidad asintótica del estimador. Contar con expresiones cerradas para el estimador puntual, así como una aproximación a la varianza hace muy sencillo, rápido y fácil obtener estimaciones por intervalo, vía el muestreo probabilístico. Cuando el supuesto de normalidad asintótica se cumple y no existen observaciones atípicas, este es el camino más recomendable.

El problema es que la distribución de muestreo del estimador, no siempre sigue una distribución normal, sino que puede estar sesgada, tener colas pesadas o estar contaminada. Esto sucede en casos en los que el tamaño de muestra es “pequeño” y/o debido a la presencia de valores atípicos. Desafortunadamente, incluso cuando una distribución normal proporcione una buena aproximación a la distribución de muestreo del estimador, suelen ocurrir problemas en la práctica. La consecuencia inmediata es que la cobertura real de los intervalos normales asintóticos será mucho menor al 95 % de confianza con el que fueron diseñados. Esto es un problema grave,

tanto metodológico como práctico. Por lo tanto, existe la necesidad de explorar métodos alternativos para obtener intervalos de confianza que alcancen la cobertura con la que fueron diseñados y que no requieran del supuesto de normalidad asintótica. Con este objetivo en mente, en esta tesis se exploran varias estrategias para estimar de manera robusta los porcentajes que pretende estimar el Conteo Rápido.

Buscando una estrategia de estimación robusta, primero se describen las ideas básicas del muestreo probabilístico, a continuación se exploran los conceptos de inferencia Bayesiana para realizar inferencia bajo poblaciones finitas, posteriormente se da una visión de ideas recientes para realizar inferencia predictiva vía martingalas. Armados con estas tres estrategias de estimación, comparamos el desempeño de cada alternativa vía simulación usando datos reales.

Capítulo 1

Introducción

La inferencia estadística bajo poblaciones finitas desempeña un papel crucial en una variedad de disciplinas, que abarcan desde la medicina y biología hasta las ciencias sociales y del comportamiento. En este contexto, tanto gobiernos como entidades privadas dependen significativamente de información relacionada con poblaciones finitas, tales como estadísticas geográficas, de salud y económicas, obtenidas a través de censos y encuestas. Estos datos no solo son fundamentales para la toma de decisiones políticas, sino que también desempeñan un papel esencial en procesos de planificación y administración.

1.1. El Conteo Rápido

La motivación principal de nuestro trabajo se centra en el Conteo Rápido organizado por el Instituto Nacional Electoral (INE). Este mecanismo estadístico tiene como objetivo proporcionar estimaciones oportunas y confiables sobre los resultados finales de un proceso electoral (o incluso de una consulta popular), basándose en la votación en una muestra aleatoria de casillas electorales.

En el contexto específico del Conteo Rápido, la población finita consta de N casillas, cada una con el número de votos a favor de cada uno de los J candidatos a un puesto de elección popular. Así, $y_{i,j}$ para $i = 1, \dots, N$ y $j = 1, \dots, J$ indica el número de votos en la casilla i a favor del candidato j . En este contexto, resulta trivial calcular los porcentajes poblacionales:

$$p_{N,j} = 100 \frac{\sum_{i=1}^N y_{i,j}}{\sum_{j=1}^J \sum_{i=1}^N y_{i,j}}, \quad \text{para } j = 1, \dots, J. \quad (1.1)$$

De las N casillas, se elige una muestra aleatoria de n casillas, y con esta información se diseña una estrategia para estimar los porcentajes poblacionales (1.1) mediante intervalos de confianza del 95 %.

En el año 2014 se promulgó la nueva Ley General de Instituciones y Procedimientos Electorales (LEGIPE), en la cual se definen y describen los programas de resultados preliminares, enfatizando su carácter informativo y no definitivo, específicamente en lo que respecta a los conteos rápidos. Dicha ley facultaba al INE y a los Organismos

Públicos Locales Electorales (OPLE) para ordenar su realización. Posteriormente, en el Reglamento de Elecciones 2016 emitido por el Consejo General del INE, se estableció la obligación de implementar conteos rápidos en elecciones de gobernadores y jefes de gobierno. Una discusión más extensa sobre la historia de los Conteos Rápidos en México se puede encontrar en Hernández Mendoza (2019).

1.1.1. Objetivo

Con el propósito de mejorar las herramientas para realizar inferencia en poblaciones finitas, esta tesis se sumergirá en el análisis de enfoques clásicos y Bayesianos, además de explorar nuevas perspectivas basadas en conceptos predictivos respaldados por martingalas. El objetivo es encontrar una estrategia que permita estimar de manera robusta los porcentajes poblacionales (1.1).

Como se mencionó anteriormente, la motivación subyacente para este trabajo de tesis surge del contexto del Conteo Rápido, con un énfasis significativo en los eventos relacionados con el Conteo Rápido de la Consulta Popular de 2021. Por este motivo, a continuación, proporcionamos una breve descripción de este ejercicio de participación ciudadana.

1.2. La Consulta Popular de 2021

La Consulta Popular de 2021 tenía como objetivo que los ciudadanos expresaran su posición respecto a la viabilidad de emprender investigaciones penales contra ex presidentes mexicanos por presunta corrupción durante sus respectivos mandatos. Sin embargo, la realización de la consulta enfrentó diversas críticas y desafíos. Entre ellos, destacan:

1. Inquietudes relacionadas con la formulación de la pregunta en la boleta, elaborada por la Suprema Corte de Justicia de la Nación (SCJN), respecto a su claridad e interpretación por parte de los ciudadanos. El presidente, Andrés Manuel López Obrador, presentó formalmente el 15 de septiembre de 2020 una petición de consulta popular al Senado de la República sobre el posible enjuiciamiento de los ex-mandatarios. Esta propuesta inicial contenía la pregunta *¿Está de acuerdo o no con que las autoridades competentes, con apego a las leyes y procedimientos aplicables, investiguen y en su caso sancionen la presunta comisión de delitos por parte de los expresidentes Carlos Salinas de Gortari, Ernesto Zedillo Ponce de León, Vicente Fox Quesada, Felipe Calderón Hinojosa y Enrique Peña Nieto antes, durante y después de sus respectivas gestiones?*. La Cámara de Senadores envió la propuesta a la SCJN para determinar su constitucionalidad y después de mucha deliberación y presión por parte del presidente, la SCJN determinó en primera instancia aprobada la constitucionalidad de la consulta, pero la pregunta se reformuló de la siguiente manera *¿Estás de acuerdo o no en que se lleven a cabo las acciones pertinentes, con apego al marco constitucional y legal, para emprender un proceso de esclarecimiento de las decisiones políticas tomadas en los años pasados por los actores políticos encaminado a garantizar la justicia y*

los derechos de las posibles víctimas?. Las distintas fuerzas políticas se manifestaron a favor o en contra de la resolución de la SCJN, pero todas coincidían en que la pregunta era “poco clara”¹.

2. Recortes presupuestarios al INE, lo que limitó la instalación de casillas a la mitad de las que se tenían originalmente planeadas. La Ley Federal de Consultas Populares determina que es el INE el responsable de su organización y desarrollo, así como de la promoción del voto. Por tal motivo el INE solicitó recursos adicionales para la organización de la consulta popular, que consistía en una petición de 1,499, 392,669.67 pesos. No obstante, un mes después, la Cámara de Diputados aprobó un recorte de 870 millones de pesos a su presupuesto. Debido a lo anterior, el INE tomó la decisión de instalar 47 mil casillas menos de las 104 mil que tenían previstas. Al final se instalaron alrededor de 57 mil casillas para la realización de la consulta popular².
3. La necesidad de una participación ciudadana mínima del 40% (de un total aproximado de 93 millones de ciudadanos en la lista nominal) para que los resultados fueran vinculantes. El artículo 35 de la Constitución Política de los Estados Unidos Mexicanos (modificado recientemente) establece que es derecho de la ciudadanía el votar en las consultas populares convocadas por el Congreso de la Unión y que si la participación corresponde al menos al 40% de la lista nominal de electores tendrán un resultado vinculatorio.
4. La realización de la consulta en medio de la tercera ola de la pandemia por COVID-19, generando desafíos logísticos y de seguridad considerables³.

1.2.1. El Conteo Rápido para la Consulta Popular de 2021

Aunque el Conteo Rápido de 2021 logró estimar de manera precisa la participación en la Consulta Popular, el intervalo del 95% arrojado por el Conteo Rápido fue de entre 7.07% y 7.74%, mientras que el valor del cómputo final fue de 7.11%, y, por ende, determinó que no sería vinculante. Las estimaciones para los porcentajes de cada una de las opciones de respuesta, presentados en la Tabla 1.1, indican que se debe hacer un análisis más cuidadoso.

El intervalo estimado para la opción SI se encuentra desplazado tan a la izquierda del valor real arrojado por los cómputos finales que no logra cubrirlo aún teniendo una longitud de casi 5 puntos porcentuales, para la opción NO igualmente se presenta un desplazamiento hacia la izquierda, sin embargo, el intervalo sí logra cubrir al valor real y presenta una longitud razonable, mientras que para los votos nulos el intervalo se desplaza a la derecha del valor real y nuevamente se tiene un intervalo muy amplio de 7 puntos porcentuales que no logra cubrir al valor real.

¹Un breviarío sobre distintas posturas se puede leer en la siguiente nota: Un “galimatías”, la pregunta para consulta popular: especialistas.

²Se puede leer al respecto en la siguiente nota: Afectaría recorte al INE consulta popular del 2021.

³Se puede leer más al respecto en la siguiente nota: COVID-19 impactará el costo de las elecciones.

Opción.	Estimaciones (%)		Computo final (%)
	L. Inf.	L. Sup.	
SI	89.36	96.28	97.72
NO	1.38	1.58	1.55
NULOS	2.19	9.21	0.73

Tabla 1.1: Comparación de resultados entre el Cuento Rápido y los cómputos finales, Consulta Popular 2021.

De la base de datos completa que consta de 57,077 computadas, podemos obtener la Tabla 1.2 en la que se incluyen las medias, las desviaciones estándar, el valor mínimo y valor máximo de los votos obtenidos para cada opción y para el total de votos.

Opción	Media.	Desv. est.	Mín.	Máx.
SI	114.08	83.47	0	1,986
NO	1.80	3.22	0	255
NULOS	0.86	11.33	0	1,887
TOTAL	116.74	85.95	0	1,986

Tabla 1.2: Resumen descriptivo de la distribución para los votos optenidos por las opciones y el total de votos por casilla, Consulta Popular 2021 base completa.

Observando el máximo número de votos por casilla se aprecia la presencia de valores atípicos muy extremos para el número de votos nulos contabilizados en alguna casilla, lo cual incrementa su varianza artificialmente, además de que al igual que la opción NO, los votos registrados se centran en valores cercanos al 0, dificultando que las distribuciones de las estimaciones que se puedan hacer entorno a ellos puedan ser simétricas. Con los valores atípicos mostrados en los votos nulos también podemos notar casillas con un participación inusualmente alta respecto a la participación global, teniendo inclusive casillas con un 100% de participación ciudadana, cuando la media es de 7.19% con una desviación estándar de apenas 5.28 puntos porcentuales. Es importante mencionar que un voto nulo se produce cuando un elector marca la boleta de manera inválida, ya sea seleccionando más opciones de las permitidas o realizando marcas no claras. Por otro lado, una boleta sin usar son aquellas que sobraron en la casilla debido a la ausencia de electores, pero no afectan directamente el resultado. Las estadísticas descritas anteriormente parecieran indicar que hubo una confusión en varias casillas con respecto a esta diferencia y se contabilizaron las boletas sin usar

como votos nulos.

Para tener un punto de referencia y poder apreciar la magnitud de lo sucedido en la Consulta Popular de 2021, en la Tabla 1.3 se presentan las estimaciones del Conteo Rápido para la Elección Presidencial de 2018 junto con los cómputos finales.

Candidato.	Estimaciones (%).		Computo final (%).
	L. Inf.	L. Sup.	
AMLO	53.0	53.8	53.19
RAC	22.1	22.8	22.27
JAMK	15.7	16.3	16.41
JHRC	5.3	5.5	5.23

Tabla 1.3: Comparación de resultados entre el Conteo Rápido y los cómputos finales, Elección Presidencial de 2018.

Nuevamente hay dos intervalos que no cubren al valor real, sin embargo, en este caso los intervalos tienen una longitud de entre 0.2 y 0.7 puntos porcentuales. Al analizar la base de datos con la votación completa, 156,631 casillas, se obtienen los resúmenes descriptivos mostrados en la Tabla 1.4.

Candidato	Media.	Desv. est.	Mín.	Máx.
AMLO	192.26	95.20	0	11,989
RAC	80.51	60.26	0	9,161
JAMK	59.31	38.59	0	1,560
JHRC	18.91	16.27	0	315
TOTAL	361.43	124.17	0	23,842

Tabla 1.4: Resumen descriptivo de la distribución para los votos obtenidos por los candidatos y el total de votos por casilla, Elección Presidencial de 2018.

Primeramente tenemos que la media para cada candidato se encuentra lo suficientemente alejada del 0 para que no afecte a la posible simetría de la distribución de un estimador que asuma normalidad, adicionalmente las desviaciones estándar se encuentran bastante contenidas por los valores mínimos y máximos. Cabe mencionar que aunque la media de la lista nominal es de 569.04 con una desviación estándar de 176.68, los valores tan grandes encontrados en el máximo de votos para los candidatos

punteros corresponden a los registros de una casilla instalada en la CDMX con una lista nominal de 28582 votantes. La media de la participación es de 63.07% con una desviación estándar de 10.98 puntos porcentuales, también se tienen registros con una participación del 100% pero en el contexto de una elección que presentó una participación ciudadana del 63.43% resulta factible. En general, en las elecciones los votos nulos y los votos en favor de candidatos no registrados son despreciables y no tienen ningún impacto en los cómputos finales ni en las estimaciones del Conteo Rápido (se consideran para obtener los porcentajes, pero no se presentan).

1.2.2. Desempeño del estimador de razón combinado

El estimador de razón combinado (descrito a detalle en el Capítulo 2) es una de las estrategias de referencia para obtener estimaciones en los Conteos Rápidos. Para evaluar el desempeño de este estimador, utilizando la base de datos con la votación de las 57,077 mesas instaladas para la Consulta Popular 2021, se realizó la siguiente simulación: Primero, se seleccionaron 200 muestras aleatorias estratificadas (MAE), cada una de 1,200 mesas, considerando a los estratos como los 300 distritos federales electorales en los que se divide al país (con un tamaño de 4 mesas seleccionadas por estrato). Segundo, con cada muestra se calcularon los intervalos del 95% de confianza para las opciones SI, NO y NULOS. Finalmente, se calculó el porcentaje de veces que los intervalos cubrieron a los valores reales (el porcentaje de cobertura real) así como la longitud de los intervalos. Los resultados para el SI se presentan en la Figura 1.1, para la opción NO en la Figura 1.2 y para la opción NULOS en la Figura 1.3, en dichas figuras se incluye el valor real para cada opción como una línea horizontal en color rojo, así como la cobertura real y la longitud media de los intervalos.

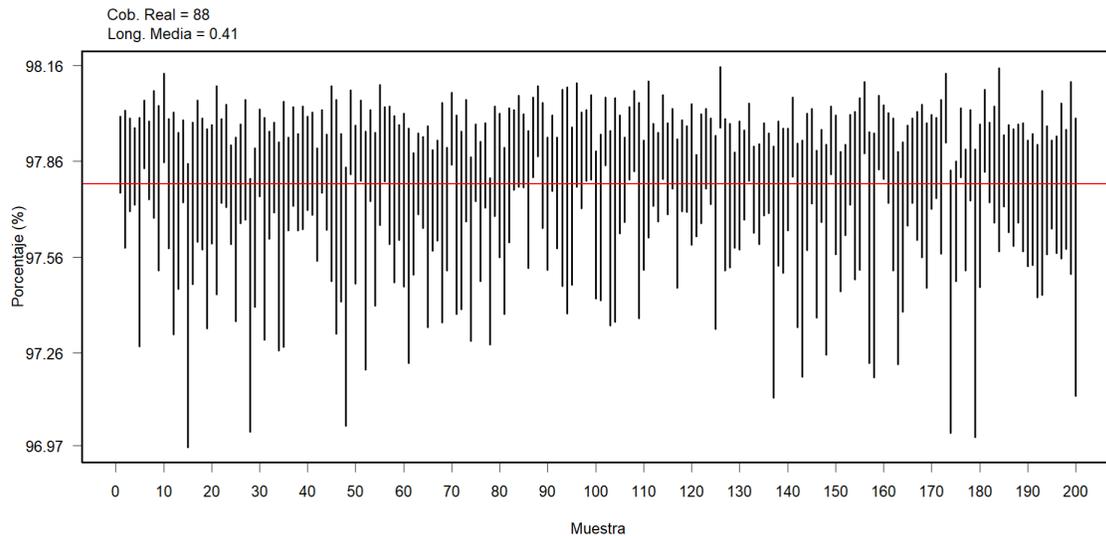


Figura 1.1: Intervalos del estimador de Razón Combinado para la opción SÍ.

Podemos notar que aun con una longitud media razonable de los intervalos, la cobertura real para la opción SI y para la opción NULOS se ve considerablemente reducida respecto a la cobertura teórica con la que fueron diseñados, siendo bastante

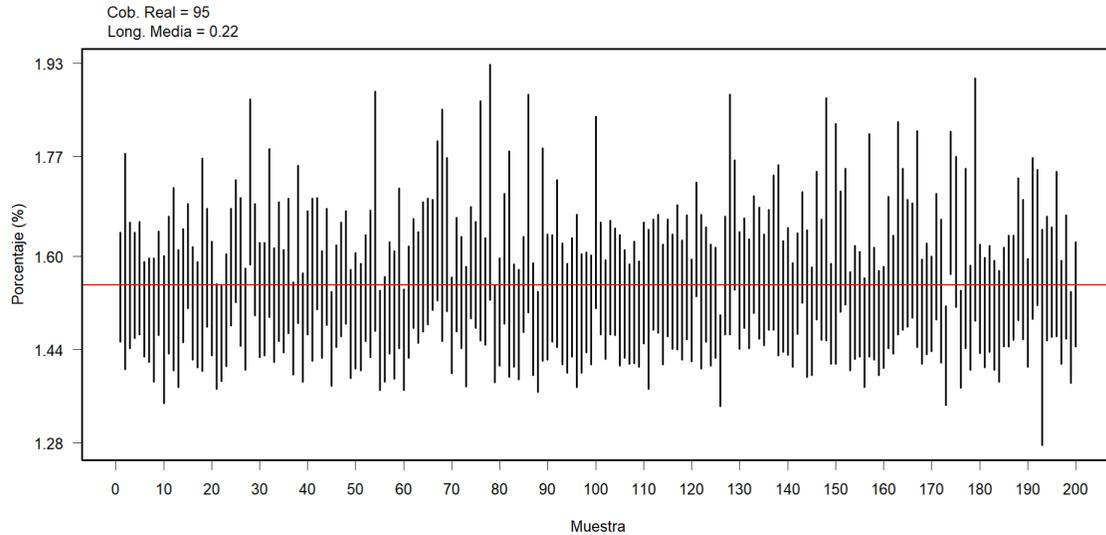


Figura 1.2: Intervalos del estimador de Razón Combinado para la opción NO.

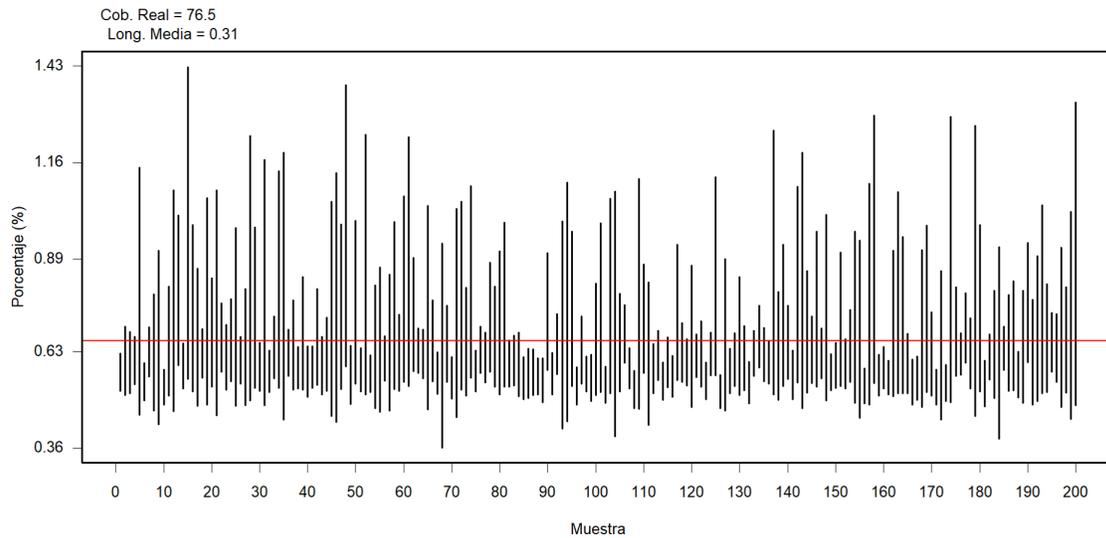


Figura 1.3: Intervalos del estimador de Razón Combinado para la opción NULOS.

más notorio el caso de la opción NULOS ya que la cobertura real obtenida es de apenas el 76.5%, dicha incongruencia podría explicarse por la presencia de valores atípicos ya mencionados así como la proximidad al 0 del valor medio de los votos para esta opción. Para la opción SI tenemos una cobertura real del 88%, en este caso podríamos considerar a la heterogeneidad de los votos para esta opción como una razón para la discrepancias entre las coberturas, ya que se presentaban muchos registros con 0 votos por la baja participación y registros con muchos votos en otras casillas.

Repetiendo el mismo ejercicio de simulación para la base de datos de la Elección Presidencial de 2018 se tiene que los porcentajes de cobertura reales para los tres candidatos punteros son de 94.5% para el candidato AMLO, 95.5% para el candidato

RAC y de 97.5% para el candidato JAMK, obviamente mucho más cercanas a la cobertura teórica del 95%.

Este análisis de lo sucedido en la Consulta Popular 2021 evidencia la necesidad de modelos robustos frente a estructuras y circunstancias complejas en los datos, así como la presencia de valores extremos (outliers) y a la alta heterogeneidad de los estratos en los que se divide la población de casillas y por ende, de la muestra.

1.3. Estimación bajo poblaciones finitas desde varios enfoques

La estadística frecuentista aborda la realización de inferencias en poblaciones finitas considerando los mecanismos de selección de la muestra aleatoria como la fuente de la incertidumbre en las estimaciones. Un análisis detallado de este enfoque se encuentra en diversas referencias bibliográficas, como por ejemplo Cochran (1977) y Lohr (2021). En esta tesis, se presenta de manera concisa la teoría y los resultados relacionados con el muestreo aleatorio simple sin reemplazo y el muestreo aleatorio estratificado. Al tratar con estimaciones sobre el porcentaje de votos para cada fuerza política (o opción de respuesta en el caso de una Consulta Popular), también se describe el estimador de razón combinado, que como ya se ha comentado es el estimador de referencia en el Conteo Rápido. Aún dentro de la perspectiva frecuentista, se presentan ideas para trabajos futuros, centrándose en el uso de información auxiliar para la propuesta de estimadores predictivos de la media de los datos no observados Yadav and Mishra (2015).

La estadística Bayesiana tradicionalmente emplea un modelo paramétrico $p(y|\theta)$ y una distribución prior $p(\theta)$ para el o los parámetros θ . Al realizar inferencia sobre poblaciones finitas, también se debe considerar un modelo para el diseño de muestreo, siendo los diseños ignorables aquellos en los que la variable que se busca estimar no tiene relación con la manera en que se selecciona la muestra. Recientemente, Fong et al. (2021) propuso una perspectiva denominada la *visión predictiva de la inferencia Bayesiana*, que aprovecha resultados sobre la convergencia de martingalas para fundamentar una metodología en la que se parte de una sucesión de distribuciones predictivas a 1-paso para realizar las inferencias buscadas, destacando la apertura de las condiciones sobre la estructura de dependencia de los datos en las cuales esta metodología puede ser aplicable. Posteriormente, Holmes and Walker (2023) retoman estas ideas para formular un algoritmo predictivo, pero en esta ocasión parten de un estimador inicial frecuentista, requiriendo que la actualización de los parámetros esté definida mediante martingalas.

1.4. Organización de la tesis

La estructura de este trabajo es la siguiente: En el Capítulo 2, se expone la teoría básica del muestreo probabilístico para la estimación frecuentista bajo poblaciones finitas, presentando el estimador de razón combinado y su aplicación al Conteo Rá-

vido como nuestra primera alternativa de comparación. En el Capítulo 3, abordamos la perspectiva clásica de la inferencia Bayesiana bajo poblaciones finitas, introduciendo dos modelos paramétricos simples. En el Capítulo 4, presentamos las ideas de Fong et al. (2021) sobre su enfoque de inferencia predictiva Bayesiana, junto con las propuestas de Holmes and Walker (2023) de un algoritmo predictivo. El Capítulo 5 proporciona la teoría necesaria sobre modelos de mezclas de distribuciones y el algoritmo Esperanza-Maximización. Conjuntamente con lo descrito en el Capítulo 4, proponemos dos modelos de mezclas de distribuciones univariadas aplicables al Conteo Rápido, completando así las tres perspectivas de inferencia bajo poblaciones finitas que se compararán usando bases de datos reales. Finalmente, en el Capítulo 6, llevamos a cabo la comparación detallada de estas perspectivas utilizando datos reales (uno de estos conjuntos de datos es la Consulta Popular de 2021), ofreciendo una evaluación crítica de las estrategias propuestas.

Capítulo 2

Teoría clásica del muestreo probabilístico

En este capítulo describiremos los conceptos básicos del muestreo probabilístico, tomando al *muestreo aleatorio simple sin reemplazo* como punto de partida para obtener resultados sobre estimadores de la media, el total y de razón entre la variable de interés y alguna otra. Posteriormente, se expondrán resultados análogos pero ahora bajo el esquema de *muestreo aleatorio estratificado*. Es importante mencionar que el muestreo probabilístico es una técnica estadística desarrollada desde el enfoque frecuentista específicamente para hacer inferencia bajo poblaciones finitas. La referencia clásica para entender los aspectos básicos del muestreo probabilístico es Cochran (1977), para una visión más actual ver por ejemplo Lohr (2021). Un trabajo de tesis de maestría con las expresiones tanto para los estimadores como para la varianza aplicados al Conteo Rápido se puede consultar en Hernández Mendoza (2019).

El objetivo del Conteo Rápido es el de estimar la proporción de votos en favor de todos los candidatos a un puesto de elección popular, básicamente un estimador de razón. Usando la teoría desarrollada en las primeras secciones del capítulo se escribirán las expresiones para hacer inferencia en el Conteo Rápido.

2.1. Introducción

Las técnicas de muestreo probabilístico son aquellas en las que se considera que todo elemento de una población dada tiene una probabilidad conocida y mayor a cero de ser seleccionado en muestra. Partimos de una población finita, cuyos N elementos quedan identificados por el conjunto de índices:

$$\mathcal{U} = \{1, 2, \dots, N\},$$

luego, fijado un tamaño de muestra n se asociará a cada una de los posibles colecciones \mathcal{S} de elementos de \mathcal{U} , de tamaño n , una probabilidad fija y conocida de ser seleccionada $P(\mathcal{S})$, así, si $\mathcal{S}(i) = \{\mathcal{S} | i \text{ está en } \mathcal{S}\}$ tenemos que la probabilidad de que el elemento

i de la población sea seleccionado queda dada por:

$$\pi_i = \sum_{S \in \mathcal{S}(i)} P(S), \quad (2.1)$$

siempre debe garantizarse que $\pi_i > 0$, para todo $i \in \mathcal{U}$.

Adicionalmente, a la medición de la característica de interés en cada uno de los elementos de la población, se le denotará como:

$$y_{1:N} = \{y_1, y_2, \dots, y_N\}.$$

2.2. Muestreo aleatorio simple (MAS)

El muestreo aleatorio simple (MAS) es la opción de muestreo probabilístico más sencilla y básica, que sirve de base para opciones más complejas y en la cual se tiene que la probabilidad de selección es la misma para todo \mathcal{S} . Existen dos alternativas para generar una *muestra aleatoria simple*, la primera es el *muestreo aleatorio simple con reemplazo* (MASCR) y la segunda es el *muestreo aleatorio simple sin reemplazo* (MASSR). En el MASCR seleccionamos n elementos de la población admitiendo reemplazo del elemento seleccionado en el paso anterior, así $\pi_i = \frac{1}{N}$ para toda $i \in \mathcal{U}$ y consecuentemente $P(\mathcal{S}) = \frac{1}{N^n}$ para cualquier muestra \mathcal{S} , lo cual es congruente con las N^n muestras posibles admitiendo reemplazo y considerando el orden (ver por ejemplo p. 14 de Casella and Berger (2002)). El MASCR cumple con la definición de muestra aleatoria vista en los cursos básicos de inferencia estadística en la licenciatura y maestría: variables aleatorias independientes e idénticamente distribuidas. Por otro lado, en el MASSR cada elemento aparece de manera única en cualquier muestra \mathcal{S} , así podemos considerar que para una muestra \mathcal{S} particular de n elementos tenemos que para la primera selección se tiene una probabilidad $\frac{n}{N}$ de seleccionar un elemento de la muestra \mathcal{S} , para la segunda selección la probabilidad es $\frac{n-1}{N-1}$ de seleccionar otro elemento de la muestra, siguiendo este razonamiento se tiene que:

$$P(\mathcal{S}) = \frac{n}{N} \frac{n-1}{N-1} \cdots \frac{1}{N-n+1} = \frac{1}{\binom{N}{n}}, \quad (2.2)$$

para cualquier \mathcal{S} , lo cual es congruente con las $\binom{N}{n}$ muestras posibles de n elementos sin reemplazo en las cuales el orden no es relevante, p. 14 de Casella and Berger (2002), consecuentemente tenemos que:

$$\pi_i = \sum_{S \in \mathcal{S}(i)} P(S) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \quad (2.3)$$

para toda $i \in \mathcal{U}$.

En este trabajo nos concentraremos en el MASSR ya que para la estimación en el Conteo Rápido no tiene sentido tener casillas repetidas en muestra. Finalmente, es importante mencionar que el MASSR no cumple con la definición de muestra aleatoria pues cada selección afecta las probabilidades de selección de los elementos restantes. En otras palabras no se cumple con la independencia.

2.2.1. Muestreo aleatorio simple sin reemplazo (MASSR)

Consideremos la siguiente variable aleatoria binaria que indica si el i -ésimo elemento de la población está o no en la muestra:

$$I_i = \begin{cases} 1 & \text{si } i \in \mathcal{S}, \\ 0 & \text{si } i \notin \mathcal{S}, \end{cases} \quad (2.4)$$

es claro que

$$P(I_i = 1) = \pi_i = \frac{n}{N}, \quad (2.5)$$

así:

$$\mathbb{E}(I_i) = \mathbb{E}(I_i^2) = \frac{n}{N}, \quad \text{para todo } i \in \mathcal{U}, \quad (2.6)$$

por lo tanto:

$$\mathbb{V}(I_i) = \frac{n}{N} \left(1 - \frac{n}{N}\right) \quad \text{para todo } i \in \mathcal{U}, \quad (2.7)$$

además, observemos que:

$$\begin{aligned} \mathbb{E}(I_i I_j) &= P(I_i = 1, I_j = 1) \\ &= P(I_i = 1 | I_j = 1) P(I_j = 1) \\ &= \left(\frac{n-1}{N-1}\right) \left(\frac{n}{N}\right) \quad \text{para } i \neq j, \end{aligned} \quad (2.8)$$

teniendo entonces que las I_i no son independientes ya que mantienen una covarianza no nula:

$$\text{Cov}(I_i, I_j) = -\frac{n}{N(N-1)} \left(1 - \frac{n}{N}\right). \quad (2.9)$$

Es importante mencionar que desde este enfoque las observaciones $y_{1:N}$ se consideran como constantes desconocidas, y la probabilidad es introducida únicamente a través de las variables aleatorias $I_{1:N} = \{I_1, \dots, I_N\}$.

2.2.2. Estimadores

Sabemos que la MASSR está formada por n índices de \mathcal{U} , que denotamos como \mathcal{S} , sin embargo hacer los cálculos con un conjunto de índices arbitrario \mathcal{S} causa muchas complicaciones. Adicionalmente, causará inconsistencias con la notación que se usará en los siguientes capítulos. Entonces, por simplicidad en la notación asumiremos que $\mathcal{S} = \{1, 2, \dots, n\}$ y por lo tanto las mediciones de la variable que fueron seleccionadas en muestra vía el MASSR estarán dadas por $y_{1:n} = \{y_1, \dots, y_n\}$. Esto de ninguna manera quiere decir que las observaciones seleccionadas por MASSR tienen que ser las que corresponden a las primeras n observaciones de \mathcal{U} , sino simplemente es una forma de simplificar la notación. La muestra siempre será un conjunto de índices \mathcal{S} obtenida por MASSR. Finalmente, mencionamos que ésta es la consideración usada por Cochran (1977) (p. 20), mientras que, por ejemplo, Lohr (2021) (p. 51) usa \mathcal{S} directamente.

Con la consideración descrita en el párrafo anterior, los estimadores se pueden escribir como:

	Media	Total	Razón
Estimador	$\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$	$t_n = N\bar{y}_n$	$r_n = \frac{\bar{y}_n}{\bar{x}_n}$

Veamos que tanto \bar{y}_n como t_n son estimadores insesgados de la media poblacional $\bar{y}_N = \frac{1}{N} \sum_{i=1}^N y_i$ y del total poblacional $t_N = N\bar{y}_N$, respectivamente. Primero, notemos que podemos reescribir a \bar{y}_n como una variable aleatoria usando a las variables binarias $I_{1:N}$:

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^N I_i y_i,$$

ya que las y_i son cantidades fijas y sólo las I_i son variables aleatorias, se tiene que:

$$\begin{aligned} \mathbb{E}(\bar{y}_n) &= \frac{1}{n} \sum_{i=1}^N y_i \mathbb{E}(I_i) \\ &= \frac{1}{n} \sum_{i=1}^N y_i \frac{n}{N} \\ &= \frac{1}{N} \sum_{i=1}^N y_i \\ &= \bar{y}_N, \end{aligned} \tag{2.10}$$

con lo anterior se tiene que en efecto \bar{y}_n es un estimador insesgado de \bar{y}_N , además, trivialmente se tiene que t_n es un estimador insesgado de t_N . Ahora, para la varianza de \bar{y}_n tenemos:

$$\begin{aligned} \mathbb{V}(\bar{y}_n) &= \frac{1}{n^2} \mathbb{V} \left(\sum_{i=1}^N y_i I_i \right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^N y_i^2 \mathbb{V}(I_i) + \sum_{i=1}^N \sum_{j \neq i}^N y_i y_j \text{Cov}(I_i, I_j) \right) \\ &= \frac{1}{n^2} \left(\frac{n}{N} \left(1 - \frac{n}{N}\right) \sum_{i=1}^N y_i^2 - \frac{n}{N(N-1)} \left(1 - \frac{n}{N}\right) \sum_{i=1}^N \sum_{j \neq i}^N y_i y_j \right) \\ &= \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N(N-1)} \left(N \sum_{i=1}^N y_i^2 - \left(\sum_{i=1}^N y_i \right)^2 \right) \\ &= \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N(N-1)} \left(N \sum_{i=1}^N (y_i - \bar{y}_N)^2 \right) \\ &= \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}, \end{aligned} \tag{2.11}$$

mientras que la varianza de t_n es:

$$\mathbb{V}(t_n) = N^2 \mathbb{V}(\bar{y}_n) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}. \quad (2.12)$$

Debido a que en aplicaciones reales difícilmente se dispondrá de la población completa para calcular $S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y}_N)^2$, se requieren estimadores de $\mathbb{V}(\bar{y}_n)$ y de $\mathbb{V}(t_n)$, así que se proponen los siguientes estimadores para las varianzas de los estimadores:

$$v(\bar{y}_n) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}, \quad (2.13)$$

$$v(t_n) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}, \quad (2.14)$$

donde

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2, \quad (2.15)$$

veamos que los estimadores propuestos son estimadores insesgados, para lo cual bastará mostrar que $\mathbb{E}(s_y^2) = S_y^2$, primero notemos que:

$$\begin{aligned} s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2, \\ &= \frac{1}{n-1} \sum_{i=1}^n \left((y_i^2 - \bar{y}_N) - (\bar{y}_n - \bar{y}_N) \right)^2, \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n (y_i - \bar{y}_N)^2 - n(\bar{y}_n - \bar{y}_N)^2 \right), \end{aligned}$$

claramente $\mathbb{E}((\bar{y}_n - \bar{y}_N)^2) = \mathbb{V}(\bar{y}_n) = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$, mientras que para el primer sumando podemos expandir la suma con las variables aleatoria I_i por lo tanto:

$$\begin{aligned} \mathbb{E} \left(\sum_{i=1}^n (y_i - \bar{y}_N)^2 \right) &= \mathbb{E} \left(\sum_{i=1}^N (y_i - \bar{y}_N)^2 I_i \right), \\ &= \sum_{i=1}^N (y_i - \bar{y}_N)^2 \frac{n}{N}, \\ &= \frac{n(N-1)}{N} S_y^2, \end{aligned}$$

teniendo finalmente que:

$$\begin{aligned} \mathbb{E}(s_y^2) &= \frac{1}{n-1} \left(\frac{n(N-1)}{N} S_y^2 - \frac{(N-n)}{N} S_y^2 \right), \\ &= S_y^2, \end{aligned}$$

por lo que en efecto tenemos estimadores insesgados.

2.2.3. Estimador de razón

En muchas ocasiones se busca estimar la razón entre dos promedios (o totales), i.e.

$$r_N = \frac{\bar{y}_N}{\bar{x}_N} = \frac{t_{N,y}}{t_{N,x}},$$

en donde X es otra variable de interés de la población bajo estudio, las mediciones en toda la población se denotan como $x_{1:N}$, mientras que la muestra estará dada por $x_{1:n}$. En este caso se usa el estimador de razón

$$r_n = \frac{\bar{y}_n}{\bar{x}_n},$$

que es la base para estimar proporciones (o porcentajes), por lo que es de especial interés para nuestro trabajo.

Obtener la esperanza, varianza y en general las propiedades probabilísticas de un cociente de variables aleatorias no es algo sencillo, la estrategia usada en el muestreo es hacer una aproximación de Taylor de primer orden (o linealización) sobre $r_n = f(\bar{y}_n, \bar{x}_n) = \frac{\bar{y}_n}{\bar{x}_n}$, así:

$$r_n \approx r_N + \frac{1}{\bar{x}_N}(\bar{y}_n - r_N \bar{x}_n), \quad (2.16)$$

como r_N y \bar{x}_N son constantes, entonces:

$$\mathbb{E}(r_n) \approx r_N + \frac{1}{\bar{x}_N}(\mathbb{E}(\bar{y}_n) - r_N \mathbb{E}(\bar{x}_n)) = r_N, \quad (2.17)$$

por lo que r_n es un estimador aproximadamente insesgado de r_N . Para calcular su varianza notemos que podemos reescribir a $\bar{y}_n - r_N \bar{x}_n$ como:

$$\bar{y}_n - r_N \bar{x}_n = \frac{1}{n} \sum_{i=1}^n (y_i - r_N x_i) = \frac{1}{n} \sum_{i=1}^n d_i,$$

definiendo una nueva variable $d_i = y_i - r_N x_i$, cuya media poblacional es $\bar{d}_N = 0$, así:

$$\begin{aligned} \mathbb{V}(r_n) &\approx \mathbb{E}((r_n - r_N)^2) \\ &\approx \frac{1}{\bar{x}_N^2} \mathbb{E}((\bar{y}_n - r_N \bar{x}_n)^2) \\ &= \frac{1}{\bar{x}_N^2} \mathbb{E}((\bar{d}_n - \bar{d}_N)^2) \\ &= \frac{1}{\bar{x}_N^2} \mathbb{V}(\bar{d}_n), \end{aligned}$$

y claramente

$$\mathbb{V}(\bar{d}_n) = \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n}, \quad (2.18)$$

entonces:

$$\mathbb{V}(r_n) \approx \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n \bar{x}_N^2}. \quad (2.19)$$

Nuevamente será necesario proponer un estimador para $\mathbb{V}(r_n)$, tomando en cuenta que si n es lo suficientemente grande entonces $\bar{x}_n \approx \bar{x}_N$ y que s_d^2 es un estimador insesgado de S_d^2 , proponemos:

$$v(r_n) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n\bar{x}_n^2}, \quad (2.20)$$

con

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d}_n)^2, \quad (2.21)$$

así, $v(r_n)$ es un buen estimador de $\mathbb{V}(r_n)$, siendo “aproximadamente” insesgado.

2.2.4. Tamaño de la muestra

Para determinar el tamaño de muestra deberemos partir de 1) un margen de error (o precisión) $\epsilon > 0$ que estemos dispuestos a aceptar en la estimación y 2) la confianza $100(1 - \alpha)\%$ con la que buscaremos obtener ese margen de error. En este trabajo el interés se centra en estimar la razón r_N usando el estimador r_n , entonces lo que se busca, en términos matemáticos, es el tamaño de muestra, n , para asegurar que

$$P(|r_n - r_N| \leq \epsilon) = 1 - \alpha, \quad (2.22)$$

tanto ϵ como α ($0 < \alpha < 1$) los fija el analista de acuerdo a lo que se busca en el estudio. Bajo el supuesto de normalidad asintótica de r_n , i.e. para n suficientemente grande se debe tener que

$$r_n \sim \mathcal{N}(r_N, \mathbb{V}(r_n)),$$

en donde la varianza, $\mathbb{V}(r_n)$, está dada en la expresión (2.19), considerando (2.22) se tiene que

$$\epsilon = z_{1-\frac{\alpha}{2}} \sqrt{\mathbb{V}(r_n)} = z_{1-\frac{\alpha}{2}} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_d^2}{n\bar{x}_N^2}},$$

en donde $z_{1-\frac{\alpha}{2}}$ es el cuantíl de la normal estándar que acumula una área de $1 - \frac{\alpha}{2}$ a la izquierda de él. Así, la expresión del tamaño de muestra n como función de la precisión ϵ y de α está dada por:

$$n = \frac{N z_{1-\frac{\alpha}{2}}^2 S_d^2}{\epsilon^2 \bar{x}_N^2 + z_{1-\frac{\alpha}{2}}^2 S_d^2}. \quad (2.23)$$

Obviamente tanto \bar{x}_N y S_d^2 son cantidades poblacionales desconocidas, lo que se hace es 1) estimarlas mediante una encuesta piloto o 2) calcularlas usando información histórica. En los Conteos Rápidos siempre se usa información de elecciones pasadas para fijar el tamaño de muestra que se usará para estimar las tendencias de la elección actual.

2.3. Muestreo aleatorio estratificado (MAE)

Usando la teoría del capítulo anterior es posible obtener resultados para un esquema de muestreo más complejo. El *muestreo aleatorio estratificado* es aquel en el que nuestra población de N elementos se divide en H sub-poblaciones llamadas estratos, cada una con sus respectivos N_1, N_2, \dots, N_H elementos de tal modo que ningún par de estrato comparte elementos y la unión de todos ellos corresponde a la totalidad de la población original. Para la selección de una muestra de tamaño n , se seleccionan muestras de tamaño n_1, n_2, \dots, n_H en cada estrato de manera que

$$n = \sum_{h=1}^H n_h, \quad (2.24)$$

en donde n_h es la muestra seleccionada en el estrato h . En general se asume que las muestras en cada estrato se obtienen de forma independiente y usando un MASSR en cada estrato.

Es importante mencionar que para definir los estratos es indispensable contar con información a nivel poblacional que nos ayude a identificar las H sub-poblaciones. Por ejemplo, en el Conteo Rápido se cuenta con información del estado, municipio, distrito federal, distrito local, tipo de casilla (urbana, rural o mixta), etc al que pertenece cada una de las N casillas instaladas para la elección. Todas estas variables (o combinaciones de ellas) pueden ser usadas para definir distintas estratificaciones.

2.3.1. Estimadores

En el caso del MAE la notación se vuelve un poco más complicada pues la medición de la variable de interés en la población ahora se registra en cada estrato, entonces se tiene $y_{i,h}$ para $i \in \{1, 2, \dots, N_h\}$ y $h \in \{1, \dots, H\}$. Por lo que $y_{i,h}$ es la observación i -ésima en el estrato h . Entonces, podemos escribir el promedio poblacional como

$$\bar{y}_N = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_{N_h} = \frac{1}{N} \sum_{h=1}^H \sum_{i=1}^{N_h} y_{i,h}.$$

De forma similar se tendrían las observaciones para la muestra, en donde en este caso $i \in \{1, \dots, n_h\}$. Con esta notación, se tiene que el promedio muestral en el estrato h está dado por

$$\bar{y}_{n_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{i,h},$$

y por lo tanto el estimador del total en el estrato h estaría dado por

$$t_{n_h} = N_h \bar{y}_{n_h}.$$

Ahora tenemos que los estimadores de la media y del total toman la siguiente forma:

	Media	Total
Estimador	$\bar{y}_n = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_{n_h}$	$t_n = \sum_{h=1}^H N_h \bar{y}_{n_h} = \sum_{h=1}^H t_{n_h}$

En ocasiones a los estimadores descritos en la tabla anterior se les pone alguna letra o distinción, por ejemplo \bar{y}_n^e , para indicar que son estimadores específicos para una selección vía MAE. Sin embargo, por simplicidad omitiremos esta distinción y esperamos que no haya confusión con los estimadores para el MASSR. Es importante mencionar que un MASSR es un caso especial del MAE cuando se tiene un sólo estrato, por lo que si $H = 1$ recuperaremos los estimadores descritos en la sección para el MASSR.

Dichos estimadores son insesgados, con las siguientes varianzas:

$$\mathbb{V}(\bar{y}_n) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_{y,h}^2}{n_h}, \quad (2.25)$$

$$\mathbb{V}(t_n) = \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_{y,h}^2}{n_h}, \quad (2.26)$$

teniendo los siguientes estimadores:

$$v(\bar{y}_n) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_{y,h}^2}{n_h}, \quad (2.27)$$

$$v(t_n) = \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_{y,h}^2}{n_h}, \quad (2.28)$$

donde $s_{y,h}^2$ es el estimador insesgado de $S_{y,h}^2$ para el MASSR del estrato h , como consecuencia los estimadores $v(\bar{y}_n)$ y $v(t_n)$ son insesgados, los detalles se pueden consultar en Hernández Mendoza (2019).

2.3.2. Estimador de razón combinado

Sean $t_{N,y}$ y $t_{N,x}$ los totales poblacionales para dos variables Y y X respectivamente, de forma general, existen dos propuestas comunes para estimar la razón

$$r_N = \frac{t_{N,y}}{t_{N,x}},$$

entre las dos variables bajo un MAE: el *estimador de razón separado* y el *estimador de razón combinado*. El primero requiere información de una variable auxiliar para toda la población, i.e. $z_{1:N}$, mientras que el segundo sólo necesita la información de las muestras $y_{1:n}$ y $x_{1:n}$ seleccionadas mediante MAE. El caso en el que se cuenta

con información auxiliar se deja como trabajo futuro, por lo que sólo describiremos el estimador de razón combinado, el cual tiene la siguiente expresión:

$$r_n = \frac{t_{n,y}}{t_{n,x}}, \quad (2.29)$$

donde $t_{n,y}$ y $t_{n,x}$ son los estimadores bajo un MAE de los totales poblacionales $t_{N,y}$ y $t_{N,x}$, respectivamente. Siguiendo el proceso de linearización de la sección (2.2.3) y dado que en cada estrato la selección se realiza de manera independiente usando un MASSR, la varianza del estimador se aproxima por:

$$\mathbb{V}(r_n) \approx \frac{1}{t_{N,x}^2} \left[\sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_{d_h}^2}{n_h} \right], \quad (2.30)$$

donde $S_{d_h}^2 = \frac{1}{N_h-1} \sum_{i=1}^{N_h} (d_{i,h} - \bar{d}_{N_h})^2$, $\bar{d}_{N_h} = \frac{1}{N_h} \sum_{i=1}^{N_h} d_{i,h}$ y $d_{i,h} = y_{i,h} - r_N x_{i,h}$.

Con un estimador “aproximadamente” insesgado (ver e.g. Cochran (1977) y Hernández Mendoza (2019)) dado por:

$$v(r_n) = \frac{1}{t_{n,x}^2} \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_{d_h}^2}{n_h}, \quad (2.31)$$

donde $s_{d_h}^2 = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (d_{i,h} - \bar{d}_{n_h})^2$, $\bar{d}_{n_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} d_{i,h}$ y $d_{i,h} = y_{i,h} - r_n x_{i,h}$.

2.4. Aplicación al Conteo Rápido

En el Conteo Rápido partimos de una población de N casillas, en donde cada casilla contiene el número de votos en favor de cada uno de los J candidatos en esa casilla. Entonces, $y_{i,j}$ para $i = 1, \dots, N$ y $j = 1, \dots, J$ indica el número de votos en la casilla i en favor del candidato j . Con lo anterior, es trivial calcular los porcentajes poblacionales

$$p_{N,j} = 100 \frac{\bar{y}_{N,j}}{\sum_{j=1}^J \bar{y}_{N,j}} = 100 \frac{\sum_{i=1}^N y_{i,j}}{\sum_{j=1}^J \sum_{i=1}^N y_{i,j}}, \quad \text{para } j = 1, \dots, J. \quad (2.32)$$

De las N casillas, se selecciona una muestra de n casillas para realizar las estimaciones respectivas. En el desarrollo de la presente tesis analizaremos el caso en el que la muestra se obtiene mediante un MAE, teniendo como caso particular un MASSR en donde los estratos se colapsan a un solo estrato.

2.4.1. Notación

En el caso del MAE la información de la muestra estaría dada por $y_{i,h,j}$, para $i = 1, \dots, n_h$, $h = 1, \dots, H$ y $j = 1, \dots, J$. Entonces, $y_{i,h,j}$ indica el número de votos en la casilla i , del estrato h y en favor del candidato j . Con esta notación $\bar{y}_{n_h,j}$, corresponde al promedio de votos en favor del candidato j en el estrato h y, análogamente, $t_{n_h,j} = N_h \bar{y}_{n_h,j}$ es la estimación para el total de votos en favor del candidato j .

2.4.2. Estimadores y varianza estimada del estimador

Ya que la cantidad a estimar en el Conteo Rápido es el porcentaje de votos obtenidos por el candidato j , i.e. la expresión (2.32), usaremos el estimador de razón combinado para realizar comparativas con otras estrategias en un capítulo posterior. Así, para el total de votos para el candidato j , denotado como $t_{n,j}$, y para el total de votos en la elección, denotado como t_n , se tiene que:

$$t_{n,j} = \sum_{h=1}^H N_h \bar{y}_{n_h,j} = \sum_{h=1}^H t_{n_h,j} \quad y \quad t_n = \sum_{j=1}^J t_{n,j}.$$

Entonces, el estimador de razón combinado para el porcentaje de votos en favor del candidato j es:

$$p_{n,j} = 100 \frac{t_{n,j}}{t_n}, \quad (2.33)$$

cuya varianza queda determinada por la expresión (2.27), es decir:

$$V(p_{n,j}) \approx \frac{100^2}{t_n^2} \left[\sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_{d_{h,j}}^2}{n_h} \right], \quad (2.34)$$

en donde t_N es el total de votos en la población de casillas, $S_{d_{h,j}}^2 = \frac{1}{N_h^2 - 1} \sum_{i=1}^{N_h} (d_{i,h,j} - \bar{d}_{N_h,j})^2$, $\bar{d}_{N_h,j} = \frac{1}{N_h} \sum_{i=1}^{N_h} d_{i,h,j}$, $d_{i,h,j} = y_{i,h,j} - p_{N,j} t_{i,h}$. En donde

$$t_{i,h} = \sum_{j=1}^J y_{i,h,j}.$$

Como la varianza del estimador depende de cantidades poblacionales, nuevamente recurrimos a estimadores muestrales de las cantidades poblacionales:

$$s_{d_{h,j}}^2 = \frac{1}{n_h^2 - 1} \sum_{i=1}^{n_h} (d_{i,h,j} - \bar{d}_{n_h,j})^2, \quad \bar{d}_{n_h,j} = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{d}_{h,i,j} \quad y \quad \hat{d}_{i,h,j} = y_{i,h,j} - p_{n,j} t_{i,h}.$$

Así, el estimador de la varianza del estimador estaría dado por:

$$v(p_{n,j}) = \frac{100^2}{t_n^2} \left[\sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_{d_{h,j}}^2}{n_h} \right], \quad (2.35)$$

nuevamente se tiene que $v(p_{n,j})$ es un estimador aproximadamente insesgado de $V(p_{n,j})$. Ahora, para construir un intervalo del $100 \times (1 - \alpha) \%$ de confianza para $p_{N,j}$, se asume normalidad asintótica para el estimador, i.e.

$$p_{n,j} \sim \mathcal{N}(p_{N,j}, v(p_{n,j})).$$

Es importante mencionar que en el caso del Conteo Rápido para la determinación del tamaño de muestra global n , siempre se asume una *asignación proporcional* para los tamaños de muestra por estrato:

$$n_h = n \left(\frac{N_h}{N} \right). \quad (2.36)$$

2.5. Trabajo futuro: estimación usando información auxiliar

En esta sección se presentan algunas ideas para realizar la estimación de $p_{N,j}$ usando información auxiliar $z_{1:N}$. Por ejemplo, la lista nominal (el número de personas que tienen derecho a votar en cada casilla) o variables socio-económicas del Censo a escala geo-electoral. Estas ideas se dejan para un trabajo futuro, pero las presentamos aquí pues están relacionadas con la teoría descrita en el presente capítulo.

Desde una perspectiva predictiva frecuentista se plantea estimar una cantidad poblacional mediante las observaciones de la muestra \mathcal{S} y una predicción de los valores no observados en $\bar{\mathcal{S}}$, la parte no observada de la población. Dicha predicción puede realizarse proponiendo un modelo asociado a la variable bajo estudio (ver por ejemplo Lohr (2021)) o mediante el uso de alguna variable auxiliar conocida. Bajo la segunda opción Yadav and Mishra (2015) desarrolló varias propuestas para estimar la media de manera robusta usando una variable auxiliar. Aplicando estas ideas al Conteo Rápido tenemos que para el candidato j el total de votos obtenidos en función de la media de votos en la muestra y su complemento por estratos es:

$$\begin{aligned}
 t_{N,j} &= \sum_{h=1}^H t_{N_h,j} \\
 &= \sum_{h=1}^H N_h \bar{y}_{N_h,j} \\
 &= \sum_{h=1}^H N_h \left[\frac{1}{N_h} \sum_{i=1}^{n_h} y_{i,h,j} + \frac{1}{N_h} \sum_{i=(n_h+1)}^{N_h} y_{i,h,j} \right] \\
 &= \sum_{h=1}^H N_h \left[\frac{n_h}{N_h} \bar{y}_{n_h,j} + \frac{N_h - n_h}{N_h} \bar{y}_{N_h - n_h,j} \right].
 \end{aligned}$$

Así, dada una variable auxiliar Z (podríamos pensar en la lista nominal o en alguna variable socio-económica que se podría obtener a través del Censo) se puede proponer un estimador predictor de $\bar{y}_{N_h - n_h,j}$ de la forma $\hat{\bar{y}}_{N_h - n_h,j} = \bar{y}_{n_h,j} f(\bar{z}_{N_h}, \bar{z}_{n_h}, \bar{z}_{N_h - n_h})$, donde f es una función de la media poblacional \bar{z}_{N_h} , la media muestral \bar{z}_{n_h} y/o la media del complemento de la muestra $\bar{z}_{N_h - n_h}$ de nuestra variable auxiliar, todas cantidades estimadas sobre el respectivo estrato h . Por lo tanto para un $\hat{\bar{y}}_{N_h - n_h,j}$ dado, tenemos el siguiente estimador del total de votos para el candidato j :

$$\hat{t}_{N,j} = \sum_{h=1}^H [n_h \bar{y}_{n_h,j} + (N_h - n_h) \hat{\bar{y}}_{N_h - n_h,j}]. \quad (2.37)$$

Como veremos en el Capítulo 3 y específicamente en la expresión (3.14) este enfoque es muy similar al usado en la inferencia Bayesiana, en donde se usa la distribución predictiva para intentar conocer la parte no observada de la población a partir de la parte observada. En cambio aquí se usa información de una variable auxiliar más la

2.5. TRABAJO FUTURO: ESTIMACIÓN USANDO INFORMACIÓN AUXILIAR 23

información de la muestra para intentar predecir el comportamiento de la parte no observada.

Naturalmente un estimador del total de votos en la elección es:

$$\hat{t}_N = \sum_{j=1}^J \hat{t}_{N,j}. \quad (2.38)$$

Por lo tanto el estimador de la proporción de votos para el candidato j es:

$$\hat{p}_j = \frac{\hat{t}_{N,j}}{\hat{t}_N}. \quad (2.39)$$

La posibilidad de obtener una expresión explícita para $\mathbb{V}(\hat{p}_j)$ dependerá del estimador predictivo propuesto.

2.5.1. Propuestas a comparar

Notemos que empleando la relación

$$\bar{z}_{N_h - n_h} = \frac{N_h \bar{z}_{N_h} - n_h \bar{z}_{n_h}}{N_h - n_h}$$

podremos obtener expresiones en términos de cantidades plenamente conocidas puesto que el total y la media poblacionales de nuestra variable auxiliar se supondrán conocidos. Por simplicidad se omitirá el subíndice j asociado al candidato y por claridad se agregará un índice numérico para identificar algunas de las propuestas de estimadores predictores para $\hat{y}_{N_h - n_h}$ encontradas en Yadav and Mishra (2015):

Estimador predictor	Sumando
$\hat{y}_{N_h - n_h}$	$[n_h \bar{y}_{n_h} + (N_h - n_h) \hat{y}_{N_h - n_h}]$
$\hat{y}_{1, N_h - n_h} = \bar{y}_{n_h} + b(\bar{z}_{N_h - n_h} - \bar{z}_{n_h})$	$N_h \bar{y}_{n_h} + b(z_{N_h} - N_h \bar{z}_{n_h})$
$\hat{y}_{2, N_h - n_h} = \bar{y}_{n_h} \left(\frac{\bar{z}_{N_h - n_h}}{\bar{z}_{n_h}} \right)$	$z_{N_h} \left(\frac{\bar{y}_{n_h}}{\bar{z}_{n_h}} \right)$
$\hat{y}_{3, N_h - n_h} = \bar{y}_{n_h} \left(\frac{\bar{z}_{n_h}}{\bar{z}_{N_h - n_h}} \right)$	$\bar{y}_{n_h} \left(\frac{n_h z_{N_h} + N_h(N_h - 2n_h) \bar{z}_{n_h}}{z_{N_h} - z_{n_h}} \right)$
$\hat{y}_{4, N_h - n_h} = \bar{y}_{n_h} \left(\frac{\bar{z}_{n_h}}{\bar{z}_{N_h}} \right)$	$\bar{y}_{n_h} \left(\frac{n_h \bar{z}_{N_h} + (N_h - n_h) \bar{z}_{n_h}}{\bar{z}_{N_h}} \right)$
$\hat{y}_{5, N_h - n_h} = \bar{y}_{n_h} \exp \left(\frac{\bar{z}_{N_h} - \bar{z}_{n_h}}{\bar{z}_{N_h} + \bar{z}_{n_h}} \right)$	$\bar{y}_{n_h} \left(n_h + (N_h - n_h) \exp \left(\frac{z_{N_h} - N_h \bar{z}_{n_h}}{z_{N_h} + (N_h - 2n_h) \bar{z}_{n_h}} \right) \right)$
$\hat{y}_{6, N_h - n_h} = \bar{y}_{n_h} \exp \left(\frac{\bar{z}_{n_h} - \bar{z}_{N_h}}{\bar{z}_{N_h} + \bar{z}_{n_h}} \right)$	$\bar{y}_{n_h} \left(n_h + (N_h - n_h) \exp \left(\frac{N_h \bar{z}_{n_h} - z_{N_h}}{z_{N_h} + (N_h - 2n_h) \bar{z}_{n_h}} \right) \right)$

Capítulo 3

Inferencia Bayesiana bajo poblaciones finitas

En este capítulo se expondrán los conceptos básicos de la inferencia Bayesiana, así como las consideraciones especiales que se deben tomar en cuenta cuando se trabaje con poblaciones finitas, centrándonos en lo referente al modelo de recolección de datos (diseño de muestreo). Este resumen se realiza principalmente con base en las ideas descritas por Bernardo and Smith (2000), Gelman et al. (2013) y Wasserman (2003). Sin embargo, es importante mencionar que una de las primeras propuestas para motivar la aplicación de la inferencia Bayesiana en poblaciones finitas fue descrita por Ericson (1969). Por ejemplo, el concepto de superpoblación que se describe más adelante, el uso del concepto de intercambiabilidad para justificar la existencia de una distribución conjunta para la población completa $y_{1:N}$, así como la idea de utilizar el enfoque predictivo para hacer inferencia para la parte que se desconoce de la población usando la información que sí se conoce. Para finalizar el capítulo y considerando la teoría descrita, se expondrán dos modelos básicos para realizar inferencia en el Conteo Rápido, tanto en el caso de un MASSR como de un MAE.

3.1. Ideas básicas de la inferencia Bayesiana

La principal diferencia entre la inferencia estadística clásica con la inferencia Bayesiana radica en la interpretación de la probabilidad. La inferencia estadística clásica ve la probabilidad como frecuencias relativas, mientras que la inferencia Bayesiana la interpreta como grados de creencia subjetiva, asumiendo que tanto los parámetros como las observaciones son variables aleatorias y actualizando el conocimiento que se tiene del problema cuando se recibe información mediante el teorema de Bayes. Por ejemplo, al lanzar un dado, la probabilidad de obtener un 6 desde el punto de vista

- Frecuentista. Se interpreta como la frecuencia relativa de obtener un 6 en el largo plazo, si se repitiera el experimento muchas veces. Si lanzamos el dado 100 veces y obtenemos un 6 en 20 ocasiones, la probabilidad frecuentista sería $20/100 = 0.2$.

- Bayesiana. Se interpreta como la medida de nuestra creencia inicial en esa posibilidad. Inicialmente, podríamos tener una creencia subjetiva basada en la forma en que sentimos el dado o en experiencias previas, quizás asignando un 10 % de probabilidad a obtener un 6. Si luego lanzamos el dado y observamos un 6, actualizaríamos nuestra creencia inicial utilizando el teorema de Bayes para reflejar esta nueva evidencia. La probabilidad Bayesiana sería una combinación de nuestra creencia inicial y la nueva información del resultado observado.

Desde el punto de vista Bayesiano las creencias iniciales del investigador se condensan en la propuesta de una densidad conjunta para los datos observados y los parámetros del modelo, dicha densidad se expresa como el producto de dos densidades propuestas; la densidad condicional de los datos observados dado los parámetros y una densidad inicial, o prior, de dichos parámetros:

$$p(y, \theta) = p(y|\theta)p(\theta). \quad (3.1)$$

La primera actualización de dichas densidades es resultado directo del teorema de Bayes, resultando en la distribución posterior de los parámetros dada la observación de y :

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y, \theta)d\theta}. \quad (3.2)$$

Así mismo, para una observación futura \tilde{y} se puede obtener la distribución posterior predictiva dada la observación de y :

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}|\theta)p(\theta|y)d\theta, \quad (3.3)$$

en donde se asume la independencia condicional

$$p(\tilde{y}, y|\theta) = p(\tilde{y}|\theta)p(y|\theta),$$

y además que $p(\tilde{y}|\theta)$ sigue el modelo propuesto inicialmente por el investigador.

3.2. Intercambiabilidad y teorema de representación

Desde la perspectiva Bayesiana es muy relevante considerar la estructura de dependencia que puedan tener los datos $y_{1:n} = \{y_1, y_2, \dots, y_n\}$ a ser analizados, la cual puede reflejarse en propiedades que satisfaga su distribución conjunta. Una de las estructuras de dependencia más simple es la que supone que el orden de recolección de las observaciones no es relevante, dicha cualidad se llama *intercambiabilidad* y se expresa matemáticamente como que la distribución conjunta $p(y_{1:n})$ satisface que:

$$p(y_{1:n}) = p(y_{\nu(1)}, y_{\nu(2)}, \dots, y_{\nu(n)}), \quad (3.4)$$

para toda permutación ν del conjunto $\{1, \dots, n\}$, más aún, una sucesión de variables aleatorias se dice *intercambiable* si cualquier subconjunto finito de ellas satisface lo

anterior.

La suposición de intercambiabilidad tiene su principal implicación expuesta en el *teorema de representación* de de Finetti, el cual dice que en el límite $n \rightarrow \infty$ una distribución conjunta intercambiabile puede expresarse con un modelo paramétrico $p(y|\theta)$ y una distribución prior $p(\theta)$ mediante:

$$p(y_{1:n}) = \int_{\Theta} \prod_{i=1}^n p(y_i|\theta)p(\theta)d\theta, \quad (3.5)$$

es decir, con una mezcla de distribuciones independientes e idénticas, sugiriendo además la consideración de una *superpoblación* para nuestros datos y_i distribuida por $p(y|\theta)$ y determinada por el parámetro desconocido θ .

3.3. Consideraciones especiales bajo poblaciones finitas

Existen diversas configuraciones “experimentales” en las cuales se debe considerar y modelar el diseño de recolección de datos, tales como encuestas por muestreo, experimentos aleatorizados, problemas de no-respuesta o datos faltantes, y estudios con datos censurados o truncados.

3.3.1. Generalidades

Partimos del supuesto de que los *datos observados* han sido recolectados de un conjunto más amplio de *datos completos* (o *datos potenciales*), por lo cual existen *datos faltantes*. Así, considerando datos univariados, sean $y_{1:N} = (y_1, \dots, y_N)$ el vector de datos completos e $I_{1:N} = (I_1, \dots, I_N)$ el vector de observación de $y_{1:N}$, en el que $I_i = 1$ indica que y_i fue observado, mientras que $I_j = 0$ indica que y_j falta. Introduzcamos la notación en la que $y_{1:n} = \{y_i | I_i = 1\}$ corresponde a nuestra muestra observada y $y_{(n+1):N} = \{y_j | I_j = 0\}$ son los datos no observados, también se asumirá que $I_{1:N}$ es observado (conocido) en su totalidad.

3.3.2. Verosimilitud-completa y verosimilitud-observada

Como se mencionó anteriormente, desde el punto de vista Bayesiano todas son variables aleatorias, por lo que es indispensable considerar a la distribución conjunta de los datos completos $y_{1:N}$ y las variables que nos dan información de cómo se ha seleccionado la muestra $I_{1:N}$. La distribución conjunta que considera un modelo para la recolección de datos se debe separar en dos partes, lo referente a los datos completos $y_{1:N}$ y lo correspondiente a la inclusión $I_{1:N}$, así podemos expresar la *verosimilitud-completa* como el producto de las verosimilitudes de ambos factores:

$$p(y_{1:N}, I_{1:N} | \theta, \phi) = p(y_{1:N} | \theta)p(I_{1:N} | \phi, y_{1:N}), \quad (3.6)$$

donde θ es el parámetro de la distribución de los datos completos $y_{1:N}$ y ϕ es un posible parámetro de la distribución de $I_{1:N}$. Ya que sólo observamos $(y_{1:n}, I_{1:N})$, la verosimilitud apropiada para realizar inferencias es la *verosimilitud-observada*:

$$p(y_{1:n}, I_{1:N} | \theta, \phi) = \int p(y_{1:N}, I_{1:N} | \theta, \phi) dy_{n+1:N}. \quad (3.7)$$

3.3.3. Superpoblación y población finita

La obtención de inferencias suele constar de dos pasos, realizar inferencias sobre la superpoblación a través de la posterior de los parámetros, $p(\theta, \phi | y_{1:n}, I_{1:N})$, para después realizar inferencias de una población finita mediante la posterior para nuestros datos faltantes $p(y_{n+1:N} | \theta, \phi, I_{1:N})$, realizando una *imputación*. En ocasiones se opta por realizar inferencias imputando mediante la distribución predictiva ya que permite comparar distintos modelos con distintas parametrizaciones.

3.3.4. Ignorabilidad

Cuando el diseño de muestreo no está relacionado con la variable que se busca estimar, entonces se cumple que

$$p(y_{1:N}, I_{1:N} | \theta, \phi) = p(y_{1:N} | \theta) p(I_{1:N} | \phi, y_{1:N}) = p(y_{1:N} | \theta) p(I_{1:N} | \phi),$$

por lo que la distribución posterior estaría dada por

$$\begin{aligned} p(\theta | y_{1:n}) &\propto p(\theta) p(y_{1:n} | \theta), \\ &\propto p(\theta) \int p(y_{1:N} | \theta) dy_{n+1:N}, \\ &\propto p(\theta) \int p(y_{1:N}, I_{1:N} | \theta, \phi) dy_{n+1:N}. \end{aligned}$$

En este caso se dice que el diseño de muestreo es *ignorable*, por lo cual las distribuciones posterior y predictiva quedan completamente determinadas de la manera usual.

3.3.5. Muestreo aleatorio simple: Diseño ignorable y conocido

Los modelos más sencillos de recolección de datos son aquellos que son ignorables y conocidos en el sentido de que:

$$p(I_{1:N} | \phi, y_{1:N}) = p(I_{1:N}), \quad (3.8)$$

es decir, no depende de $y_{1:N}$ y además no hay parámetro desconocido ϕ . Así, sólo la verosimilitud $p(y_{1:n} | \theta)$ y la prior $p(\theta)$ son necesarias para realizar inferencias. En el caso de un MASSR en el que se obtiene una muestra \mathcal{S} de tamaño n , es claro que:

$$I_i = \begin{cases} 1 & \text{si } y_i \in \mathcal{S}, \\ 0 & \text{en otro caso,} \end{cases} \quad (3.9)$$

además, si se define que:

$$p(I_{1:N}) = \begin{cases} \binom{N}{n}^{-1} & \text{si } \sum_{i=1}^N I_i = n, \\ 0 & \text{en otro caso,} \end{cases} \quad (3.10)$$

entonces el diseño es ignorable, por lo cual la inferencia es sencilla.

3.3.6. Muestreo aleatorio estratificado

Ahora considerando la división de $y_{1:N}$ en H estratos con sus respectivos N_1, N_2, \dots, N_H elementos, tal que $N = \sum_{i=1}^H N_i$, debemos diseñar un MAE con una muestra \mathcal{S} de n elementos, en el cual en cada estrato se selecciona una muestra \mathcal{S}_h con n_h elementos mediante un MASSR, teniendo ahora que $I_{1:N}$ es una matriz, cuya i -ésima columna, I_i , es un vector con H componentes determinadas por:

$$I_{i,h} = \begin{cases} 1 & \text{si } y_i \in \mathcal{S}_h, \\ 0 & \text{en otro caso,} \end{cases} \quad (3.11)$$

al tener un MASSR en cada estrato podemos definir naturalmente

$$p(I_{1:N_1}, I_{1:N_2}, \dots, I_{1:N_H}) = \begin{cases} \left(\prod_{h=1}^H \binom{N_h}{n_h} \right)^{-1} & \text{si } \sum_{i=1}^N I_{i,h} = n_h \forall h \in \{1, \dots, H\}, \\ 0 & \text{en otro caso,} \end{cases} \quad (3.12)$$

para tener un diseño ignorable.

3.4. Aplicación al Conteo Rápido

Los conteos que modelaremos están dados en forma de matriz. Con los votos totales en favor de cada fuerza política, i.e. $y_{i,j}$, para $i = 1, \dots, N$ y $j = 1, \dots, J$. En donde $y_{i,j}$ es el número de votos en la casilla i en favor del candidato j .

Con lo anterior, es trivial calcular los porcentajes poblacionales

$$p_{N,j} = 100 \frac{\bar{y}_{N,j}}{\sum_{j=1}^J \bar{y}_{N,j}} = 100 \frac{\sum_{i=1}^N y_{i,j}}{\sum_{j=1}^J \sum_{i=1}^N y_{i,j}}, \quad \text{para } j \in \{1, \dots, J\}. \quad (3.13)$$

El objetivo en esta sección será describir modelos para estimar los porcentajes (3.13) y usar el enfoque Bayesiano para obtener estimaciones por intervalo de las mismas. Por simplicidad omitiremos el sub-índice j y nos concentraremos en describir la manera en que modelaremos \bar{y}_N . Una vez hecho lo anterior vía simulación es sencillo obtener estimaciones por intervalo de los porcentajes poblacionales. Adicionalmente, asumiremos que la muestra de votos en las casillas seleccionadas en muestra $y_{1:n}$ se obtiene vía un MASSR que de acuerdo a la sección anterior es un diseño de muestreo ignorable.

Empezamos descomponiendo el promedio poblacional \bar{y}_N como un promedio ponderado entre el promedio de la parte observada y la no observada, i.e.

$$\bar{y}_N = \left(\frac{n}{N}\right)\bar{y}_n + \left(\frac{N-n}{N}\right)\bar{y}_{N-n} \quad (3.14)$$

en donde \bar{y}_n es la media de la parte observada de la población y \bar{y}_{N-n} es la media de la parte no observada de la población.

Utilizando la expresión (3.14) obtendremos la posterior predictiva

$$p(\bar{y}_{N-n}|y_{1:n}) = \int p(\bar{y}_{N-n}|\theta)p(\theta|y_{1:n})d\theta,$$

en donde $p(\theta|y_{1:n}) \propto p(y_{1:n}|\theta)p(\theta)$ es la distribución posterior.

3.4.1. Modelo normal univariado

En este caso se asumirá que los votos recibidos $y_{1:n}$ de las casillas seleccionadas son una m.a. de una población $Y \sim \mathcal{N}(\mu, \sigma^2)$. Por lo tanto, la verosimilitud asociada a la muestra se puede expresar como:

$$\begin{aligned} p(y_{1:n}|\mu, \sigma^2) &\propto \sigma^{-n} e^{-\frac{1}{2\sigma^2}(\sum_{i=1}^n (y_i - \mu)^2)}, \\ &\propto \sigma^{-n} e^{-\frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y}_n - \mu)^2)}, \\ &\propto p(\bar{y}_n, s^2|\mu, \sigma^2), \end{aligned}$$

con $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2$. Asumiendo una distribución inicial poco informativa, $p(\mu, \sigma^2) \propto \sigma^{-2}$, la distribución posterior está dada por:

$$\begin{aligned} p(\mu, \sigma^2|\bar{y}_n, s^2) &\propto p(\mu, \sigma^2)p(\bar{y}_n, s^2|\mu, \sigma^2), \\ &\propto \sigma^{-(n+2)} e^{-\frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y}_n - \mu)^2)}. \end{aligned}$$

Permitiéndonos un cambio de notación para facilitar el siguiente desarrollo tendremos que $\bar{y}_{N-n} = \bar{y}_r = \frac{1}{r} \sum_{i=1}^r y_i$, donde claramente $r = N - n$, entonces la posterior predictiva para \bar{y}_r está dada por

$$p(\bar{y}_r|\bar{y}_n, s^2) = \int_0^\infty \int_{-\infty}^\infty p(\bar{y}_r|\mu, \sigma^2)p(\mu, \sigma^2|\bar{y}_n, s^2)d\mu d\sigma^2,$$

por el supuesto de normalidad se tiene que

$$p(\bar{y}_r|\mu, \sigma^2) = \mathcal{N}\left(\mu, \frac{\sigma^2}{r}\right),$$

por lo tanto tenemos que la posterior predictiva para la media es:

$$\begin{aligned} p(\bar{y}_r|\bar{y}_n, s^2) &= \int_0^\infty \int_{-\infty}^\infty p(\bar{y}_r|\mu, \sigma^2)p(\mu, \sigma^2|\bar{y}_n, s^2)d\mu d\sigma^2, \\ &\propto \int_0^\infty \int_{-\infty}^\infty \frac{1}{(\sigma^2)^{\frac{n+3}{2}}} e^{-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}_n - \mu)^2 + r(\bar{y}_r - \mu)^2]} d\mu d\sigma^2, \\ &\propto \int_0^\infty \frac{1}{(\sigma^2)^{\frac{n+2}{2}}} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + \frac{nr(\bar{y}_r - \bar{y}_n)^2}{n+r}\right)} d\sigma^2. \end{aligned}$$

Sea $z = \frac{A}{2\sigma^2}$ con $A = \left[(n-1)s^2 + \frac{nr(\bar{y}_r - \bar{y}_n)^2}{n+r} \right]$. Entonces se tiene que:

$$\begin{aligned} p(\bar{y}_r | \bar{y}_n, s^2) &\propto A^{-\frac{n}{2}} \int_0^\infty z^{n-2} e^{-z} dz, \\ &\propto \left[1 + \frac{(\bar{y}_r - \bar{y}_n)^2}{(n-1)\left(\frac{n+r}{nr}\right)s^2} \right]^{-\frac{n}{2}}, \\ &\propto t_{n-1} \left(\bar{y}_n, \left(\frac{n+r}{nr} \right) s^2 \right). \end{aligned}$$

En donde $t_\nu(\mu, \sigma^2)$ denota una t-Student con media μ , varianza σ^2 y ν grados de libertad. Considerando la expresión (3.14) en donde se descompone a la media de la población como una parte observada y una parte no observada. Se tiene que \bar{y}_{N-n} es una cantidad aleatoria, pero \bar{y}_n es una cantidad conocida pues ya se observó. Así, por el desarrollo anterior se tiene que

$$p(\bar{y}_{N-n} | \bar{y}_n, s^2) = t_{n-1} \left(\bar{y}_n, \left(\frac{N}{n(N-n)} \right) s^2 \right),$$

e inmediatamente se sigue que

$$p(\bar{y}_N | \bar{y}_n, s^2) = t_{n-1} \left(\bar{y}_n, \left(1 - \frac{n}{N} \right) \frac{s^2}{n} \right).$$

Con esta expresión será sencillo simular muchos promedios de votos para cada candidato y por simulación estimaremos los porcentajes (3.13).

3.4.2. Modelo Poisson

Dado el contexto discreto propio de un conteo de votos, resulta natural considerar un modelo que parta de alguna distribución discreta, como es el caso de la distribución Poisson. Así, ahora se asumirá que los votos recibidos $y_{1:n}$ de las casillas seleccionadas son una m.a. de una distribución $Y \sim \text{Pois}(\lambda)$. Por lo tanto, la verosimilitud asociada a la muestra se puede expresar como:

$$p(y_{1:n} | \lambda) \propto \lambda^{t_n} e^{-n\lambda}, \quad (3.15)$$

en donde $t_n = \sum_{i=1}^n y_i$.

Asumiendo una distribución inicial Gamma, i.e.

$$p(\lambda | \alpha, \beta) \propto \lambda^{\alpha-1} e^{-\beta\lambda},$$

la distribución posterior estaría dada por:

$$\begin{aligned} p(\lambda | y_{1:n}) &\propto p(\lambda | \alpha, \beta) p(y_{1:n} | \lambda), \\ &\propto \lambda^{\alpha+t_n-1} e^{-(\beta+n)\lambda}. \end{aligned}$$

Sabemos que para una m.a. $y_{1:r}$ de v.a. $\text{Pois}(\lambda)$, se tiene que el total muestral $t_r = \sum_{i=1}^r y_i \sim \text{Pois}(r\lambda)$. Entonces, la distribución predictiva para t_r está dada por

$$\begin{aligned} p(t_r|y_{1:n}) &\propto \int_0^\infty p(t_r|\lambda)p(\lambda|y_{1:n})d\lambda, \\ &\propto \frac{r^{t_r}}{t_r!} \int_0^\infty \lambda^{\alpha+t_n+t_r-1} e^{-(n+\beta+r)\lambda} d\lambda, \end{aligned}$$

usando el siguiente cambio de variable $u = (n + \beta + r)\lambda$, tenemos que:

$$\begin{aligned} p(t_r|y_{1:n}) &\propto A \int_0^\infty u^{\alpha+t_n+t_r-1} e^{-u} du, \\ &\propto \frac{\Gamma(\alpha + t_n + t_r)}{t_r! \Gamma(\alpha + t_n)} \left(\frac{n + \beta}{n + \beta + r} \right)^{\alpha+t_n} \left(\frac{r}{n + \beta + r} \right)^{t_r} \\ &\propto NB \left(t_r \mid \frac{n + \beta}{n + \beta + r}, \alpha + t_n \right). \end{aligned}$$

Donde $A = \frac{r^{t_r} (n + \beta)^\alpha}{t_r! \Gamma(\alpha + t_n) (n + \beta + r)^{\alpha+t_n+t_r}}$.

Entonces, para generar muestra aleatorias de la predictiva $p(\bar{y}_N|y_{1:n})$, bajo el modelo Poisson, primero generamos

$$p(t_{N-n}|y_{1:n}) = NB \left(t_{N-n} \mid \frac{n + \beta}{n + \beta + r}, \alpha + t_n \right),$$

para posteriormente aplicar la transformación

$$\bar{y}_N = \left(\frac{n}{N} \right) \bar{y}_n + \left(\frac{1}{N} \right) t_{N-n}.$$

Para verificar que el proceso anterior en efecto genera muestras aleatorias de $p(\bar{y}_N|y_{1:n})$, simplemente hay que notar que $t_{N-n} = (N - n)\bar{y}_{N-n}$ por lo que la expresión anterior y la expresión (3.14) son equivalentes.

3.4.3. Muestreo aleatorio estratificado (MAE)

Los modelos normal y Poisson pueden aplicarse directamente a muestras obtenidas mediante MASSR ya que es un diseño de muestro ignorable. Sin embargo, para extender estas ideas a diseños de muestreo estratificado, se requieren conceptos adicionales. En este caso, asumimos que hay N unidades divididas en H estratos, donde el estrato h contiene N_h unidades, con $N = \sum_{h=1}^H N_h$. Se selecciona de manera independiente una muestra aleatoria simple sin reemplazo de unidades en cada estrato, y los tamaños de muestra dentro de cada estrato se denotan como n_h , para $h = 1, \dots, H$, siendo el tamaño de muestra total $n = \sum_{h=1}^H n_h$. Este procedimiento obviamente también produce un diseño de muestreo ignorable a nivel estrato. En donde se asume que $y_{1:n_h}$

es la muestra observada del estrato h , para $h = 1, \dots, H$. Bajo un diseño de muestreo estratificado, la media poblacional \bar{y}_N se puede expresar como

$$\bar{y}_N = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_{N_h}, \quad (3.16)$$

con

$$\bar{y}_{N_h} = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{i,h}.$$

Por lo tanto, podemos utilizar los métodos descritos en la sección anterior para generar muestras de $p(\bar{y}_{N_h} | y_{1:n_h})$, y a través de (3.16) se obtienen muestras de $p(\bar{y}_N | y_{1:n})$ de manera sencilla. Obviamente este proceso es computacionalmente demandante, ya que se tiene que replicar en todos los estratos, pero esto no es de ninguna manera un obstáculo para poder aplicarlo en conjuntos de datos reales.

Capítulo 4

Inferencia predictiva

4.1. Motivación y panorama general

La idea de inferencia predictiva vía martingalas descrita por Fong et al. (2021) utiliza el enfoque Bayesiano para la cuantificación de la incertidumbre, pero la interpreta bajo un contexto de datos faltantes. Supongamos que tenemos una población infinita $y_{1:\infty}$ y que observamos una m.a. $y_{1:n}$. Entonces, se define un modelo predictivo de la siguiente manera

$$p(y_{n+1:\infty}|y_{1:n}) = \prod_{i=n}^{\infty} p(y_{i+1}|y_{1:i}),$$

con

$$p(y_{i+1}|y_{1:i}) = \int p(y_{i+1}|\theta)p(\theta|y_{1:i})d\theta,$$

i.e. la distribución predictiva usual Bayesiana. En este caso, se realiza la predicción a un paso y se actualiza la predictiva: se genera la predicción para la observación $n+1$, con esta nueva información se actualiza la predictiva y se genera la predicción para $n+2$. De forma puntual, la actualización se realiza vía

$$y_{i+1} \sim p(y_{i+1} | y_{1:i}), \quad (4.1)$$

$$\theta_{i+1} = \mathbb{E}(\theta | y_{1:i+1}), \quad (4.2)$$

para $i \in \{n, n+1, \dots\}$. Con este proceso, se imputa la población completa $(y_{1:n}, y_{n+1:\infty})$. Además, se asume que el parámetro que se desea conocer está dado por $\theta_{\infty} = \theta(y_{1:n}, y_{n+1:\infty})$. Por lo tanto, si se conoce la población completa, se conoce θ_{∞} .

Obviamente, no es posible imputar un número infinito de observaciones, $y_{n+1:\infty}$, pero sí es posible imputar $y_{n+1:N}$, para N suficientemente grande. Así, con la información $(y_{1:n}, y_{n+1:N})$ se aproxima θ_{∞} , mediante algún funcional estadístico, $\theta_N = \theta(y_{1:n}, y_{n+1:N})$. Desde la perspectiva Bayesiana la esperanza posterior $\theta_N = \mathbb{E}(\theta|y_{1:N})$, intuitivamente debería ser un buen estimador puntual para θ_{∞} . Usando la teoría sobre aplicaciones de las martingalas, desarrollada por Doob (1949), Fong et al. (2021) confirma lo anterior observando que de hecho

$$\theta_N \xrightarrow{a.s.} \theta_{\infty}, \quad (4.3)$$

en donde, además de algunas condiciones no restrictivas, el supuesto clave es que al actualizar θ_{N-1} a θ_N se tenga una martingala, i.e.

$$\mathbb{E}(\theta_N | y_{1:N-1}) = \theta_{N-1}, \quad a.s. \quad (4.4)$$

esto se verifica en la expresión (4.2) ya que

$$\mathbb{E}(\theta_{i+1} | y_{1:i}) = \mathbb{E}(\mathbb{E}(\theta | y_{1:i+1}) | y_{1:i}) = \mathbb{E}(\theta | y_{1:i}) = \theta_i.$$

Posteriormente, Holmes and Walker (2023) notaron que desde la inferencia Bayesiana y usando como estimador puntual la esperanza posterior, siempre se cumple (4.4). Pero demostraron que el resultado de Doob, también se podría aplicar desde la perspectiva frecuentista. En este caso, el modelo predictivo estaría dado por

$$p(y_{n+1:\infty} | x_{1:n}) = \prod_{i=n}^{\infty} p(y_{i+1} | \hat{\theta}_i).$$

El punto de partida, $\hat{\theta}_n$, podría ser el estimador máximo verosímil basado en la muestra $y_{1:n}$. La actualización sería similar a la descrita para la versión Bayesiana, i.e.

$$y_{i+1} \sim p(y_{i+1} | \hat{\theta}_i), \quad (4.5)$$

$$\hat{\theta}_{i+1} = \psi(\hat{\theta}_i, y_{i+1}), \quad (4.6)$$

para $i \in \{n, n+1, \dots\}$, con ψ alguna función que cumpla la condición de la martingala, i.e.

$$\mathbb{E}(\hat{\theta}_{i+1} | y_{1:i}) = \hat{\theta}_i.$$

Adicionalmente, para garantizar la convergencia (4.3), es necesario demostrar que la varianza, $Var(\hat{\theta}_\infty)$, es finita.

4.2. Enfoque predictivo bajo poblaciones finitas

La adaptación del marco de inferencia predictiva de martingalas propuesto por Fong et al. (2021), a una población finita $y_{1:N}$, es trivial. Supongamos que $y_{1:n}$ es una muestra aleatoria observada de la población finita. Se asume un modelo paramétrico $p(y|\theta)$ para los datos, donde el valor verdadero es $\theta(y_{1:N}) = \theta_N$. La parte faltante de la población $y_{n+1:N}$ se imputa mediante cualquiera de las dos estrategias predictivas (Bayesiana o frecuentista). En resumen, este es el mismo procedimiento que se describiría en Fong et al. (2021) con la diferencia obvia de que detenemos el proceso en N (el tamaño de nuestra población), mientras que Fong et al. (2021) con $N = \infty$ esperaba la convergencia del martingala. Adicionalmente, en la versión frecuentista, en nuestro caso no es necesario demostrar la existencia de θ_∞ asegurando que $Var(\theta_\infty) < \infty$.

Es muy importante mencionar que debido a que el enfoque predictivo está relacionado fuertemente con el paradigma Bayesiano, para poder aplicar la estrategia predictiva es indispensable que el diseño de muestreo sea ignorable.

Finalmente, es interesante señalar que dado que la población completa $\{y_{1:n}, Y_{n+1:N}\}$ se imputa, es posible estimar virtualmente cualquier $\eta_N = \eta(y_{1:N})$ a través de $\hat{\eta}_N = \eta(y_{1:n}, Y_{n+1:N})$. Aquí, θ representa el parámetro que rige el comportamiento del modelo asumido para los datos, mientras que η es una estadística funcional que es el foco de la inferencia. A través de las m poblaciones imputadas, es posible obtener $\hat{\eta}_N^1, \hat{\eta}_N^2, \dots, \hat{\eta}_N^m$. Nos centraremos en la estimación de proporciones poblacionales (3.13). Aunque este proceso puede ser computacionalmente exigente, el procesamiento en paralelo facilita la implementación.

4.3. Metodología

La perspectiva predictiva vía martingalas está estrechamente relacionada con el paradigma Bayesiano, por lo que obviamente el uso de un modelo $p(y|\theta)$, así como el supuesto de intercambiabilidad sobre $y_{1:n}$ son relevantes.

En este caso, para N suficientemente grande, se imputan B poblaciones completas. Con cada población se estiman los parámetros $\hat{\eta}_N^1, \hat{\eta}_N^2, \dots, \hat{\eta}_N^B$. En donde η es alguna función de los elementos de toda la población y $\hat{\eta}_N$ es algún funcional estadístico basado en la N elementos de la población. Pueden utilizarse los cuantiles correspondientes de la muestra aleatoria $\hat{\eta}_N^1, \hat{\eta}_N^2, \dots, \hat{\eta}_N^B$ para cuantificar la incertidumbre sobre la estimación de η . Los Algoritmos 1 y 2 describen tanto la aproximación Bayesiana como frecuentista del enfoque predictivo vía martinagalas.

Algoritmo 1: Bayesiano	Algoritmo 2: Frecuentista
<p>Input: $y_{1:n}, n, N, B$;</p> <p>1 for $h \leftarrow 1$ to B do</p> <p>2 $y_{1:n,h} \leftarrow y_{1:n}$;</p> <p>3 for $i \leftarrow n$ to $N - 1$ do</p> <p>4 $y_{i+1,h} \sim p(y_{i+1} y_{1:i})$;</p> <p>5 $\hat{\eta}_N^h \leftarrow \eta(y_{1:n,h}, y_{n+1:N,h})$;</p> <p>Output: $\hat{\eta}_N^1, \hat{\eta}_N^2, \dots, \hat{\eta}_N^B$;</p>	<p>Input: $y_{1:n}, n, N, B$;</p> <p>1 for $h \leftarrow 1$ to B do</p> <p>2 $y_{1:n,h} \leftarrow y_{1:n}$;</p> <p>3 $\hat{\theta}_n \leftarrow T(y_{1:n,h})$;</p> <p>4 for $i \leftarrow n$ to $N - 1$ do</p> <p>5 $y_{i+1,h} \sim p(y_{i+1} \hat{\theta}_i)$;</p> <p>6 $\hat{\theta}_{i+1} \leftarrow \psi(\hat{\theta}_i, y_{i+1,h})$;</p> <p>7 $\hat{\eta}_N^h \leftarrow \eta(y_{1:n,h}, y_{n+1:N,h})$;</p> <p>Output: $\hat{\eta}_N^1, \hat{\eta}_N^2, \dots, \hat{\eta}_N^B$;</p>

4.3.1. Modelo normal univariado

Para ejemplificar las dos visiones del enfoque predictivo vía martingalas se asume que $y_{1:n}$ es una muestra de una distribución normal $\mathcal{N}(y|\theta, \sigma^2)$, con media y varianza desconocidas.

- Enfoque predictivo Bayesiano. Es sencillo verificar que asumiendo una prior no informativa $p(\mu, \sigma^2) \propto \sigma^{-2}$ la predictiva está dada por

$$y_{n+1} \sim p(y_{n+1} | y_{1:n}) = t_{n-1} \left(\bar{y}_n, \left(1 + \frac{1}{n} \right) s_n^2 \right).$$

Así, la actualización de la predictiva a un paso esta dada por

$$y_{i+1} \sim p(y_{i+1} | y_{1:i}) = t_{i-1} \left(\bar{y}_i, \left(1 + \frac{1}{i} \right) s_i^2 \right),$$

para $i \in \{n, n+1, \dots, N\}$. Cada vez que se genera una predicción, la actualización se realiza de manera eficiente definiendo: $a_i = \sum_{r=1}^i y_r$ y $b_i = \sum_{r=1}^i y_r^2$. Entonces, la actualización de la media y la varianza considerando la nueva observación se hace vía $\bar{y}_{i+1} = (a_i + y_{i+1}) / (i+1)$ y $s_{i+1}^2 = (b_i + y_{i+1}^2 - (i+1)\bar{y}_{i+1}^2) / i$. Observemos que $s_i^2 = \left(\sum_{r=1}^i y_r^2 - i\bar{y}_i^2 \right) / (i-1)$.

- Enfoque predictivo frequentista. Holmes and Walker (2023) obtuvieron el algoritmo predictivo que se describe a continuación.

Se inicializan $\hat{\mu}_n = \bar{y}_n$ y $\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{\mu}_n)^2$. Así, para $i \in \{n, \dots, N-1\}$ tenemos que

$$y_{i+1} \sim \mathcal{N}(y | \hat{\mu}_i, \hat{\sigma}_i^2), \quad (4.7)$$

en donde la actualización de los parámetros está dada por

$$\hat{\mu}_{i+1} = \frac{1}{i+1} (i\hat{\mu}_i + y_{i+1}), \quad y \quad \hat{\sigma}_{i+1}^2 = \frac{1}{k_{i+1}} \left[k_i \hat{\sigma}_i^2 + \frac{i+1}{i+2} (\hat{\mu}_i - y_{i+1})^2 \right],$$

con $k_i = \sum_{j=1}^i \frac{j}{j+1}$.

Es sencillo verificar que en efecto se obtienen martingalas, para la media se tiene:

$$\begin{aligned} \mathbb{E}(\hat{\mu}_{i+1} | y_{1:i}) &= \frac{1}{i+1} [i\mathbb{E}(\hat{\mu}_i | y_{1:i}) + \mathbb{E}(y_{i+1} | y_{1:i})], \\ &= \frac{1}{i+1} (i\hat{\mu}_i + \hat{\mu}_i), \\ &= \hat{\mu}_i, \end{aligned}$$

para la varianza calculamos

$$\begin{aligned} \mathbb{E}(\hat{\sigma}_{i+1}^2 | y_{1:i}) &= \frac{1}{k_{i+1}} \left[k_i \mathbb{E}(\hat{\sigma}_i^2 | y_{1:i}) + \frac{i+1}{i+1} (\mathbb{E}((\hat{\mu}_i - y_{i+1})^2 | y_{1:i})) \right], \\ &= \frac{1}{k_{i+1}} \left[k_i \hat{\sigma}_i^2 + \frac{i+1}{i+1} \hat{\sigma}_i^2 \right], \\ &= \hat{\sigma}_i^2, \end{aligned}$$

por lo que en efecto se tienen martingalas en ambos casos.

En ambos casos se imputa la población completa $\{y_{1:n}, y_{n+1:N}\}$. Por lo que una vez con la población completa, sería posible estimar cualquier funcional estadístico η_N , vía $\hat{\eta}_N = \eta(y_{1:n}, y_{n+1:N})$.

4.4. Inferencia predictiva bajo MASSR y MAE

Bajo el marco predictivo de martingala para imputar la población completa, procedemos imputando la población de cada estrato, es decir, $\{y_{1:n_h}, Y_{n_h+1:N_h}\}$, para $h = 1, \dots, H$. Con la población completa imputada, ahora podemos estimar prácticamente cualquier parámetro de interés poblacional, denotado por η_N , que es una función de la población completa. Específicamente, podemos estimar η_N mediante $\eta_N = \eta(y_{1:n_1}, Y_{n_1+1:N_1}, y_{1:n_2}, Y_{n_2+1:N_2}, \dots, y_{1:n_H}, Y_{n_H+1:N_H})$.

Capítulo 5

Inferencia predictiva usando modelos de mezclas

Después de haber explorado el enfoque predictivo mediante martingalas en el capítulo previo, el presente capítulo se centra en el desarrollo de un modelo del número de votos para los candidatos que combine excelentes habilidades predictivas pero que nos brinde la posibilidad de realizar cálculos de manera analítica. El modelo de mezclas, ver por ejemplo McLachlan and Peel (2000) y Bouveyron et al. (2019), emerge como una elección destacada, cumpliendo con ambos requisitos. No obstante, desde una perspectiva Bayesiana, la implementación de Métodos Monte Carlo vía Cadenas de Markov (MCMC) para generar muestras de la posterior de todos los parámetros, ver por ejemplo Robert and Casella (1999), es esencial. En cambio, desde la visión frecuentista, simplemente requerimos un estimador inicial y definir martingalas que actualicen los parámetros. Al considerar ambas aproximaciones, nuestra preferencia recae en la vía frecuentista, donde el mayor desafío radica en obtener el estimador inicial. Este último se puede calcular de manera efectiva mediante el algoritmo Esperanza Maximización (EM), Dempster et al. (1977). A continuación, se presenta la descripción del modelo de mezclas, seguido por la estrategia predictiva vía martingalas frecuentista aplicada a este modelo.

El presente capítulo está basado en una versión preliminar de un artículo de mi supervisor de tesis junto con los doctores Stephen G. Walker y Ramsés H. Mena.

5.1. Modelos de mezclas de distribuciones

Los modelos de mezcla constituyen una clase versátil y potente de modelos estadísticos, distinguiéndose por su habilidad para capturar poblaciones heterogéneas al asumir que los datos se originan a partir de una combinación de varios subgrupos o componentes subyacentes. Además de su función en la modelación de la heterogeneidad poblacional, los modelos de mezcla son ampliamente utilizados en la estimación de densidades de probabilidad, permitiendo una representación más precisa de la distribución de los datos. Esta capacidad de modelar densidades resulta particularmente valiosa en aplicaciones predictivas, donde la flexibilidad inherente de los modelos de

mezcla se convierte en una herramienta eficaz para describir la complejidad de fenómenos subyacentes y mejorar la calidad de las predicciones. Esta es la capacidad de los modelos de mezclas que buscaremos explotar vía el enfoque predictivo.

El modelo de mezclas con k componentes, y asumimos que k es conocido, se expresa como

$$p(y|\mathbf{w}, \boldsymbol{\theta}) = \sum_{j=1}^k w_j f(y|\theta_j), \quad (5.1)$$

donde $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^k$, $\mathbf{w} = \{w_j\}_{j=1}^k$ y $f(y|\theta)$ es una distribución de probabilidad. Para que (5.1) sea una distribución, los pesos, \mathbf{w} , deben ser no negativos y sumar uno.

Dada una muestra aleatoria $y_{1:n}$ del modelo de mezcla (5.1), una idea útil es incluir variables de asignación latentes $z_{1:n}$, de modo que dada z_i , se conozca el componente del cual se ha extraído y_i , es decir,

$$p(y_i|\mathbf{w}, \boldsymbol{\theta}, z_i = j) = f(y_i|\theta_j).$$

Así, *a priori*, cada z_i se extrae de manera independiente de la distribución

$$p(z_i = j|\mathbf{w}) = w_j, \quad \text{para } j = 1, \dots, k.$$

Integrando la variable latente z_i de la distribución conjunta, se obtiene

$$\begin{aligned} \sum_{j=1}^k p(y_i, z_i = j|\mathbf{w}, \boldsymbol{\theta}) &= \sum_{j=1}^k p(z_i = j|\mathbf{w}, \boldsymbol{\theta}) p(y_i|\mathbf{w}, \boldsymbol{\theta}, z_i = j), \\ &= \sum_{j=1}^k p(z_i = j|\mathbf{w}) p(y_i|\mathbf{w}, \boldsymbol{\theta}, z_i = j), \\ &= p(y_i|\mathbf{w}, \boldsymbol{\theta}), \end{aligned}$$

el modelo original.

Al elegir distribuciones para modelos de mezcla, es crucial considerar aquellas que simplifiquen la estimación de los parámetros del modelo. Distribuciones como la normal y Poisson son preferidas por su simplicidad matemática. Optar por estas distribuciones simplifica la implementación de métodos computacionales que nos permitan, primero la estimación de los parámetros del modelo y a continuación el realizar predicciones de manera sencilla. En las siguientes secciones se describen las ideas predictivas usando los modelos de mezclas normal y Poisson.

5.2. Modelo de mezclas de normales univariadas

Se propondrá ahora un modelo de una mezcla de k variables normales univariadas, con k conocida, por lo que ahora tenemos:

$$f(y|\boldsymbol{\theta}) = \sum_{j=1}^k w_j \mathcal{N}(y|\mu_j, \sigma_j^2), \quad (5.2)$$

donde las $\mathcal{N}(y|\mu_j, \sigma_j^2)$ corresponden a densidades normales, por lo que ahora $\theta_i = (\mathbf{w}_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$, $\mathbf{w}_i = (w_{i,1}, \dots, w_{i,k})$, $\boldsymbol{\mu}_i = (\mu_{i,1}, \dots, \mu_{i,k})$ y $\boldsymbol{\sigma}_i^2 = (\sigma_{i,1}^2, \dots, \sigma_{i,k}^2)$. En este caso se hará uso del algoritmos EM (ver la Sección 5.4) para la obtención de $\hat{\mathbf{w}}_n$, $\hat{\boldsymbol{\mu}}_n$ y de $\hat{\boldsymbol{\sigma}}_n^2$, que a su vez emplea variables latentes $z_{1:n}$ en un modelo aumentado, lo que nos permitirá estimar el número de observaciones asociadas a cada componente de la mezcla $m_{n,j} = \sum_{i=1}^n \mathbb{1}\{z_i = j\}$ para $j \in \{1, \dots, k\}$. Así, la predicción estaría dada por

$$\begin{aligned} z_{i+1} &\sim p(z|\hat{\mathbf{w}}_i), \\ y_{i+1} &\sim \mathcal{N}(y|\hat{\mu}_{i,z_{i+1}}, \hat{\sigma}_{i,z_{i+1}}^2), \end{aligned}$$

mientras que la actualización de los parámetros para $i \in \{n, \dots, N-1\}$ se obtiene con:

$$\begin{aligned} \hat{w}_{i+1,j} &= \frac{1}{i+1} (i\hat{w}_{i,j} + \mathbb{1}\{z_{i+1} = j\}), \quad \text{para } j \in \{1, \dots, k\}, \\ \hat{\mu}_{i+1,j} &= \begin{cases} \frac{1}{m_{i,j} + 1} (m_{i,j}\hat{\mu}_{i,j} + y_{i+1}), & \text{si } j = z_{i+1}, \\ \hat{\mu}_{i,j}, & \text{si } j \neq z_{i+1}, \end{cases} \\ \hat{\sigma}_{i+1,j}^2 &= \begin{cases} \frac{1}{d_{i+1,j}} \left[d_{i,j}\hat{\sigma}_{i,j}^2 + \left(\frac{m_{i,j} + 1}{m_{i,j} + 2} \right) (\hat{\mu}_{i,j} - y_{i+1})^2 \right], & \text{si } j = z_{i+1}, \\ \hat{\sigma}_{i,j}^2, & \text{si } j \neq z_{i+1}, \end{cases} \end{aligned}$$

en donde el número de predicciones generadas por cada componente se actualiza mediante

$$m_{i+1,j} = m_{i,j} + \mathbb{1}\{z_{i+1} = j\}, \quad \text{para } j \in \{1, \dots, k\},$$

donde $d_{i,j} = \sum_{l=1}^{m_{i,j}} \frac{l}{l+1}$. Ahora comprobemos que las actualizaciones definan martingalas, primero para los pesos tenemos que:

$$\begin{aligned} \mathbb{E}(\hat{w}_{i+1}|z_{1:i}) &= \frac{1}{i+1} (i\mathbb{E}(\hat{w}_i|z_{1:i}) + \mathbb{E}(\mathbb{1}\{z_{i+1} = j\}|z_{1:i})), \\ &= \frac{1}{i+1} (i\hat{w}_{i,j} + P(z_{i+1} = j|z_{1:i})), \\ &= \frac{1}{i+1} (i\hat{w}_{i,j} + \hat{w}_{i,j}), \\ &= \hat{w}_{i,j}. \end{aligned}$$

Además de que se mantiene la condición de que los pesos sumen 1, en efecto:

$$\sum_{j=1}^k \hat{w}_{i+1,j} = \frac{1}{i+1} \left(i \sum_{j=1}^k \hat{w}_{i,j} + \sum_{j=1}^k \mathbb{1}\{z_{i+1} = j\} \right) = 1.$$

Para las medias la martingala se satisface trivialmente si $z_{i+1} \neq j$, por lo tanto si

$z_{i+1} = j$ tenemos:

$$\begin{aligned}\mathbb{E}(\hat{\mu}_{i+1,j}|y_{1:i}) &= \frac{1}{m_{i,j} + 1} (m_{i,j}\mathbb{E}(\hat{\mu}_{i,j}|y_{1:n}) + \mathbb{E}(y_{i+1}|y_{1:i})), \\ &= \frac{1}{m_{i,j} + 1} (m_{i,j}\hat{\mu}_{i,j} + \hat{\mu}_{i,j}), \\ &= \hat{\mu}_{i,j}.\end{aligned}$$

Para las varianzas nuevamente la martingala se satisface trivialmente si $z_{i+1} \neq j$, ahora si $z_{i+1} = j$ notemos que $m_{i,j} + 1 = m_{i+1,j}$ por lo tanto tenemos que:

$$\begin{aligned}\mathbb{E}(\hat{\sigma}_{i+1,j}^2|y_{1:i}) &= \frac{1}{d_{i+1,j}} \left[d_{i,j}\mathbb{E}(\hat{\sigma}_{i,j}^2|y_{1:i}) + \left(\frac{m_{i+1}}{m_{i+1,j} + 1} \right) \mathbb{E}((\hat{\mu}_{i,j} - y_{i+1})^2|y_{1:i}) \right], \\ &= \frac{1}{d_{i+1,j}} \left[d_{i,j}\hat{\sigma}_{i,j}^2 + \frac{m_{i+1,j}}{m_{i+1,j} + 1} \hat{\sigma}_{i,j}^2 \right], \\ &= \hat{\sigma}_{i,j}^2.\end{aligned}$$

5.3. Modelo de mezclas Poisson

Nuevamente, debido al contexto discreto propio de un conteo de votos procedemos a considerar una mezcla de distribuciones discretas, específicamente una mezcla de k distribuciones Poisson, con k fija y conocida, es decir:

$$f(y|\boldsymbol{\theta}) = \sum_{j=1}^k w_j \text{Pois}(y|\lambda_j), \quad (5.3)$$

donde las $\text{Pois}(y|\lambda_j)$ corresponde a distribuciones Poisson, teniendo ahora $\boldsymbol{\theta}_i = (\mathbf{w}_i, \boldsymbol{\lambda}_i)$, $\mathbf{w}_i = (w_{i,1}, \dots, w_{i,k})$ y $\boldsymbol{\lambda}_i = (\lambda_{i,1}, \dots, \lambda_{i,k})$. Con el algoritmo EM se obtienen $\hat{\mathbf{w}}_n$ y $\hat{\boldsymbol{\lambda}}_n$, así, las predicciones para $i \in \{n, \dots, N-1\}$ se obtienen con:

$$\begin{aligned}z_{i+1} &\sim p(z|\hat{\mathbf{w}}_i), \\ y_{i+1} &\sim \text{Pois}(y|\hat{\lambda}_{i,z_{i+1}}),\end{aligned}$$

mientras que la actualización de los parámetros se realiza vía

$$\begin{aligned}\hat{w}_{i+1,j} &= \frac{1}{i+1} (i\hat{w}_{i,j} + \mathbb{1}\{z_{i+1} = j\}), \text{ para } j \in \{1, \dots, k\}, \\ \hat{\lambda}_{i+1,j} &= \begin{cases} \frac{1}{m_{i,j} + 1} (m_{i,j}\hat{\lambda}_{i,j} + y_{i+1}), & \text{si } j = z_{i+1} \\ \hat{\lambda}_{i,j}, & \text{si } j \neq z_{i+1}, \end{cases}\end{aligned}$$

y el número de predicciones generado por cada componente

$$m_{i+1,j} = m_{i,j} + \mathbb{1}\{z_{i+1} = j\}, \text{ para } j \in \{1, \dots, k\}. \quad (5.4)$$

Mostremos que la actualización para $\hat{\lambda}_{i+1,j}$ define una martingala. Si $z_{i+1} \neq j$ se satisface la martingala trivialmente, así, supongamos que $z_{i+1} = j$, entonces:

$$\begin{aligned}\mathbb{E}(\hat{\lambda}_{i+1,j}|y_{1:i}) &= \frac{1}{m_{i,j} + 1} \left(m_{i,j} \mathbb{E}(\hat{\lambda}_{i,j}|y_{1:n}) + \mathbb{E}(y_{i+1}|y_{1:i}) \right), \\ &= \frac{1}{m_{i,j} + 1} \left(m_{i,j} \hat{\lambda}_{i,j} + \hat{\lambda}_{i,j} \right), \\ &= \hat{\lambda}_{i,j}.\end{aligned}$$

5.4. Algoritmo Esperanza-Maximización (EM)

En general, al estimar los parámetros de un modelo bajo el enfoque frecuentista, buscamos el estimador máximo verosímil. Es bien sabido que los estimadores obtenidos por método de máxima verosimilitud poseen propiedades como consistencia, invarianza, normalidad asintótica, entre otras (ver, por ejemplo, Wasserman (2003)). Para encontrar el estimador máximo verosímil, se busca el máximo global de la función de verosimilitud $L(\theta|y_{1:n}) = p(y_{1:n}|\theta)$, que es equivalente al máximo global de la log-verosimilitud (la función logaritmo es estrictamente creciente):

$$l(\theta|y_{1:n}) = \log(p(y_{1:n}|\theta)). \quad (5.5)$$

En general, lograr lo anterior de forma analítica solo es posible para modelos simples; en casos más complejos, se vuelve imprescindible el uso de métodos computacionales para aproximar el máximo global. Uno de estos métodos es el algoritmo de Esperanza-Maximización (EM) Dempster et al. (1977), que se presenta como una herramienta importante para la estimación de parámetros en modelos complejos.

El algoritmo EM es un procedimiento iterativo que, a partir de un valor inicial, busca máximos locales y se basa en la presencia de datos incompletos o latentes. En su primera fase, conocida como la fase de esperanza (E), el algoritmo calcula las esperanzas de los valores latentes condicionales dados los datos observados y las estimaciones actuales de los parámetros. Posteriormente, en la fase de maximización (M), ajusta los parámetros del modelo para maximizar la verosimilitud condicional utilizando los valores esperados calculados en la fase anterior. Estos pasos se repiten iterativamente hasta que los parámetros convergen a un máximo local de la log-verosimilitud del modelo.

Es importante resaltar dos de los comentarios realizados anteriormente:

- El algoritmo EM es especialmente útil cuando se pueden introducir datos faltantes en el problema. En el caso de los modelos de mezclas, estos datos faltantes están representados por las variables latentes $z_{1:n}$. Por esta razón, el algoritmo EM suele ser una opción popular para aproximar los estimadores de máxima verosimilitud del modelo de mezclas.
- El algoritmo EM encuentra máximos locales a partir de un valor inicial (véase, por ejemplo, Robert and Casella (1999)). Por lo tanto, el valor o semilla inicial es

de crucial importancia (ver por ejemplo Bouveyron et al. (2019)). En la Sección 5.4.5, describimos las estrategias que utilizamos para encontrar semillas iniciales para el algoritmo EM en el caso de un modelo de mezclas.

Finalmente, recordamos que en nuestro caso utilizaremos el algoritmo EM para obtener estimadores iniciales, es decir, \boldsymbol{w}_n y $\boldsymbol{\theta}_n$, y a partir de estos aplicaremos el enfoque predictivo mediante martingalas, como se describe en las secciones anteriores.

5.4.1. Teoría general

Como mencionamos el algoritmo EM encuentra máximos locales de la función (5.5), en caso en el que algunas variables no son observadas. La estrategia parte de una propuesta inicial y cada iteración consiste de la obtención de una función $Q(\theta, \theta_s)$ mediante la esperanza de la log-verosimilitud completa $l(\theta|y, z)$ bajo la condicional de las variables no observadas $p(z|\theta_s, y)$ dada las variables observadas y la estimación actual θ_s de θ , a dicha etapa se le denomina el paso E y el paso M consiste en la maximización de $Q(\theta, \theta_s)$.

Sea (Y, Z) nuestras variables completas en las que sólo Y es observada, sea $p(y, z|\theta)$ la densidad conjunta, si Z es discreta tenemos que la marginal de la variable observada es:

$$p(y|\theta) = \sum_{z \in \mathcal{Z}} p(y, z|\theta) = \sum_{z \in \mathcal{Z}} p(y|\theta, z)p(z|\theta),$$

y la condicional de la variable no observada es:

$$p(z|\theta, y) = \frac{p(y, z|\theta)}{p(y|\theta)}.$$

Dada una muestra $\{y_1, \dots, y_n\}$ podemos construir la log-verosimilitud completa asumiendo que las (y_i, z_i) son i.i.d. como:

$$l_C(\theta|y_{1:n}, z_{1:n}) = \sum_{i=1}^n \log(p(y_i, z_i|\theta)),$$

mientras que la log-verosimilitud observada corresponde a:

$$l_O(\theta|y_{1:n}) = \sum_{z_{1:n} \in \mathcal{Z}^n} l_C(\theta|y_{1:n}, z_{1:n}) = \sum_{i=1}^n \log(p(y_i|\theta)).$$

Definamos para $s \in \{1, \dots, T\}$, con T un índice de término:

$$Q(\theta, \theta_s) := \mathbb{E}_{p(z|\theta_s, y)}(l_C(\theta|Y, Z)) = \sum_{z \in \mathcal{Z}} l_C(\theta|y, z)p(z|\theta_s, y)dz,$$

como la esperanza condicional de la log-verosimilitud completa dada la variable observada y y la estimación actual θ_s , es decir $\theta_s = \arg \max_{\theta} Q(\theta, \theta_{s-1})$.

Proposición: Para $s \in \{1, \dots, T-1\}$ tenemos que:

$$\log(p(y|\theta_{s+1})) \geq \log(p(y|\theta_s)).$$

Demostración. Partiendo de la ecuación (4.16) tenemos que:

$$\log(p(z|\theta, y)) = \log(p(y, z|\theta)) - \log(p(y|\theta)),$$

multiplicando por $p(z|\theta_s, y)$ y sumando sobre el rango de Z tenemos:

$$\sum_{z \in \mathcal{Z}} p(z|\theta_s, y) \log(p(z|\theta, y)) = \sum_{z \in \mathcal{Z}} p(z|\theta_s, y) \log(p(y, z|\theta)) - \sum_{z \in \mathcal{Z}} p(z|\theta_s, y) \log(p(y|\theta)),$$

dado que $p(y|\theta)$ no depende de z , tenemos que:

$$\log(p(y|\theta)) = \sum_{z \in \mathcal{Z}} \log(p(y, z|\theta)) p(z|\theta_s, y) - \sum_{z \in \mathcal{Z}} \log(p(z|\theta, y)) p(z|\theta_s, y),$$

es decir:

$$\log(p(y|\theta)) = Q(\theta, \theta_s) - \sum_{z \in \mathcal{Z}} \log(p(z|\theta, y)) p(z|\theta_s, y),$$

ahora, dada la ecuación anterior tenemos que:

$$\log(p(y|\theta)) - \log(p(y|\theta_s)) = Q(\theta, \theta_s) - Q(\theta_s, \theta_s) - \sum_{z \in \mathcal{Z}} \log \left(\frac{p(z|\theta, y)}{p(z|\theta_s, y)} \right) p(z|\theta_s, y),$$

por la desigualdad de Jensen tenemos que:

$$\sum_{z \in \mathcal{Z}} \log \left(\frac{p(z|\theta, y)}{p(z|\theta_s, y)} \right) p(z|\theta_s, y) \leq \log \left(\sum_{z \in \mathcal{Z}} \frac{p(z|\theta, y)}{p(z|\theta_s, y)} p(z|\theta_s, y) \right) = \log(1) = 0,$$

por lo tanto:

$$\log(p(y|\theta)) - \log(p(y|\theta_s)) \geq Q(\theta, \theta_s) - Q(\theta_s, \theta_s),$$

si $\theta = \theta_{s+1}$, entonces:

$$\log(p(y|\theta_{s+1})) - \log(p(y|\theta_s)) \geq Q(\theta_{s+1}, \theta_s) - Q(\theta_s, \theta_s) \geq 0,$$

ya que $\theta_{s+1} = \arg \max_{\theta} Q(\theta, \theta_s)$. □

5.4.2. Algoritmo EM para mezclas

El modelo de mezclas con K componentes fijas dado por:

$$f(y|\mathbf{w}, \boldsymbol{\theta}) = \sum_{k=1}^K w_k f(y|\theta_k), \quad (5.6)$$

donde $\mathbf{w} = (w_1, \dots, w_K)$ son los pesos de cada componente de la mezcla, puede interpretarse con una perspectiva de variables latentes en la cual se introduce la variable Z , correspondiente a una multinomial de un sólo ensayo sobre los K grupos

con vector de probabilidades \mathbf{w} , es decir $Z \sim \text{Multinomial}(1, K, \mathbf{w})$. Así, para la pareja (y_i, z_i) tenemos que $z_i = (z_{i,1}, \dots, z_{i,K})$ donde:

$$z_{i,k} = \begin{cases} 1, & \text{si } y_i \text{ pertenece al grupo } k, \\ 0, & \text{en otro caso,} \end{cases} \quad (5.7)$$

y para y_i asumiremos que su distribución condicional dada z_i es:

$$f(y_i|\theta, z_i) = \prod_{k=1}^K f(y_i|\theta_k)^{z_{i,k}}, \quad (5.8)$$

por lo tanto la log-verosimilitud completa es:

$$l_C(\mathbf{w}, \boldsymbol{\theta}, z_{1:n}|y_{1:n}) = \sum_{i=1}^n \sum_{k=1}^K z_{i,k} \log(w_k f(y_i|\theta_k)). \quad (5.9)$$

Notemos que:

$$p(z|\mathbf{w}, \boldsymbol{\theta}, y) = \frac{p(y, z|\mathbf{w}, \boldsymbol{\theta})}{\sum_{k=1}^K w_k f(y|\theta_k)}, \quad (5.10)$$

así, tenemos que:

$$\begin{aligned} Q(\mathbf{w}, \mathbf{w}_s, \boldsymbol{\theta}, \boldsymbol{\theta}_s) &= \sum_{i=1}^n \sum_{k=1}^K \left(\sum_{z \in \mathcal{Z}} z_{i,k} p(z|\mathbf{w}^s, \boldsymbol{\theta}^s, y) \right) \log(w_k f(y_i|\theta_k)) \\ &= \sum_{i=1}^n \sum_{k=1}^K \tau_{i,k}^s (\log(w_k) + \log(f(y_i|\theta_k))), \end{aligned} \quad (5.11)$$

donde $\tau_{i,k}^s$ es la probabilidad condicional de que y_i provenga de la componente k bajo los parámetros del paso s , dada por:

$$\tau_{i,k}^s = \frac{w_k^s f(y_i|\theta_k^s)}{\sum_{l=1}^K w_l^s f(y_i|\theta_l^s)}. \quad (5.12)$$

Ahora podremos estimar los parámetros en el paso M, considerando que se tiene la restricción:

$$\sum_{k=1}^K w_k = 1, \quad (5.13)$$

procedemos con multiplicadores de Lagrange, así, para w_k tenemos:

$$\frac{\partial}{\partial w_k} (Q(\mathbf{w}, \mathbf{w}^s, \boldsymbol{\theta}, \boldsymbol{\theta}^s) - \lambda (\sum_{l=1}^K w_l - 1)) = \sum_{i=1}^n \frac{\tau_{i,k}^s}{w_k} - \lambda = 0,$$

para λ tenemos que:

$$\frac{\partial}{\partial \lambda} (Q(\mathbf{w}, \mathbf{w}^s, \boldsymbol{\theta}, \boldsymbol{\theta}^s) - \lambda (\sum_{l=1}^K w_l - 1)) = \sum_{l=1}^K w_l - 1 = \sum_{l=1}^K \frac{\sum_{i=1}^n \tau_{i,l}^s}{\lambda} - 1 = 0,$$

por lo tanto:

$$\lambda = \sum_{i=1}^n \sum_{k=1}^K \tau_{i,k}^s = n,$$

así,

$$w_k^{s+1} = \frac{\sum_{i=1}^n \tau_{i,k}^s}{n}. \quad (5.14)$$

Para los parámetros θ es preferible asignar un modelo concreto.

5.4.3. Mezclas de normales univariadas

Para el caso particular en el que tenemos desindades normales:

$$f(y|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2\sigma_k^2}(y-\mu_k)^2} \quad \text{para } k \in \{1, \dots, K\}, \quad (5.15)$$

tenemos que:

$$\frac{\partial}{\partial \mu_k} (Q(\mathbf{w}, \mathbf{w}^s, \boldsymbol{\theta}, \boldsymbol{\theta}^s) - \lambda(\sum_{l=1}^K w_l - 1)) = \sum_{i=1}^n \frac{\tau_{i,k}^s}{\sigma_k^2} (y_i - \mu_k) = 0,$$

por lo tanto:

$$\mu_k^{s+1} = \frac{\sum_{i=1}^n \tau_{i,k}^s y_i}{\sum_{i=1}^n \tau_{i,k}^s}, \quad (5.16)$$

para σ_k^2 tenemos que:

$$\frac{\partial}{\partial \sigma_k^2} (Q(\mathbf{w}, \mathbf{w}^s, \boldsymbol{\theta}, \boldsymbol{\theta}^s) - \lambda(\sum_{l=1}^K w_l - 1)) = \sum_{i=1}^n \left(-\frac{\tau_{i,k}^s}{2\sigma_k^2} + \frac{(y_i - \mu_k)^2}{2(\sigma_k^2)^2} \right) = 0,$$

por lo tanto:

$$(\sigma_k^2)^{s+1} = \frac{\sum_{i=1}^n \tau_{i,k}^s (y_i - \mu_k^{s+1})^2}{\sum_{i=1}^n \tau_{i,k}^s}. \quad (5.17)$$

Para una mezcla de distribuciones Poisson se obtiene una versión simplificada del algoritmo descrito para la mezcla de distribuciones normales, por lo que se omite su descripción.

5.4.4. Mezclas de distribuciones Poisson

Si ahora tenemos densidades Poisson:

$$f(y|\lambda_k) = \frac{\lambda_k^y}{y!} e^{-\lambda_k} \quad \text{para } k \in \{1, \dots, K\}, \quad (5.18)$$

entonces:

$$\frac{\partial}{\partial \lambda_k} (Q(\mathbf{w}, \mathbf{w}^s, \boldsymbol{\theta}, \boldsymbol{\theta}^s) - \lambda(\sum_{l=1}^K w_l - 1)) = \sum_{i=1}^n \left(\tau_{i,k}^s \frac{y_i}{\lambda_k} - 1 \right) = 0,$$

por lo que:

$$\lambda_k^{s+1} = \frac{\sum_{i=1}^n \tau_{i,k}^s y_i}{\sum_{i=1}^n \tau_{i,k}^s}. \quad (5.19)$$

5.4.5. Consideraciones adicionales

Debido a que la verosimilitud de un modelo de mezcla en general no es una función convexa pueden existir máximos locales por lo que las soluciones del algoritmo EM pueden depender fuertemente de los valores iniciales propuestos, ver por ejemplo Bouveyron et al. (2019). Para una correcta inicialización del algoritmo EM se propone realizar una primera agrupación de los datos para así, realizar estimaciones de los valores iniciales, para lograr aquello existen por un lado métodos que no requieren un modelo a priori para agrupar los datos, los cuales se denominan métodos aglomerativos de agrupación jerárquica, por otro lado tenemos a aquellos que si hacen uso de un modelo para definir el criterio de aglomeración, los cuales se denominan métodos de agrupación jerárquica basada en modelos.

Métodos aglomerativos de agrupación jerárquica.

Estos métodos parten de la partición trivial en la que cada elemento es un grupo en sí mismo para posteriormente ir comparando similitudes (o diferencias) conforme a un criterio específico para combinar grupos sucesivamente. La implementación usual de estos métodos se describe en el siguiente algoritmo:

Algoritmo 3: Agrupación jerárquica aglomerativa.

Input: Vectores $\{\mathbf{x}_i\}_{i=1}^n$, criterio de distancia entre grupos $\mathcal{D}(\mathcal{G}, \mathcal{G}')$, $\mathcal{A} \leftarrow \emptyset$;

- 1 **for** $i \leftarrow 1$ **to** N **do**
- 2 $\mathcal{A} \leftarrow \mathcal{A} \cup \{x_i\}$;
- 3 $\mathcal{T} \leftarrow \mathcal{A}$;
- 4 **while** $|\mathcal{A}| > 1$ **do**
- 5 $\mathcal{G}_1^*, \mathcal{G}_2^* \leftarrow \arg \max_{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{A}} \mathcal{D}(\mathcal{G}_1, \mathcal{G}_2)$;
- 6 $\mathcal{A} \leftarrow (\mathcal{A} - \{\mathcal{G}_1^*\}) - \{\mathcal{G}_2^*\}$;
- 7 $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathcal{G}_1^* \cup \mathcal{G}_2^*\}$;
- 8 $\mathcal{T} \leftarrow \mathcal{T} \cup \{\mathcal{G}_1^* \cup \mathcal{G}_2^*\}$;

Output: \mathcal{T} ;

Donde \mathcal{T} es un árbol cuyas hojas son los singletones $\{\mathbf{x}_i\}$ para $i \in \{1, \dots, n\}$, su raíz es el conjunto formado por todas las observaciones $\{x_i\}_{i=1}^n$ y sus nodos intermedios son grupos intermedios. Se parte de seleccionar una distancia apropiada entre las observaciones, $d(\mathbf{x}_i, \mathbf{x}_j)$, para así poder definir $\mathcal{D}(\mathcal{G}_1, \mathcal{G}_2)$ para dos grupos $\mathcal{G}_1 = \{\mathbf{x}_i\}_{i=1}^p$ y $\mathcal{G}_2 = \{\mathbf{y}_j\}_{j=1}^q$ según el criterio a considerar.

Criterio de enlace-simple: Sean $\mathcal{G}_1, \mathcal{G}_2$ dos grupos, entonces:

$$\mathcal{D}_{S-L}(\mathcal{G}_1, \mathcal{G}_2) = \min_{\mathbf{x} \in \mathcal{G}_1, \mathbf{y} \in \mathcal{G}_2} d(\mathbf{x}, \mathbf{y}). \quad (5.20)$$

Criterio de enlace-completo: Sean $\mathcal{G}_1, \mathcal{G}_2$ dos grupos, entonces:

$$\mathcal{D}_{C-L}(\mathcal{G}_1, \mathcal{G}_2) = \max_{\mathbf{x} \in \mathcal{G}_1, \mathbf{y} \in \mathcal{G}_2} d(\mathbf{x}, \mathbf{y}). \quad (5.21)$$

Criterio de enlace-medio: Sean $\mathcal{G}_1, \mathcal{G}_2$ dos grupos, entonces:

$$\mathcal{D}_{A-L}(\mathcal{G}_1, \mathcal{G}_2) = \frac{1}{|\mathcal{G}_1||\mathcal{G}_2|} \sum_{\mathbf{x} \in \mathcal{G}_1} \sum_{\mathbf{y} \in \mathcal{G}_2} d(\mathbf{x}, \mathbf{y}). \quad (5.22)$$

Criterio de centroide: Sean $\mathcal{G}_1, \mathcal{G}_2$ dos grupos, entonces:

$$\mathcal{D}_C(\mathcal{G}_1, \mathcal{G}_2) = d\left(\frac{1}{|\mathcal{G}_1|} \sum_{\mathbf{x} \in \mathcal{G}_1} \mathbf{x}, \frac{1}{|\mathcal{G}_2|} \sum_{\mathbf{y} \in \mathcal{G}_2} \mathbf{y}\right). \quad (5.23)$$

Métodos de agrupación jerárquica basada en modelos.

Ahora el criterio para mezclar grupos es la maximización de la *verosimilitud de clasificación*:

$$\mathcal{L}_{CL}(\theta, z_{1:n}|y_{1:n}) = \prod_{i=1}^n z_i f(y_i|\theta_{z_i}), \quad (5.24)$$

al mezclar sobre todas las posibles parejas de grupos. Si no se cuenta con información respecto al número y tamaño de los grupos, análogamente a los métodos aglomerativos se comienza tratando a cada observación como un singletón. La mayoría de estos métodos realiza conjuntamente estimaciones de los parámetros asociados a los grupos correspondientes para cada iteración partiendo de alguna propuesta inicial, la cual puede obtenerse con una partición aleatoria de los datos o con una muestra aleatoria de parámetros Bouveyron et al. (2019).

Capítulo 6

Comparaciones

En los capítulos anteriores se describieron tres perspectivas para abordar la estimación de medias y porcentajes bajo poblaciones finitas, comenzado con el enfoque frecuentista y el estimador de razón combinado, continuando con la inferencia Bayesiana y finalizando con una visión moderna centrada en distribuciones predictivas e imputación vía martingalas. En este capítulo, el objetivo será comparar el desempeño de varias estrategias desarrolladas mediante cada perspectiva para la estimación de los porcentajes, en particular buscaremos estimar los porcentajes de votos en favor de cada uno de los candidatos a un puesto de elección popular. De manera concisa, supongamos que se tienen N casillas electorales instaladas y J candidatos que compiten por un puesto de elección popular. Entonces, con la información de la votación en cada una de las casillas que conforman la población se calcula como

$$p_{N,j} = 100 \frac{\sum_{i=1}^N y_{i,j}}{\sum_{j=1}^J \sum_{i=1}^N y_{i,j}} = 100 \frac{N \bar{y}_{N,j}}{N \sum_{j=1}^J \bar{y}_{N,j}}, \quad (6.1)$$

para $j = 1, 2, \dots, J$. Así, $y_{i,j}$ es el número de votos en la casilla i -ésima en favor del candidato j y por lo tanto $p_{N,j}$ es el porcentaje poblacional de votos en favor del candidato j .

6.1. Estrategias a comparar

Asumiremos que se obtienen muestras de n casillas, i.e. $y_{i,j}$, para $i = 1, 2, \dots, n$ y $j = 1, \dots, J$. Las muestras se extraerán usando un MASSR y un MAE. Para describir de manera sencilla los estimadores se asume que la muestra se seleccionó por MASSR, pero los casos para el muestreo estratificado pueden consultarse en los capítulos correspondientes. Las estrategias de estimación con las que buscaremos aproximar los porcentajes (6.1) son:

1. **Estimador de razón combinado (RC).** Éste es el estimador del porcentaje de votos en favor del candidato j obtenido mediante muestreo probabilístico y descrito en el Capítulo 2. El estimador está dado por

$$p_{n,j} = 100 \frac{t_{n,j}}{\hat{t}_n}, \quad (6.2)$$

en donde el total de votos en favor del candidato j y el total de votos en favor de todos los candidatos se estiman mediante

$$t_{n,j} = N\bar{y}_{n,j} \quad \text{y} \quad t_n = \sum_{j=1}^J t_{n,j},$$

con $\bar{y}_{n,j}$ el promedio muestral de votos en favor del candidato j . La varianza de $p_{n,j}$ se aproxima mediante

$$v(p_{n,j}) \approx \frac{100^2 N^2}{t_n^2} \left[\left(1 - \frac{n}{N}\right) \frac{s_{d_j}^2}{n} \right], \quad (6.3)$$

en donde $s_{d_j}^2 = \frac{1}{n-1} \sum_{i=1}^n (d_{i,j} - \bar{d}_j)^2$, $d_{i,j} = y_{i,j} - p_{n,j}t_i$ y $t_i = \sum_{j=1}^J y_{i,j}$. En este caso se asume que normalidad asintótica del estimador para calcular los intervalos de confianza.

2. **Estimador Bayesiano vía el modelo normal (*NormB*)**. Esta estrategia se describió en el Capítulo 3 y en particular en la Sección 3.4.1. Asumiendo que los votos de la muestra en favor del candidato j siguen una distribución normal y asumiendo una prior no informativa, se tiene que

$$p(\bar{y}_{N,j} | y_{1:n}) = t_{n-1} \left(\bar{y}_{n,j}, \left(1 - \frac{n}{N}\right) \frac{s_j^2}{n} \right). \quad (6.4)$$

Para realizar la estimación simplemente se generan digamos $m = 10,000$ muestras de (6.4), para $j = 1, 2, \dots, J$. Obteniendo una matriz de $m \times J$ de predicciones para $\bar{y}_{N,j}$, con $j = 1, \dots, J$. Con cada renglón de esta matriz se pueden estimar las proporciones (6.1), obteniendo así una matriz de $m \times J$ con estimaciones $\hat{p}_{N,j}$, para $j = 1, \dots, J$ en cada renglón. Para obtener el intervalo creíble para $p_{N,j}$, simplemente se toman los cuantiles por columna de esta matriz.

3. **Estimador Bayesiano vía el modelo Poisson (*PoisB*)**. Esta estrategia se describió en el Capítulo 3 y en particular en la Sección 3.4.2. En este caso se asume que la muestra del número de votos en favor del candidato j sigue una distribución Poisson y se asume una distribución Gamma. Entonces, para generar muestras aleatorias de la predictiva $p(\bar{y}_{N,j} | y_{1:n,j})$, primero generamos predicciones para la parte no observada del total de votos en favor del candidato j , denotada como $t_{N-n,j}$, i.e.

$$p(t_{N-n,j} | y_{1:n,j}) = \mathcal{NB} \left(t_{N-n,j} \mid \frac{n + \beta}{n + \beta + N}, \alpha + t_{n,j} \right). \quad (6.5)$$

Posteriormente, se aplica la transformación

$$\bar{y}_{N,j} = \left(\frac{n}{N}\right) \bar{y}_{n,j} + \left(\frac{1}{N}\right) t_{N-n,j}.$$

Existen algunas propuestas bien establecidas de distribuciones Gamma no informativas, entre las cuales se encuentra $\text{Gamma}(\epsilon, \epsilon)$ con $\epsilon \approx 0$, dicha propuesta es la que se tomó en cuenta con $\epsilon = 0.1$. Para obtener intervalos creíbles para $p_{N,j}$ se sigue una estrategia similar al caso anterior.

4. **Estimador predictivo asumiendo una mezcla de distribuciones normales (*MixNormP*)**. Esta estrategia se describió en el capítulo 5 y en particular en la sección 5.2. Se asume ahora que los votos en favor del candidato j son distribuidos por una mezcla de k distribuciones normales, la k se asume fija y conocida. Se imputan los votos en favor de cada candidato para las N casillas, i.e. se cuenta con información del número de votos en favor de cada candidato en la muestra $y_{i:n,j}$, para $j = 1, \dots, J$. Con base en una estrategia predictiva basada en un modelo de mezclas, se imputa la parte no observada $\{y_{n+1,N,j}\}$, para $j = 1, \dots, J$. Con los votos en cada una de las N casillas es trivial estimar $\hat{p}_{N,j}$. Para obtener los intervalos creíbles para los $p_{N,j}$, bastará repetir todo el procedimiento m veces, resultando en la matriz descrita para la estrategia Bayesiana vía el modelo normal y nuevamente tomando los cuantiles apropiados.
5. **Estimador predictivo asumiendo una mezcla de distribuciones Poisson (*MixPoisP*)**. Esta estrategia se describió en el capítulo 5 y particularmente en la sección 5.3. Ahora los votos obtenidos por el candidato j se suponen distribuidos por una mezcla de k distribuciones Poisson y nuevamente se imputa la población completa. Se sigue un procedimiento completamente análogo a la estrategia anterior.

6.2. Evaluación de las estrategias de estimación vía simulación

Para realizar una evaluación empírica de las estrategias de estimación discutidas en la sección anterior, se utilizaron dos conjuntos de datos reales:

- Los cálculos finales para la Consulta Popular de 2021, con $N = 57,077$ casillas. Este ejercicio de participación ciudadana, muchos de los retos para llevarlo a la práctica e incluso análisis descriptivos exploratorios fueron descritos en la introducción de esta tesis.
- Los cálculos finales para la elección presidencial de 2018, con $N = 156,840$ casillas. Este conjunto de datos simplemente es usado como referencia para tener un punto de comparación contra la Consulta Popular de 2021.

Para evaluar las estrategias, se seleccionaron 200 muestras vía MASSR de tamaño n del primer conjunto de datos, y se aplicaron todas las estrategias para obtener estimaciones puntuales y de intervalo del 95% para las proporciones de la población (6.1). Dado que la población completa está disponible, se conocen las verdaderas proporciones de la población ($p_{N,j}$ para $j = 1, \dots, J$), lo que permite calcular el error

porcentual medio (EPM) y los porcentajes reales de cobertura logrados por cada método. El EPM se calcula utilizando la siguiente fórmula:

$$\text{EPM}(\hat{p}_{N,j}) = \frac{1}{2} \sum_{l=1}^{200} \frac{|\hat{p}_{N,j}^l - p_{N,j}|}{p_{N,j}}, \quad (6.6)$$

donde $\hat{p}_{N,j}^l$ para $j = 1, \dots, J$ son las estimaciones puntuales obtenidas considerando la muestra l , para $l = 1, \dots, 200$. Los porcentajes reales de cobertura (PRC) se aproximan por

$$\text{PRC} = \frac{1}{2} \sum_{l=1}^{200} \mathbf{1}(p_{N,j} \in (\hat{p}_{L,j}^l, \hat{p}_{U,j}^l)), \quad (6.7)$$

donde $(\hat{p}_{L,j}^l, \hat{p}_{U,j}^l)$ es la estimación del intervalo del 95% para la proporción de la población $p_{N,j}$ utilizando la muestra l , para $l = 1, \dots, 200$.

Para identificar la sensibilidad de las distintas estrategias se utilizaron distintos tamaños de muestra n . Posteriormente se hizo el mismo ejercicio pero en esta ocasión las muestras se seleccionaron vía MAE.

Se repitió el ejercicio anterior utilizando el segundo conjunto de datos.

6.3. Consulta Popular 2021

De las 57,077 actas computadas para la Consulta Popular 2021, se filtró una base de datos constituida por 56,865 registros. Del análisis descriptivo se detectó que 212 mesas tenían una participación inusualmente alta, superior al 39.96%, cifra mayor al 26.1% resultado del criterio usual para detectar valores atípicos extremos, y se determinó que debían ser errores pues se confundieron los votos nulos con las boletas no utilizadas así que se eliminaron del análisis.

En este caso cada una de los 56,865 registros cuenta con información del número de votos en favor de la opción *SI*, el número de votos para la opción *NO*, el número de votos nulos *NULOS* y el identificador del distrito federal electoral al que pertenece la casilla, dichos distritos serán los estratos a tomar en cuenta en un MAE.

6.3.1. Resultados: MASSR

Las 200 muestras de tamaño n , se seleccionaron con $n \in \{1, 200; 3, 000; 6, 000\}$. En la Tabla 6.1 se presenta el PRC de los intervalos del 95% de confianza (95 de probabilidad) para todas las estrategias a comparar.

Es evidente que el modelo normal Bayesiano presenta un gran desempeño en la cobertura real para los tres tamaños de muestra, mejorando los resultados del tradicional estimador de razón combinado y no implicando un incremento sustancial en la complejidad de su implementación ni en el tiempo de cómputo. Los modelos de mezclas de distribuciones normales tienen igualmente un mejor desempeño que el estimador de razón combinado, habiendo una ligera mejora en el tamaño de muestra $n = 1, 200$ para $k = 3$ respecto a $k = 2$, sin embargo, es digno de mencionar que

n	1, 200				3, 000				6, 000			
Propuesta	SI	NO	NULOS	Media	SI	NO	NULOS	Media	SI	NO	NULOS	Media
R.C.	93	95	81.5	89.8	88.5	92	86.5	89	93.5	93.5	91.5	92.8
PoisB	56.2	72	42	56.7	48.5	62.5	36.5	49.2	47	66.5	33	48.8
NormB	93	97.5	83.5	91.3	90.5	96	87	91.2	93.5	97.5	92	94.3
MixPoisP ($k = 3$)	21.5	23	23	22.3	30.5	30	30.5	30.3	40.5	47	38	41.8
MixPoisP ($k = 2$)	15	15.5	8.5	13	18.5	20	13	17.2	26	36	19.5	27.2
MixNormP ($k = 3$)	94.5	98	82.5	91.7	92.5	95.5	87.5	91.8	92	97	90.5	93.2
MixNormP ($k = 2$)	93	98.5	84.5	92	92	95.5	87	91.5	94	95.5	91	93.5

Tabla 6.1: PRC para la Consulta Popular calculado con 200 muestras seleccionadas vía MASSR.

la implementación de dichos modelos sí incrementa la complejidad y el tiempo de cómputo significativamente. Por último, queda patente el pobre desempeño de los modelos basados en la distribución de Poisson a pesar de estar en un contexto discreto.

En la Tabla 6.2 encontramos los EPM obtenidos para cada estrategia.

n	1, 200				3, 000				6, 000			
Propuesta	SI	NO	NULOS	Media	SI	NO	NULOS	Media	SI	NO	NULOS	Media
R.C.	0.085	3.155	9.888	4.376	0.060	2.208	6.945	3.071	0.042	1.384	4.994	2.140
PoisB	0.085	3.165	9.886	4.379	0.060	2.210	6.954	3.075	0.042	1.387	4.987	2.139
NormB	0.085	3.157	9.914	4.385	0.060	2.222	6.921	3.068	0.042	1.380	4.991	2.138
MixPoisP ($k = 3$)	0.085	3.148	9.910	4.381	0.060	2.214	6.921	3.065	0.042	1.374	4.997	2.139
MixPoisP ($k = 2$)	0.085	3.154	9.873	4.371	0.060	2.205	6.925	3.063	0.042	1.387	4.995	2.141
MixNormP ($k = 3$)	0.087	3.155	10.025	4.422	0.061	2.195	6.997	3.084	0.042	1.417	5.010	2.156
MixNormP ($k = 2$)	0.085	3.132	10.027	4.415	0.060	2.231	6.901	3.064	0.042	1.395	4.987	2.141

Tabla 6.2: EPM para la Consulta Popular calculado con 200 muestras seleccionadas vía MASSR.

En el caso de los EPM, los resultados no son tan drásticos y directos, podemos observar que el estimador de razón combinado tiene un buen desempeño en los tres tamaños de muestra, sorpresivamente ahora el modelo Poisson Bayesiano es quien lo

sigue muy de cerca al igual que el modelo normal Bayesiano. En cuanto a los modelos de mezclas se tiene que los errores correspondientes a mezclas normales son mayores que los correspondientes a mezclas Poisson. Puntualizamos además que los votos nulos son aquellos en los que todos los modelos tienen el peor desempeño.

6.3.2. Resultados: MAE

Para el ejercicio usando muestras seleccionadas vía MAE es importante mencionar que México se divide en 300 distritos federales electorales, por lo que para alcanzar tamaños de muestra globales de $n \in \{1, 200; 3, 000; 6, 000\}$ se decidió simplemente dividir el tamaño global entre el número de estratos, $n_h = n/300$. De esta manera, los tamaños por estrato estarían dados por $n_h \in \{4, 10, 20\}$, con $h = 1, \dots, 300$, para cada tamaño de muestra global. Es importante mencionar que esto representará un reto para las estrategias basadas en modelo ya que cada modelo se deberá ajustar en cada estrato, con el tamaño de muestra n_h , de manera independiente. La Tabla 6.3 muestra el PCR de los modelos propuestos ahora implementados en un MAE.

n	1, 200				3, 000				6, 000			
	SI	NO	NULOS	Media	SI	NO	NULOS	Media	SI	NO	NULOS	Media
R.C.	88	95	76.5	86.5	96	95	87	92.7	93	94	92.5	93.2
PoisB	52	74.5	40	55.5	51	77.5	32	53.5	48	75.5	29	50.8
NormB	98.5	100	90.5	96.3	98	98	90	95.3	94.5	95.5	93.5	94.5
MixPoisP ($k = 3$)	18	19	17	18	29.5	41.5	25	32	43.5	49.5	40	44.3
MixPoisP ($k = 2$)	11.5	17	9.5	12.7	16	33	14	21	22.5	35.5	13	23.7
MixNormP ($k = 3$)	85	90	74.5	83.2	95.5	97	85.5	92.7	91.5	94	93.5	93
MixNormP ($k = 2$)	83	90.5	74.5	82.7	94.5	95.5	86	92	92	95	92.5	93.2

Tabla 6.3: PRC para la Consulta Popular calculado con 200 muestras seleccionadas vía MAE.

El modelo normal Bayesiano alcanza el mejor PCR, inclusive mejorando ligeramente su desempeño respecto a la implementación en el MASSR para todos los tamaños de muestra, el estimador de razón combinado igualmente muestra una ligera mejora para $n = 3, 000$ y $n = 6, 000$, mientras que para $n = 1, 200$ se reduce la cobertura real, un comportamiento similar se tiene para el resto de modelos, salvo la mezcla Poisson con $k = 2$ en la cual la cobertura se redujo o se mantuvo. La Tabla 6.4 presenta los EPM correspondientes.

Resulta llamativo que todos los modelos mostraron un incremento de los EP obtenidos bajo un MAE respecto a los EP obtenidos bajo un MASSR para muestras con

n	1, 200				3, 000				6, 000			
Propuesta	SI	NO	NULOS	Media	SI	NO	NULOS	Media	SI	NO	NULOS	Media
$\mathcal{R.C.}$	0.093	2.968	11.392	4.818	0.055	1.603	6.771	2.810	0.039	1.236	4.906	2.060
PoisB	0.098	3.030	11.361	4.830	0.057	1.673	6.836	2.855	0.040	1.216	4.948	2.068
NormB	0.093	2.978	11.444	4.838	0.055	1.604	6.757	2.805	0.040	1.243	4.905	2.063
MixPoisP ($k = 3$)	0.093	2.970	11.390	4.818	0.055	1.592	6.752	2.800	0.039	1.244	4.897	2.060
MixPoisP ($k = 2$)	0.093	2.965	11.396	4.818	0.055	1.603	6.778	2.812	0.039	1.237	4.909	2.062
MixNormP ($k = 3$)	0.094	2.979	11.553	4.875	0.055	1.619	6.842	2.839	0.039	1.231	4.950	2.073
MixNormP ($k = 2$)	0.095	3.014	11.459	4.856	0.054	1.624	6.669	2.782	0.039	1.239	4.901	2.060

Tabla 6.4: EPM para la Consulta Popular calculado con 200 muestras seleccionadas vía MAE.

$n = 1, 200$, mientras que para los otros dos tamaños de muestras los EP se redujeron para el MAE respecto al MASSR.

6.4. Elección Presidencial de 2018

De las 156, 840 actas computadas para la Elección Presidencial del 2018 se tienen registros faltantes en 209 de ellas, por lo que finalmente la base de datos empleada para las comparativas consta de 156, 631 casillas con el número de votos en favor de Andrés Manuel López Obrador (AMLO), el número de votos para Ricardo Anaya Cortés (RAC), el número de votos para José Antonio Meade Kuribreña (JAMK), el número de votos para Jaime Heliodoro Rodríguez Calderón (JHRC), una columna denominada OTROS que condensa los votos nulos, los votos para candidatos no registrados y los votos de la candidata independiente Margarita Zavala que canceló su registro, así como el identificador del distrito federal electoral al que pertenece la casilla.

Para la comparación de estrategias sólo se muestran los resultados obtenidos para los candidatos con mayor número de votos a favor, i.e. AMLO, RAC y JAMK, aunque también se cuenta con las estimaciones para JHRC y OTROS. Nuevamente se seleccionan 200 muestras de tamaño n , para $n \in \{1, 200, 3, 000, 6, 000\}$.

6.4.1. Resultados: MASSR

En la Tabla 6.5 se presenta la cobertura real de intervalos del 95% de confianza (0.95 de probabilidad) para los tres candidatos punteros de las alternativas comparadas en un MASSR.

n	1, 200				3, 000				6, 000			
Propuesta	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media
R.C.	96.5	96.5	97.5	96.8	95.5	96.5	96.5	96.2	95	94	93	94
PoisB	23.5	21.5	30.5	25.2	27	23.5	32	27.5	21	19	27.5	22.5
NormB	92.5	95.5	98	95.3	92.5	95	95.5	94.3	90.5	92	93	91.8
MixPoisP ($k = 3$)	1.5	1.5	2	1.7	1.5	2.5	5.5	3.2	4	4.5	4.5	4.3
MixPoisP ($k = 2$)	2	1.5	3	2.2	0.5	2.5	6.5	3.2	4.5	4.5	4.5	4.5
MixNormP ($k = 3$)	92	97	96.5	95.2	90.5	94	95.5	93.3	92	92.5	93	92.5
MixNormP ($k = 2$)	92.5	96.5	97.5	95.5	93	94	97	94.7	90	90.5	93	91.2

Tabla 6.5: PRC para la Elección Presidencial de 2018 calculado con 200 muestras seleccionadas vía MASSR.

Salta a la vista que el estimador de razón combinado tiene un desempeño excelente para todos los tamaños de muestras, siendo inclusive ligeramente mejor su desempeño para $n = 1, 200$ y más aún, el PRC es bueno o muy bueno para todos los candidatos, inclusive para aquellos que obtuvieron pocos votos. El modelo Bayesiano basado en la distribución Poisson vuelve a tener un desempeño muy pobre, sin embargo es digno de mencionar que para esta base de datos en las cuales no se presentaron registros atípicos considerables, este modelo tuvo su mejor desempeño en los candidatos que obtuvieron pocos votos y se aprecia una mejora para muestras más grandes. Los modelos de mezclas de distribuciones Poisson nuevamente tuvieron el peor desempeño, al igual que el modelo Bayesiano Poisson el mejor PRC lo tuvieron los candidatos con pocos votos, en estos modelos se aprecia una mejora sostenida en los PRC conforme incrementa el tamaño de la muestra y no hay diferencia significativa entre usar 2 o 3 componentes. El modelo normal Bayesiano presenta un desempeño ligeramente menor que el estimador de razón combinado, sin embargo para los primeros dos tamaños de muestra los PRC son superiores al 95%, teniendo su peor PRC para el candidato puntero en los tres tamaños de muestra. Para los modelos de mezclas normales tenemos un PRC medio muy cercano al 95% en los tres tamaños de muestra si se usan 3 componentes y curiosamente en el caso de usar 2 componentes el PRC medio disminuye a 91.2% para muestras con $n = 6, 000$.

La Tabla 6.6 muestra los EPM obtenidos para los tres candidatos punteros.

A diferencia de los EPM obtenidos para la Consulta Popular 2021 ahora los EPM son pequeños, no superando el 1.5% en registros particulares y no superando el 1.2% la media de los registros. Sin embargo, las diferencias de los EPM obtenidos para cada propuesta vuelve a ser marginal, teniendo que los modelos de mezclas de distribuciones normales son los que tienen los EPM mayores.

n	1, 200				3, 000				6, 000			
Propuesta	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media
R.C.	0.763	1.421	1.332	1.172	0.456	0.911	0.870	0.746	0.349	0.706	0.644	0.566
PoisB	0.763	1.421	1.331	1.172	0.457	0.910	0.870	0.746	0.349	0.706	0.645	0.567
NormB	0.763	1.415	1.332	1.170	0.457	0.907	0.869	0.744	0.348	0.702	0.643	0.564
MixPoisP ($k = 3$)	0.763	1.421	1.333	1.172	0.456	0.911	0.870	0.746	0.349	0.706	0.645	0.567
MixPoisP ($k = 2$)	0.763	1.420	1.332	1.172	0.456	0.911	0.870	0.746	0.349	0.706	0.644	0.566
MixNormP ($k = 3$)	0.760	1.425	1.353	1.179	0.458	0.904	0.883	0.748	0.347	0.476	0.486	0.436
MixNormP ($k = 2$)	0.782	1.409	1.362	1.184	0.459	0.928	0.883	0.757	0.355	0.714	0.651	0.573

Tabla 6.6: EPM para las Elecciones Federales de 2018 calculado con 200 muestras seleccionadas vía MASSR.

6.4.2. Resultados: MAE

En la Tabla 6.7 se presenta los PRC de intervalos del 95% de confianza (95 de probabilidad) para los tres candidatos punteros de las alternativas comparadas en un MAE.

n	1, 200				3, 000				6, 000			
Propuesta	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media
R.C.	94.5	95.5	97.5	95.8	96	96	95.5	95.8	96.5	95	95	95.5
PoisB	34.5	27	30.5	30.7	33.5	37	41	37.2	33.5	30	40	34.5
NormB	100	99	100	99.7	98	98	98.5	98.2	97	96.5	96	96.5
MixPoisP ($k = 3$)	4	3	4	3.7	2.5	5	9.5	5.7	10	5.5	12	9.2
MixPoisP ($k = 2$)	4	3.5	4	3.8	2.5	5.5	8	5.3	9.5	5	12.5	9
MixNormP ($k = 3$)	93.5	95	93.5	94	96.5	96	94	95.5	95	95	95	95
MixNormP ($k = 2$)	89	93.5	89	90.5	94.5	94.5	94	94.3	96.5	95.5	96.5	97.7

Tabla 6.7: PRC para la Elección Presidencial de 2018 calculado con 200 muestras seleccionadas vía MAE.

Nuevamente se tiene un buen desempeño del estimador de razón combinado en los tres tamaños de muestra. El modelo Bayesiano de distribución Poisson tiene una

mejora significativa en los PRC obtenidos bajo un MAE respecto a los obtenidos bajo un MASSR, manteniendo que su mejor cobertura se logra en candidatos con pocos votos. Igualmente los modelos de mezcla de distribuciones Poisson presentan una ligera mejora respecto a los resultados para un MASSR y manteniendo el mejor desempeño en candidatos con pocos votos y una mejora sostenida conforme incrementa el tamaño de las muestras. El modelo Bayesiano basado en la distribución normal es el que mejor se desempeñó, teniendo PRC 100% para casi todos los candidatos en muestras con $n = 1,200$ y para los otros tamaños de muestras los PRC fueron superiores al 95%. Para las mezclas de distribuciones normales tenemos que los PRC para $n = 1,200$ son menores al 95% y que reduce más en el caso de usar 2 componentes, ya en los otros dos tamaños de muestra tenemos buenos PRC para ambos modelos e inclusive para $n = 6,000$ ambos presentan PCR muy cercanos.

La Tabla 6.8 contiene los EPM para los tres candidatos punteros obtenidos en un MAE.

n	1, 200				3, 000				6, 000			
Propuesta	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media	AMLO	RAC	JAMK	Media
R.C.	0.538	1.084	1.172	0.931	0.318	0.623	0.635	0.525	0.231	0.470	0.477	0.393
PoisB	0.540	1.082	1.169	0.930	0.318	0.624	0.635	0.526	0.232	0.471	0.477	0.393
NormB	0.540	1.078	1.181	0.933	0.318	0.626	0.638	0.527	0.231	0.471	0.478	0.393
MixPoisP ($k = 3$)	0.539	1.084	1.172	0.932	0.318	0.623	0.635	0.525	0.231	0.470	0.477	0.393
MixPoisP ($k = 2$)	0.538	1.084	1.171	0.932	0.318	0.623	0.636	0.526	0.231	0.470	0.477	0.393
MixNormP ($k = 3$)	0.539	1.076	1.177	0.931	0.326	0.624	0.640	0.530	0.234	0.476	0.486	0.399
MixNormP ($k = 2$)	0.539	1.103	1.173	0.938	0.323	0.621	0.651	0.532	0.237	0.478	0.483	0.399

Tabla 6.8: EPM para la Elección Presidencial de 2018 calculado con 200 muestras seleccionadas vía MAE.

Hay una reducción considerable en los EPM obtenidos bajo un MAE respecto a los obtenidos bajo un MASSR para los tres tamaños de muestras, sin embargo, la comparativa de los errores porcentuales para cada propuesta bajo un MAE es igualmente difícil ya que las diferencias son marginales, además que no hay una misma propuesta que tenga el menor error porcentual medio en los tres tamaños de muestras ni una que tenga el mayor error porcentual. Podemos notar que los modelos basados en las distribuciones normales presentan los mayores errores porcentuales, mientras que el estimador de razón combinado y los modelos basados en la distribución porcentual tienen errores porcentuales muy parecidos.

Capítulo 7

Conclusiones

Es evidente que en una tarea aparentemente simple, como la estimación de intervalos de confianza para el porcentaje de votos obtenidos por cada fuerza política (u opción de respuesta en un ejercicio democrático), pueden surgir diversas complicaciones logísticas y técnicas. Por lo tanto, resulta esencial identificar propuestas metodológicas de estimación que sean robustas y eficientes. Este aspecto cobra relevancia al considerar el objetivo primordial del Conteo Rápido: proporcionar estimaciones oportunas y confiables sobre los resultados de los ejercicios democráticos.

Ciertamente, el estimador de razón combinado se mantiene como una alternativa apropiada cuando se busca hacer inferencia sobre porcentajes. Sin embargo, es notable el efecto negativo que la estratificación de la población, y por ende de la muestra, tiene sobre la calidad de las estimaciones que proporciona. Este impacto se acentúa aún más cuando los datos presentan características singulares, como los valores atípicos presentes en los registros de votos nulos durante la Consulta Popular de 2021.

Las estrategias teóricamente más adecuadas, que proponen modelos basados en distribuciones discretas, parecen enfrentar dificultades para cuantificar la incertidumbre de los datos. Es bien sabido que este tipo de modelos establecen supuestos sobre la varianza de manera muy restrictiva: en el modelo Poisson, la media y la varianza son iguales; en el Binomial, la varianza es una función de la media y siempre es menor que la media; para el Binomial negativo, la varianza es una función de la media y, en este caso, la varianza es mayor que la media, etc.

Una opción para superar estas restricciones sería buscar distribuciones discretas que ofrezcan mayor flexibilidad, esencialmente generalizaciones de las distribuciones mencionadas anteriormente. Esta podría ser una dirección prometedora para futuras investigaciones. Otro enfoque es utilizar mezclas de distribuciones discretas, lo que proporciona la flexibilidad necesaria para modelar datos heterogéneos. De hecho, la mezcla de distribuciones Poisson demuestra un mejor rendimiento a medida que tanto el tamaño de la muestra, n , como el número de componentes, k , aumentan. Sin embargo, en casos de tamaños de muestra muy pequeños, no tiene sentido incrementar el número de componentes.

Específicamente, en las elecciones presidenciales, es necesario repartir el tamaño de muestra total, n , entre numerosos estratos, resultando en tamaños de muestra por estrato, n_h , muy reducidos. Esto sin mencionar la complejidad computacional de

la implementación, que se vuelve especialmente desafiante cuando k es pequeño, sin lograr mejorar los ya de por sí modestos resultados obtenidos con la alternativa más simple, un modelo Poisson Bayesiano tradicional.

Cuando se opta por una distribución que permite un ajuste más flexible en la varianza, como la distribución normal, se observa una mejora sustancial en los resultados. En la tesis, se exploraron dos modelos: el modelo normal Bayesiano “tradicional” y el modelo de mezclas de normales univariadas. Resulta interesante destacar que el modelo normal Bayesiano logra mejores resultados en comparación con los modelos predictivos basados en mezclas de normales univariadas. En el caso del MASSR, podría afirmarse que la mezcla supera de manera muy marginal al modelo normal Bayesiano. En cambio, cuando se trata del MAE, el modelo normal Bayesiano se erige como la mejor opción.

Considerando nuestro estimador de referencia, el estimador de razón combinado, también se puede afirmar que el modelo normal Bayesiano muestra un rendimiento destacado, especialmente cuando se enfrenta a observaciones atípicas, como las presentes en la Consulta Popular de 2021 y bajo un MAE.

De hecho, en general, la estrategia más robusta resultó ser el modelo normal Bayesiano “tradicional”. Al analizar con mayor detalle este punto, observamos que la base de esta estrategia consiste en modelar de manera robusta \bar{y}_N . Esto se confirma debido a que, bajo un MASSR, en el Capítulo 3, vimos que la distribución predictiva para la media poblacional \bar{y}_N está dada por

$$p(\bar{y}_N | y_{1:n}) = t_{n-1} \left(\bar{y}_n, \left(1 - \frac{n}{N}\right) \frac{s^2}{n} \right).$$

La pregunta obligada es entonces, ¿cómo modela el muestreo probabilístico la media poblacional? De los resultados del Capítulo 2 y asumiendo normalidad asintótica, se tiene que

$$\hat{\bar{y}}_N \sim \mathcal{N} \left(\bar{y}_n, \left(1 - \frac{n}{N}\right) \frac{s^2}{n} \right),$$

donde $\hat{\bar{y}}_N = \frac{1}{N} \sum_{i=1}^N I_i y_i$ es el estimador del muestreo clásico de la media poblacional, y en este caso, lo estamos viendo como una variable aleatoria directamente. Sabemos que la distribución t de Student tiene colas más pesadas que la normal, y esto se acentúa particularmente para tamaños de muestra pequeños. Esta es la razón por la que en el MASSR observamos una mejora marginal en las coberturas por parte del modelo normal Bayesiano.

Pero, ¿qué sucede en el caso del MAE?, donde, por ejemplo, con $n = 1,200$, el tamaño de muestra por estrato es $n_h = 4$ casillas para cada uno de los 300 estratos. Aquí es donde vemos la diferencia en los resultados y ahora es fácil entender lo que sucede: en este caso, el modelo normal va a modelar de manera independiente cada uno de los 300 estratos usando distribuciones t de Student con 3 grados de libertad, mientras que el muestreo clásico seguirá usando una normal asintótica (con media y varianza descritas en el Capítulo 2).

En el contexto específico de un Conteo Rápido, el diseño de la muestra se concibe para evitar ejercer una presión excesiva sobre los CAEs (Capacitadores y Asistentes

Electorales). Los CAEs son responsables de recopilar y enviar los resultados de las casillas seleccionadas en la muestra. No obstante, la función principal de los CAEs el día de la elección es apoyar a los funcionarios de casilla en cualquier problema que pueda surgir. Por lo tanto, se busca la muestra más pequeña posible que proporcione la mejor precisión; consulte la expresión 2.23 en el Capítulo 2 para más detalles.

En consecuencia, bajo un MAE con un gran número de estratos, es necesario tener en cuenta que los modelos basados en mezclas deberán ajustarse a tamaños de muestra muy reducidos, es decir, $n_h \in \{4, 5, 6, \dots, 20\}$.

Finalmente, para cualquier estrategia que se siga, es indispensable considerar el tiempo de procesamiento. Las complejidades y complicaciones añadidas de un ejercicio a gran escala con consideraciones legales y técnicas tan específicas, como lo es el Conteo Rápido, son precisamente lo que inspira y motiva propuestas de solución desde distintas perspectivas y visiones, ya sea dentro de la inferencia frecuentista o dentro de la inferencia Bayesiana.

Bibliografía

- Bernardo, J. and Smith, A. (2000). *Bayesian Theory*, Wiley Series in Probability and Statistics, Wiley.
- Bouveyron, C., Celeux, G., Murphy, T. and Raftery, A. (2019). *Model-based Clustering and Classification for Data Science: With Applications in R*, Cambridge: University Press.
- Casella, G. and Berger, R. (2002). *Statistical Inference*, Duxbury advanced series in statistics and decision sciences, Thomson Learning.
- Cochran, W. G. (1977). *Sampling Techniques*, John Wiley.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1): 1–38.
- Doob, J. L. (1949). Application of the theory of martingales, *Actes du Colloque International Le Calcul des Probabilite's et ses applications*, Paris CNRS, Lyon, pp. 23–27.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics* **7**(1): 1 – 26.
- Ericson, W. A. (1969). Subjective bayesian models in sampling finite populations, *Journal of the Royal Statistical Society. Series B (Methodological)* **31**(2): 195–233.
- Ericson, W. A. (1988). 9 bayesian inference in finite populations, in P. Krishnaiah and C. Rao (eds), *Sampling*, Vol. 6 of *Handbook of Statistics*, Elsevier, pp. 213–246.
- Fong, E., Holmes, C. and Walker, S. G. (2021). Martingale posterior distributions, Disponible vía <https://arxiv.org/abs/2103.15671>.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd edn, Chapman and Hall/CRC.
URL: <http://www.stat.columbia.edu/~gelman/book/>
- Hernández Mendoza, J. (2019). Esquemas de muestreo para el conteo rápido bajo restricciones operativas, Disponible vía <https://bindani.izt.uam.mx/concern/tesiuams/9019s249g?locale=es>.

- Holmes, C. C. and Walker, S. G. (2023). Statistical inference with exchangeability and martingales, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **381**(2247): 1–17.
- Lohr, S. L. (2021). *Sampling: Design and Analysis*, 3rd edn, Chapman and Hall/CRC.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New York, NY.
- Moala, F. A., Ramos, P. L. and Achcar, J. A. (2013). Bayesian inference for two-parameter gamma distribution assuming different noninformative priors, *Revista Colombiana de Estadística* **36**(2): 319–336.
URL: <https://revistas.unal.edu.co/index.php/estad/article/view/44351>
- Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer, New York, NY.
- Wasserman, L. (2003). *All of Statistics. A Concise Course in Statistical Inference*, Springer.
- Yadav, S. and Mishra, S. (2015). Developing improved predictive estimator for finite population mean using auxiliary information, *Statistica* **95**(1): 76–85.