UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

MUESTREO ESTRATIFICADO PREVIA SEGMENTACIÓN SOCIODEMOGRÁFICA:
UNA PROPUESTA PARA ESTUDIOS ELECTORALES

REPORTE DE TRABAJO PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE

ACTUARIA

PRESENTA

TANIA MAYELLI SARMIENTO TORRES

TUTOR:

M. EN C. JOSÉ SALVADOR ZAMORA MUÑOZ

2019

1. Datos del alumno
Sarmiento
Torres
Tania Mayelli
5534979711
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
305019684

2. Datos del tutor
M en C
Salvador
Zamora
Muñoz

3. Datos del sinodal 1
Dr
Carlos Erwin
Rodríguez
Hernández-Vela

4. Datos del sinodal 2
Dra
Sofía
Villers
Gómez

5. Datos del sinodal 3
Dra
Lizbeth
Naranjo
Albarrán

6. Datos del sinodal 4
M en C
Patricia Isabel
Romero
Mares

7. Datos del trabajo escrito
Muestreo Estratificado Previa Segmentación Sociodemográfica: una propuesta para estudios electorales
94 p
2019

Winwood Reade escribe muy bien acerca del tema -dijo Holmes-. Hace observar que mientras el hombre, tomado individualmente, es un acertijo irresoluble, el conjunto de los hombres se convierte en una certidumbre matemática. No puede usted, por ejemplo, anunciar de antemano qué es lo que hará un hombre determinado, pero se puede prever con precisión lo que hará la mayoría de ellos. Eso es lo que dice la estadística.

Sherlock Holmes

El signo de los cuatro (Sherlock Holmes No. 2)
Sir Arthur Conan Doyle

Gracias...

A mi tutor, el M. en C. Salvador Zamora Muñoz por su paciencia, por su apoyo, su guía y enseñanzas para la realización del presente trabajo.

A mis sinodales, la M. en C. Patricia Romero, la Dra. Lizbeth Naranjo, la Dra. Sofía Villers y el Dr. Carlos Erwin Rodríguez, por su tiempo y disposición. Por sus opiniones, que sin duda ayudaron a enriquecer este trabajo.

A la UNAM, por brindarme un segundo hogar, por permitirme formar parte de una comunidad extraordinaria, por el desarrollo profesional, por forjar parte de mi desarrollo ideológico y de mi identidad como persona. Más que mi alma mater ha sido también parte fundamental de mi vida, aún recuerdo la primera vez que visité la UNAM, tenía como 5 años y me maravilló; desde entonces se volvió un sueño, después una meta y finalmente una realidad.

A la Facultad de Ciencias, a mis maestros y ayudantes por todo lo aprendido; por mostrarme la grandeza del conocimiento, por abrazar la diversidad de las ideas, por enseñarme que no hay mayor rebeldía que el saber y que nunca se deja de aprender.

A mis padres por su infinito amor y por siempre estar ahí, por ser los mejores padres del mundo. Gracias mamá por tu apoyo y por todo tu amor, por hacer de mí la persona que soy, por educarme con cariño, por enseñarme el valor de la responsabilidad y de la libertad; por inculcar en mí el amor por mi país, su historia, su cultura, sus sabores, sus colores y su gente; por transmitirme lo invaluable de la justicia, de la equidad y de la democracia; sin duda, estas enseñanzas influyeron para desarrollarme profesionalmente por este camino y transmitirlo en este trabajo. Gracias papá por ser mi ejemplo de disciplina, constancia, de sacrificio y de esfuerzos; por mostrarme el valor de la información y la lectura; por tu herencia revolucionaria y de superación constante.

A mi hermana por su complicidad, por acompañarme en los momentos más difíciles de mi formación académica, por inspirarme, por su apoyo y por su cariño incondicional, por ser la mejor hermana que uno puede desear.

A Manuel Rodríguez Woog (q.e.p.d.) por ser mi mentor en este camino de los estudios aplicados a elecciones y desarrollo de marcas; por ser un maestro, un amigo y un consejero; por impulsar mi carrera y siempre creer en mí; por tu confianza y tu cariño. A Integra, por las enseñanzas tanto personales como profesionales, por darme la oportunidad de crecer en todos los sentidos, por los retos, por la libertad que me dieron para experimentar con nuevas técnicas, sin esto no habría sido posible la realización de este trabajo.

A toda mi familia, mis tíos, mis primos y mis sobrinos por todo su cariño y por toda su alegría.

A mis amigos, los que estuvieron presentes en los momentos más importantes de la carrera; a los que estuvieron desde el inicio, desde aquella primera clase de álgebra; a los que aumentaron mi pasión por la probabilidad y la estadística, con aquellas pláticas interminables; a los cómplices de aquella jardinera cercana al estacionamiento y a mi aliado miembro de aquel comité; a mis amigos del trabajo, por darle a mi día a día la alegría que necesitaba; a todos ustedes, gracias por formar parte de mi vida.

Y a los que se me olviden, a los que no tuvieron un paso afortunado por mi vida, a ustedes que directa o indirectamente forjaron la mujer que soy.

Índice general

Índice de figuras

Capítulo 1

Introducción

Durante la última década los estudios demoscópicos han cobrado particular relevancia en el mundo de la política nacional e internacional, es a partir de ellos, que se toman muchas de las decisiones más importantes en un proceso electoral o durante una gestión de gobierno. Conocer la posición frente a la de sus adversarios, además de entender la manera en la que opinan los ciudadanos, resulta fundamental para cualquier persona que busca un puesto de poder dentro de la política, principalmente aquellos que compiten por cargos de elección popular.

Las encuestas de opinión pública, principalmente las electorales, son utilizadas como medio de propaganda por los políticos, candidatos y gobiernos en general, lo que se ha traducido en el aumento de la oferta de este tipo de estudios; sin embargo, muchos de ellos no cumplen con los requerimientos mínimos de una investigación estadística formal. El universo de estudio se selecciona por conveniencia, se malinterpreta el significado de error muestral, se generan intervalos de confianza sin seguir los principios de la teoría estadística, e incluso en muchas ocasiones, se generan conclusiones a partir de una muestra incompleta y sin ninguna justificación.

El desarrollo tecnológico y el avance en las comunicaciones, han permitido que la labor estadística encuentre nuevas formas para generar información de valor, además de realizar estimaciones y modelos en tiempo récord. En apenas unos minutos se pueden generar modelos sumamente robustos y eficientes. Varios periodistas e investigadores aseguran que vivimos en *"La Era de la Información"* como consecuencia de la evolución exponencial de la digitalización en las sociedades modernas, ha sido a partir de esto, que la estadística se ha fortalecido y apreciado; sin embargo, también ha crecido el número de detractores y oportunistas.

Los estudios demoscópicos, por su parte, han encontrado en este desarrollo tecnológico un sinfin de metodologías que permiten realizar investigaciones a bajo costo y con la promesa de obtener resultados prácticamente en tiempo real. Si bien es cierto, existen empresas que han enfocado sus esfuerzos en esto para obtener resultados precisos y encontrar un sello distintivo, la mayoría no sólo no lo logra, sino que sus conclusiones resultan por demás fraudulentas; por si esto no fuera suficiente, han usado la coincidencia y el azar para validar sus métodos y promocionarse en el sector.

Todo lo anterior se suma a la poca precisión que ha existido en ejercicios anteriores, las elecciones en Estados Unidos del 2016, donde la mayor parte de las encuestas anticipaban una victoria para la candidata demócrata Hillary Clinton y finalmente la victoria fue para Donald Trump; el plebiscito en Colombia para el acuerdo de paz con las FARC, en el que se pensaba que ganaría el "Sí" porque así lo decían los estudios previos o la consulta en Reino Unido para su separación de la Unión Europea, donde los medios de comunicación alertaban que ganaría el "No", con base en las encuestas publicadas. En México las elecciones de Nuevo León en 2015, donde sorpresivamente el candidato independiente Jaime Rodríguez Calderón obtuvo la gubernatura, pese a que las encuestas veían un triunfo holgado para la candidata del PRI; o las elecciones presidenciales de 2012 en México, en las que la mayoría de las encuestadoras fallaron en sus estimaciones, incluso hubo quienes salieron a aceptar públicamente haberse equivocado, otras fueron señaladas de favorecer intencionalmente al candidato del PRI; lo cierto es que la mayoría decía que Enrique Peña Nieto ganaría con una ventaja cómoda, pero la diferencia con Andres Manuel López Obrador que obtuvo el segundo lugar, fue de menos de 6 puntos porcentuales. Todos los anteriores son solo algunos ejemplos de las fallas que ha tenido el sector en materia de estimación, razón por la cual los estudios demoscópicos, principalmente las encuestas electorales, han sido objeto de diversas críticas y descalificaciones, algunas de ellas impulsadas por los propios protagonistas de la contienda.

En algunos casos, las elecciones han sido tan cerradas que cualquier escenario planteado tenía una probabilidad importante de ocurrir; aquí es cuando la técnica se ve rebasada y es labor del investigador precisar en la interpretación de los resultados. Con el prestigio y la confianza a la baja, ha surgido la necesidad de reinvención en las empresas, con lo que se han desarrollado algunas soluciones; las más populares se basan en el agregado y ponderación de las encuestas públicas (*poll of polls*), muchas están basadas en el trabajo de Nate Silver, quien ganó popularidad en 2008 cuando pronosticó la victoria de Barack Obama en Estados Unidos, aún cuando la mayoría de las encuestas lo ubicaban en segundo lugar.

Algunas de las principales soluciones para mejorar la estimación de resultados, se han basado en la aplicación de técnicas empleadas tradicionalmente en otras disciplinas, que van desde modelos financieros hasta modelos multivariados con aplicaciones médicas y biológicas por mencionar algunas; si bien es cierto, todas ellas funcionan como alternativas a la estimación y así plantear diferentes escenarios probables, todas parten de suponer que se tiene una muestra confiable. Con base en la experiencia del sector, cuando se trata de modificar el muestreo se recurre a aumentar el tamaño de la muestra, lo que aumenta considerablemente el costo, otros se niegan radicalmente a la sustitución de unidades muestrales; y siguiendo esta lógica, algunos otros generan una selección aleatoria (al menos teórica) sin reemplazo de la última unidad de muestreo (el informante), lo cual además de aumentar los costos de operación, genera incertidumbre sobre el tamaño de la muestra.

De acuerdo a nuestras experiencias realizando estudios de este tipo, hemos decidido centrar nuestro trabajo en la selección de la muestra para estudios electorales, pensando que sin una buena muestra, cualquier esfuerzo posterior por obtener estimaciones adecuadas se quedará corto. Partiendo del hecho que la mayoría de las empresas con metodologías sólidas, usa una selección polietápica,

estratificada y por conglomerados, maximizando los recursos destinados para la investigación; la propuesta se enfoca por lo tanto, en la estratificación de la muestra usando los datos e indicadores sociodemográficos del INEGI, los cuales permiten identificar perfiles de votantes (o de informantes óptimos) y en consecuencia segmentarlos conforme al potencial de participación en las elecciones. Originalmente la estratificación se basa en características geográficas, lo cual aumenta el riesgo de subestimar la intención de voto del candidato, cuyos votantes se encuentran en un espacio geográfico "pequeño" pero densamente poblado.

Capítulo 2

Las Encuestas

2.1. ¿Qué son?

Desde el punto de vista estadístico, una encuesta debe estar precedida de un muestreo que sea representativo de la población de interés; es decir, que las características de ésta sean parecidas a la población, además se debe garantizar la aleatoriedad de las unidades de muestreo, ambas características son clave para un correcto análisis e interpretación de los resultados.

Las encuestas son el instrumento utilizado dentro de una investigación, que tiene como objetivo obtener información acerca de alguna población de interés, a partir de la aplicación de un cuestionario diseñado previamente; el cuestionario debe ser idéntico para todos los participantes de la investigación, se pretende que la aplicación del mismo sea clara y puntual, sin dejar espacio a interpretaciones sobre una misma variable.

Las encuestas son sin duda el insumo por excelencia para toda investigación social que requiera de evidencia estadística, fundamentada en el método científico, es por eso que resulta de vital importancia que este insumo sea atendido con rigurosidad, tanto desde el punto de vista social como en el punto de vista estadístico. Hablando del diseño del cuestionario, existen necesidades básicas, por ejemplo, que la extensión y dinámica de su aplicación permitan que el entrevistado mantenga interés en todo momento, evitando así patrones en las respuestas. Se quiere que la mayor parte del instrumento contenga respuestas de opción múltiple que mantengan una escala con crecimiento homogéneo, obteniendo paridad entre positivos y negativos sin lugar a posibles sesgos; las preguntas abiertas deben ser limitadas, ya que las respuestas y la captura de ellas, conlleva una labor subjetiva que afecta directamente el análisis estadístico.

2.2. ¿Qué tipo de estudios se hacen en un proceso electoral?

En la mayoría de los estudios de opinión pública se parte de modelos teóricos para obtener estimaciones y conclusiones; es decir, cada estimación presentada se toma como cierta debido a las propiedades probabilísticas de los estimadores, pero difícilmente existe un ejercicio que recoja

las opiniones de la población total para corroborar su precisión (a excepción de los censos). En el caso de los estudios electorales, los escenarios planteados a partir del análisis de la información son contrastados con el resultado final de la jornada electoral, todos los resultados obtenidos son evaluados y juzgados por expertos en la materia, además de la opinión ciudadana.

Por estas razones, se debe ser muy cuidadoso al realizar un estudio electoral, pues el prestigio de cualquier empresa o investigador, e incluso del sector en general, se pone en juego. Existen diferentes estudios de opinión pública que se hacen durante un proceso electoral, y aunque la metodología en todos los casos es similar, los objetivos e interpretación de resultados en cada caso varía. A continuación se describen los principales:

- **Estudios *Benchmark* o de posicionamiento:** Estos estudios comprenden la primer etapa de un proceso electoral, se realizan previo a la designación de candidatos, tienen como objetivo conocer las posibilidades de un aspirante a un cargo de elección popular de resultar elegido, además de contrastar sus puntos a favor y en contra con los de otros aspirantes, sin importar si pertenecen al mismo partido o no.

- **Estudios preelectorales:** Se trata de encuestas y sondeos realizados previo al inicio de las campañas, su objetivo es conocer la posición de los candidatos antes del arranque de las campañas, estos estudios son los primeros en ser expuestos públicamente, pues funcionan para verificar las posibilidades de competencia de alguno de los aspirantes, en general reflejan el nivel de conocimiento y popularidad que se tiene entre la ciudadanía.

- **Estudios electorales:** En general son los de mayor popularidad entre la ciudadanía, se realizan durante el periodo de campañas, en ellos se presentan escenarios probables a partir de la estimación de la intención de voto; para el planteamiento de escenarios se usan modelos y variables que exploren la probabilidad de voto, así como la seguridad en la decisión.

- **Encuesta de Salida o *Exit Poll*:** Se realiza durante la jornada electoral, es decir, desde la apertura de las casillas y hasta el cierre de las mismas, su objetivo es conocer las tendencias de los resultados en tiempo real, esto se realiza a través de una breve encuesta a la salida de las casillas.

- **Conteo Rápido:** Para muchos el estudio más confiable y preciso para estimar los resultados reales, se trata del único estudio que también realiza públicamente la autoridad electoral, en el caso de México, el Instituto Nacional Electoral (INE). La estimación de resultados se basa en evidencia sólida del conteo previo de los votos emitidos, a diferencia de las anteriores, no depende de la subjetividad de la respuesta de los informantes, pues las últimas unidades de muestreo son las casillas electorales.

En ocaciones se confunde a los sondeos con las encuestas, los primeros son estudios ágiles que se realizan con pocas variables (1 ó 2) para obtener resultados instantáneos sin la necesidad de procesar la información, pues la muestra no es representativa de la población y se trata de un acercamiento general al comportamiento de las tendencias de votación.

Para las encuestas realizadas previas a la jornada electoral, pueden existir varias metodologías de recolección de datos, además de distintos diseños de muestreo. Se pueden realizar cara a cara en viviendas, a través de una llamada telefónica a números fijos y/o móviles, en línea a partir de la invitación por correo electrónico, redes sociales o publicidad; aunque menos comunes, también se realizan a través de correo convencional o en puntos de afluencia. En todos los casos la población de estudio es definida por el investigador, así como la generalización de resultados a cierto sector de la población o a toda ella; debido a las limitantes en algunos métodos de recolección de datos, se considera que la encuesta cara a cara en viviendas genera resultados más precisos y confiables, sobre todo en un país como México, donde la disposición a tecnologías de la información aún es limitada.

2.3. ¿En dónde y cómo surgen las encuestas?

No se conoce con exactitud el origen de las encuestas, aunque los romanos ya realizaban censos poblacionales durante su imperio, no solían usarlos como medio de información sobre las opiniones del pueblo, se realizaban exclusivamente con fines fiscales. Se piensa que las primeras encuestas se llevarían a cabo con fines políticos por *"Los Reformadores Sociales"* de la Revolución Industrial en el siglo XVIII; este grupo conformado principalmente por burgueses, tenía un especial interés en plantear normas de seguridad y salud debido a las condiciones en las que vivían los obreros, llegando a causar un problema de salubridad en todo el territorio. De esta manera se acercaron a los obreros para conocer su opinión y postura a través de una entrevista homogeneizada, para agrupar y presentar las respuestas en forma de iniciativa.

Así pues, es en Europa del Este donde se tienen los primeros registros de encuestas en el mundo; de manera paralela iniciarían las primeras investigaciones sociales, como se puede observar, no podría entenderse una sin la otra. Surgirían como una forma de darle voz a la clase trabajadora frente a los gobiernos y evidenciar las malas decisiones que se habían tomado, pensando únicamente en el desarrollo de la industria. Karl Marx es considerado como la primer persona en realizar una encuesta por correo, revolucionando así la manera de recabar la información; en uno de sus estudios sobre la calidad de vida de los trabajadores envió 25,000 cuestionarios para fundamentar sus ideas; sin embargo, fracasaría en este intento, pues la respuesta fue prácticamente nula.

A estos primeros intentos por usar una encuesta como recurso para investigaciones científicas, les seguirían estudios psicológicos, incluso, se piensa que los médicos desarrollaron de esta manera su experiencia para identificar patrones de enfermedades y tratamientos a seguir. Pero fue hasta inicios del siglo XX, cuando se desarrolla la encuesta como un procedimiento válido para los estándares de una investigación científica de calidad, los avances en la materia se deben en definitiva, al desarrollo de la teoría de probabilidad y las bases del muestreo estadístico. Arthur L. Bowley y A. R. Burnett-Hurt [1] usarían el muestreo para obtener a sus informantes y serían los primeros

[1] **En el Estudio *Livelihood and Poverty* realizado con una encuesta a 115 personas:** La encuesta. P. López-Roldán y S. Fachelli, 2015. Metodología de la Investigación Social Cuantitativa. Bellaterra (Cerdanyola del Vallés): Dipósit Digital de Documents, Universitat Autónoma de Barcelona. Capítulo II.3. Edición digital: http://ddd.uab.cat/record/163567

en realizar una selección aleatoria de los mismos; sin embargo, fue hasta los años 30 cuando los estudios de Jerzy Neyman, concluirían con el aporte de conceptos clave para el muestreo actual: error muestral, población finita e infinita, distribución muestral, la estratificación de la muestra, por mencionar algunos, a ellos se sumarían los estudios realizados por Mahalanobis, Yates, Deming, Cochran o Kish para dar paso así a la consolidación de este tipo de investigaciones con el método científico aplicado de manera correcta, asimismo el desarrollo de técnicas estadísticas serían cada vez más relevantes.

En 1936 durante las elecciones presidenciales de Estados Unidos, George Gallup anticiparía el triunfo de Franklin Delano Roosevelt con la primer encuesta electoral, aplicada a 5,000 personas alrededor de todo el territorio Estadounidense. Gallup trasladaría su conocimiento y experiencia en la investigación de mercados para concretar resultados en su encuesta electoral; vale la pena destacar que sus resultados se contraponían a lo publicado por la prestigiosa revista *Literary Digest* quien había usado una muestra de 2.3 millones de personas; evidenciando que una muestra más allá del tamaño, depende de un buen diseño metodológico. A partir de ese momento iniciaría un crecimiento exponencial de este tipo de estudios; por su parte Gallup extendería su trabajo por Reino Unido y Francia, donde presentaba tendencias inesperadas que finalmente se cumplían.

Ante el papel que poco a poco cobraban las encuestas, los candidatos no dudarían en usarlas para el planteamiento de su estrategia de campaña. John F. Kennedy[2] sería el primer candidato en usar las conclusiones obtenidas a partir de las encuestas, para delinear su perfil con base en los adeptos que la gente veía en él; de igual forma, usaría las encuestas para identificar de mejor manera los puntos débiles de sus contrincantes y forjar una sólida campaña que lo llevaría a la Casa Blanca.

Hoy en día los avances en las tecnologías de la información y comunicación, permiten que se puedan aplicar encuestas en muy poco tiempo, además de obtener resultados casi en tiempo real; sin embargo, la falta de claridad en el perfil del informante pone en duda la viabilidad de su metodología, sobre todo cuando se realizan a través de las redes sociales.

2.4. ¿Cuándo se inician las encuestas en México?

En nuestro país, el estudio de la opinión pública no era de particular interés, sobre todo tratándose de temas relacionados con elecciones presidenciales, pues hasta antes de la década de los 80 se conocía de antemano el resultado de todos los procesos electorales, el poder político era monopolizado por un solo partido político que se imponía ante cualquier intento de ejercicio democrático; por esta razón, durante mucho tiempo este tipo de estudios eran realizados únicamente por las instituciones gubernamentales, para mejorar su imagen entre la opinión ciudadana, así como plantear políticas públicas a partir de las necesidades de la población. El INEGI se encargaría de generar indicadores sociodemográficos que permitían exponer el panorama de los mexicanos frente a la clase política, además de generar información y conocimiento de la sociedad en general; no sería una sorpresa que muchos de los investigadores del INEGI sentaran las bases de las

[2]Historia de las encuestas en el mundo:https://aprendeenlinea.udea.edu.co/revistas/index.php/ceo/article/viewFile/6549/5999

encuestas electorales, una vez que su apertura y posición fue clave en los procesos democráticos.

A pesar de las limitantes en la investigación a partir de encuestas, que se vivía en México, el Dr. László Radványi[3], académico e investigador de la UNAM, realizó en los años 40 decenas de estudios de opinión pública con temas muy variados, con lo que destacaría la importancia del planteamiento en las preguntas del instrumento, evitando así cualquier sesgo, además destacó la importancia de seleccionar correctamente una muestra; tiempo más tarde, Radványi se convertiría en uno de los miembros fundadores de la Asociación Mundial de Investigación en Opinión Pública (WAPOR).

En 1988, durante las elecciones presidenciales, Miguel Basañéz[4] realizó las que se consideran las primeras encuestas electorales públicas, en la primera de dos, realizada en la Ciudad de México, anticipaba un cómodo triunfo de Cuauhtémoc Cárdenas como candidato a la presidencia del país, en la segunda, realizada en todo el país, colocaba a Carlos Salinas de Gortari como ganador apenas con pocos puntos de ventaja sobre Cárdenas; el resultado obtenido en estas encuestas, sumado a otras realizadas principalmente por universidades, pondrían en duda la veracidad de los resultados finales de la elección, pues la tendencia indicaba una victoria de quien hasta entonces iba en segundo lugar. En esa época los estudios carecían de las facilidades tecnológicas de hoy, así como la dedicación en el diseño metodológico de la selección de muestra, por lo que es digno de reconocer el esfuerzo realizado para el levantamiento de la información y el procesamiento de la misma.

Durante el proceso electoral presidencial de 2006, las encuestas sufrirían del primer descalabro en reputación. En la campaña la mayoría de las encuestas ponían al frente a Andrés Manuel López Obrador, pero conforme se acercaba el día de la elección, la ventaja del candidato se reducía drásticamente; sin embargo, días antes de la elección los estudios anticipaban la victoria de López Obrador. Llegado el día de la elección el IFE (Instituto Federal Electoral) decidió no dar a conocer las estimaciones de sus conteos, por lo que la ciudadanía comenzó a desconfiar de la veracidad del resultado; al iniciar los cómputos distritales AMLO se puso al frente de Felipe Calderón y mantuvo la ventaja por varias horas, durante la madrugada se presentó un cruce en las tendencias que fue severamente cuestionado por los especialistas en estadística, ya que un fenómeno de esta naturaleza, solo se podía explicar con votaciones atípicas en ciertas casillas, conteos equivocados en un alto porcentaje de casillas o un fraude prefabricado. La incertidumbre en los resultados y la diferencia con las estimaciones de las encuestas, llevaron a AMLO a declarar su desconfianza en las encuestas, misma que transmitiría a sus simpatizantes, el resto de la ciudadanía tampoco estaba convencida de la fiabilidad de éstas, sobre todo por lo acontecido durante la noche de la jornada electoral. Esta desconfianza crecería con las elecciones de 2012, principalmente por la mala práctica y falta de ética de algunos investigadores; y el proceso local en Nuevo León de 2015 para elegir gobernador, una elección con características muy particulares. Por este motivo, en las elecciones de 2018 se usarían distintas alternativas para mejorar la calidad de la estimación y la interpretación de resultados, mismas que se abordarán más adelante.

[3] https://academic.oup.com/ijpor/article/21/1/3/776405

[4] http://www.eluniversal.com.mx/entrada-de-opinion/articulo/andrew-selee/nacion/politica/2015/08/8/miguel-basanez-hombre-democratico

2.5. ¿Por qué fallan?

Es la pregunta que ha rondado la cabeza de los investigadores durante los últimos años, una encuesta falla si los resultados finales distan mucho de la estimación o la contradicen; a partir de los resultados obtenidos en ejercicios como las elecciones en Estados Unidos en el 2016, el Brexit en Reino Unido, el plebiscito de pacificación con las FARC en Colombia, así como diferentes procesos electorales en Latinoamérica o las últimas elecciones en Francia, en los casos anteriores la mayoría de las encuestas quedaron lejanas al resultado real. En el caso de México, las elecciones federales de 2006, 2012 y más recientemente la de 2018 donde a pesar de la precisión en la posición de los candidatos, la diferencia entre ellos fue radicalmente diferente a la presentada por la mayoría de las encuestas; uno de los casos más graves ocurrió en la elección para gobernador de Nuevo León, donde los resultados fueron por demás sorpresivos para la mayoría de los investigadores.

Una de las razones de la creciente desconfianza en este tipo de estudios, obedece a la falta de regulación en el tema. Previo a 2014[5], la regulación de estudios electorales era prácticamente nula, razón por la cual existían cientos de publicaciones con referencias a estudios, muchas veces no realizados, en otros casos los estudios publicados no cumplían con los criterios de un protocolo de investigación estadística, pues la muestra se centraba en un sector de la población y no en toda, además de usar métodos que favorecían a un candidato sobre otro. Además de esto, el uso de las encuestas como propaganda política, intensificó las críticas y desconfianza en el método; actualmente existe una regulación que prohibe la publicación de encuestas que no hayan sido verificadas por el INE a partir de su metodología y objetivos, a pesar de ello, existen estudios no verificados que se viralizan a través de las redes sociales.

Otra de las limitantes de los estudios electorales, responde a las declaraciones de candidatos que cuestionan el método, esto tiene como consecuencia un aumento significativo en el rechazo de las encuestas o bien el crecimiento atípico de la no respuesta. Ambos problemas conllevan a una mala estimación, producto de un levantamiento deficiente; a pesar de que esto no es responsabilidad del investigador, se han tenido que desarrollar metodologías previas al levantamiento y modelos estadísticos posteriores a la integración de la base de datos. El estudio de la no respuesta es uno de los grandes retos del sector, pues dentro de esta información se encuentra la clave de los escenarios probables en la elección; algunos investigadores y empresas han dedicado importantes esfuerzos para evitar la pérdida de información, usando modelos y técnicas utilizadas típicamente en otros campos, como finanzas, salud, etc. Las regresiones logísticas, clasificación de los individuos para generar segmentos, análisis factorial para generar índices, simulaciones Monte Carlo, *Bootstraping*, cada vez son más frecuentes en los estudios electorales.

La mala interpretación de los resultados es una constante, en muchas ocasiones se pone en duda la representatividad de la muestra debido a su tamaño, sin entender que ésta depende en mayor medida de otros conceptos como la aleatoriedad de la selección, así como la cobertura y dispersión de la misma. Como Gallup lo demostró, una muestra de 5,000 casos puede ser más precisa que

[5] https://www.ine.mx/la-regulacion-encuestas-electorales/

una de 2.3 millones de informantes; en la población se tienen conceptos totalmente equivocados, por ejemplo, la probabilidad de selección es muy pequeña, por lo que no es raro que una persona no haya sido entrevistada jamás; extrañamente, hay quienes consideran representativa una muestra conforme a la relación del tamaño de ella y el de la población, llegando a conclusiones como decir que una muestra es representativa si se estudia al 10 % del total, lo cual representaría un estudio de alrededor de 8,000,000 de personas, imposible de costear para la mayoría de instituciones, empresas e investigadores; lamentablemente los conceptos no son sencillos de explicar para entender que muestras de 1,000 personas, pueden ser altamente precisas, conforme al diseño de muestreo. Otro problema de interpretación corresponde al resultado presentado como pronóstico, cuando en realidad es un indicador del comportamiento ciudadano, en un momento en particular, funciona como medidor de tendencias pero no como oráculo; los resultados obtenidos plantean los escenarios más probables para el día de la elección, pronosticar un resultado depende del análisis de la tendencia en estudios realizados bajo los mismos criterios, pero aún así pueden variar debido a lo vulnerable de la opinión pública.

Actualmente se ofrecen metodologías que aseguran un levantamiento de la información y resultados en tiempo real, aunque esto es teóricamente posible, la brecha tecnológica entre la ciudadanía complica la precisión de resultados, han surgido personas y empresas sin fundamentos estadísticos, que realizan y publican estudios realizados en *Facebook* o *Twitter*, cuando el perfil de usuarios de redes sociales y tecnologías de la información en general, sesga los resultados en favor de algún candidato en particular, cayendo en este fenómeno de la nueva era de la información: sensacionalismo o notas falsas que parecen verdaderas.

Desde el punto de vista técnico, en la mayoría de los casos se siguen usando formas de estimación descriptiva, que no profundizan en la riqueza de la información, se generan indicadores de la misma manera en la que se hacían hace 25 años; a partir del crecimiento en la popularidad del *Big Data* o la Minería de Datos, los investigadores deberían retomar estas técnicas y aplicarlas en sus estudios; es cierto que existen esfuerzos para brindar alternativas al estudio de la información, pero aún carecen del potencial de estimación que se necesita. Durante décadas, la creatividad era el motor del desarrollo de metodologías, actualmente se ha dejado de lado debido a la necesidad de publicar resultados antes que nadie.

Finalmente, la poca atención al muestreo, objetivo principal de este trabajo. La manera en la que se selecciona la muestra difícilmente puede sufrir modificaciones, debido a que es la mejor manera de optimizar recursos sin perder precisión en los resultados; algunos investigadores han planteado una alternativa que consiste en la selección de una muestra de tamaño considerablemente mayor al usual, de esta manera se busca hacer contacto con el informante seleccionado, sin opción de sustitución en caso de negativa; este tipo de levantamiento tiene la ventaja de controlar la selección hasta la última unidad, pero no se conoce el tamaño de la muestra sino hasta terminado el levantamiento, por lo que no resulta una opción popular dado el alto costo e incertidumbre en algunos aspectos. Otros han intentado innovar en la manera en la que se levanta la información, usando tabletas electrónicas para disminuir las fallas en la aplicación del cuestionario; sin embargo, además del aumento en el costo, no existe evidencia suficiente en la mejora de la estimación o la disminu-

ción de la no respuesta. Por otro lado, hay quienes han intentado evitar a toda costa la sustitución de unidades de muestreo, ya que esto afecta directamente la aleatoriedad de la misma; sin embargo, al existir zonas inaccesibles por el terreno o bien por la inseguridad que se vive, hacen inviable estas prácticas.

Es por todo lo anterior y tomando como referencia la experiencia que se tiene en estos estudios, que se centra la atención de este trabajo en el muestreo, pues a pesar de lo valioso de las técnicas y modelos estadísticos usados en el análisis de la información por algunos investigadores, no se puede tener certeza de su potencial si la muestra no es seleccionada de manera adecuada. A continuación se presenta una propuesta de alternativa al muestreo, para la realización de encuestas de opinión pública.

Capítulo 3

Objetivo y descripción de los recursos

3.1. Objetivo del reporte

Plantear una alternativa sobre la estratificación de una muestra para estudios de opinión pública, principalmente estudios electorales previos a la jornada electoral.

Generalmente en el diseño de la muestra para estos estudios se usa el modelo estratificado; para definir los estratos relevantes en el estudio se recurre a la división planteada por el INE en estrato rural, mixto y urbano (en ocasiones urbano, no urbano) por las características de la zona donde se encuentran las secciones electorales; además de las circunscripciones, distritos federales y/o locales. El INE divide el territorio nacional en secciones que dependen de algunos factores como la densidad de población, la extensión del territorio y las características del entorno; sin embargo, esta estratificación tiene algunas limitantes:

- La división es principalmente geográfica, en consecuencia la muestra obtenida será mayor en estados con muchas secciones electorales que en estados con mayor participación. Por ejemplo Oaxaca, que tiene un número elevado de secciones electorales, pero con poca influencia en los resultados finales.

- Al estratificar de esta manera, se da un peso particular a las secciones rurales y/o mixtas; se limita así la selección de la muestra, restando probabilidad de selección a secciones con mayor participación o bien con menor riesgo de no respuesta.

- Aunque anteriormente existía homogeneidad dentro de estos estratos, en la actualidad no hay evidencia de esto. Existen resultados para asumir homogeneidad en perfiles similares, basados en más indicadores que solo el tipo de sección.

- Cada vez es más relevante el estudio de votantes probables, hay mayor precisón en la estimación si se entrevista a personas con mayor probabilidad de ir a votar, los estratos rural y/o mixto, aportan más informantes con menor probabilidad de votar, sesgando los resultados a favor de algún candidato.

- Esta estratificación otorga una probabilidad homogénea de selección a las unidades en muestra (secciones); con esto, la probabilidad de selección de un informante rural y/o mixto resulta mucho mayor a la de un informante urbano.
- Ponderar los resultados por sexo y edad, después del levantamiento es una práctica recurrente. Si se tiene información disponible, se debería usar a favor, para limitar el impacto de los ponderadores, por lo que una alternativa dentro de la estratificación resultaría de ayuda.
- Se generan conclusiones para toda la población mayor de edad y no sobre los electores; poseer una identificación del INE no es suficiente para ser un informante adecuado en este tipo de estudios.

La propuesta consiste en usar información oficial, pública y disponible para generar una estratificación *ad-hoc* a cada tipo de estudio de opinión pública, con base en los objetivos planteados en cada investigación y usar esto como la base de la estratificación para la muestra. En un proceso electoral presidencial, se tiene información previa que permite identificar el perfil de ciudadanos con mayor participación, así se obtendrían segmentos de ciudadanos con similitud en la probabilidad de participación, además de ello, usar los indicadores sociodemográficos obtenidos de censos y conteos del Instituto Nacional de Estadística y Geografía (INEGI) aportan mayor información a los perfiles, delineando mejor sus características; con esto existiría una mayor probabilidad de que los ponderadores finales impacten lo menos posible en la información, pues en teoría aumentaremos la probabilidad de entrevistar ciudadanos que sí votarán.

La estratificación que se propone, se basa en la agrupación de los municipios del país conforme a indicadores sociodemográficos relevantes en este tipo de estudios: el sexo, edad, educación, situación de empleo, ingresos y acceso a internet. Diversos estudios señalan que estos indicadores permiten observar patrones de voto en los ciudadanos. Para realizar la segmentación se usará un análisis de conglomerados no jerárquicos por k-medias, más adelante se ahondará en este tema.

El resultado que se busca es una división de las secciones electorales conforme a los indicadores municipales; sin embargo, la muestra no será proporcional al número de secciones dentro de cada segmento, sino al número de ciudadanos (potenciales informantes) que habita en cada partición. De esta manera la probabilidad de selección aumenta en las zonas densamente pobladas, así como en aquellas donde el perfil insinúa una mayor participación.

3.2. Descripción de los recursos empleados en el presente reporte

3.2.1. Base de Datos de Indicadores Sociodemográficos del INEGI (2015)

Base de datos elaborada a partir de la información de la Encuesta Intercensal de 2015, contiene los estimadores de población total, porcentaje de hombres y mujeres, índice de edad, índice de escolaridad, tasa de empleo y desempleo, nivel de ingreso y uso de internet. Los índices fueron construidos a partir de la información del INEGI, para manejar variables ordinales.

Como parte de la necesidad de información actualizada en periodos más cortos que los diez años que comprende la realización de los Censos de Población y Vivienda del INEGI, para 2015 se realizó la primer encuesta intercesal. A diferencia de los conteos, la Encuesta Intercensal del INEGI, se realizó para actualizar la información demográfica y socioeconómica de México, es decir, obtener indicadores más específicos y robustos que los que se obtenían con los conteos.

El marco muestral, fue definido en 2014 a partir del Recorrido de Actualización del Marco Geoestadístico Nacional, del Entorno Urbano y las Características de las Localidades, con la que se actualizó la cartografía urbana y rural del país.

Las unidades de observación fueron viviendas particulares habitadas y sus habitantes; la encuesta se realizó del 2 al 27 de marzo de 2015, siendo el 15 de marzo de 2015 la fecha de referencia de la información. En cada vivienda en muestra, se realizó una entrevista a un informante definido como el jefe o la jefa de la vivienda, en caso de no estar disponible, se optó por un informante de 18 años o más que fuera residente habitual de la vivienda y conociera los datos de todos los residentes. Para cada vivienda, se realizaban hasta cuatro visitas en diferentes horarios y días.

El tamaño de la muestra fue de 6.1 millones de viviendas, obtenido a partir del diseño estadístico del estudio. Además durante el mes de junio del mismo año se realizó una verificación del levantamiento de la información.

Con este diseño de la muestra para la Encuesta Intercensal 2015 se logra actualizar las estimaciones sobre el volumen, la composición y distribución de la población y de las viviendas particulares habitadas en el país, además de los estimadores de proporciones, tasas y promedios de las variables de interés. La muestra consigue estimar los totales de viviendas particulares habitadas y la población que las habita en los siguientes dominios de estudio:

- **Nacional:** Estados Unidos Mexicanos
- **Estatal:** 32 entidades federativas
- **Municipal:** 2,457 municipios o delegaciones en el caso de la Ciudad de México
- **Localidades:** Para cada una de las localidades de 50 mil o más habitantes

Se obtienen también estimadores de proporciones, tasas y promedios de las variables de estudio, en cada uno de los dominios mencionados.

El tamaño de muestra mínimo por cada municipio para obtener estimadores precisos y confiables fue de 1,300 viviendas habitadas; razón por la cual se censan todos los municipios que en el censo de 2010 contaban con este número de viviendas o menor, además de aquellos con características de particular interés como los municipios con población en pobreza extrema, aquellos con rezago social, con población afromexicana y algunas localidades con población hablante de lengua indígena, sobre todo aquellas donde la lengua en cuestión se encuentra en peligro de desaparecer. Debido a esta consideración se obtuvo un tamaño de muestra final de 7.9 millones de viviendas, de

las que 5.9 millones eran habitadas.

El diseño de la muestra fue estratificado por conglomerados y en una sola etapa, las localidades fueron agrupadas en estratos que se clasificaron según el tamaño y nivel socioeconómico; se seleccionaron áreas geográficas completas utilizando muestreo aleatorio simple, estas áreas fungen como los conglomerados del muestreo; para las localidades con más de 50 mil habitantes, la estratificación fue realizada por nivel socioeconómico en las áreas geoestadísticas básicas (AGEB), en cada AGEB en muestra se seleccionaron dos manzanas.

3.2.2. Catálogo de Secciones Electorales del INE (2015)

El Instituto Nacional Electoral (INE) realiza una representación gráfica de la organización del Marco Geoestadístico Electoral (MGE) conocida como cartografía electoral, esta representación se encarga de conocer la distribución de los ciudadanos con derecho a votar dentro del territorio nacional.

Para la organización electoral, el territorio nacional se divide en diferentes capas de información:

- Circunscripciones electorales: 5
- Entidades federativas: 32
- Distritos uninomiales: 300
- Municipios: 2,458: 2,457 municipios incluidos en el INEGI (2015) + 1 incluido en años posteriores
- Secciones electorales: 68,364

Además de localidades urbanas, rurales y manzanas. La información usada en este estudio corresponde al corte realizado para las elecciones federales de 2015. El INE realiza actualizaciones a esta información durante la planeación de los procesos electorales federales, efectuados en el país cada 3 años.

A pesar de existir un corte de febrero de 2018, correspondiente a las últimas elecciones federales, para los estudios previos a la integración de este marco se usó el corte de 2015; debido a la necesidad de comparación de los resultados obtenidos durante todo el proceso, se ignora la división del corte de 2018.

La base de datos está integrada por el total de las secciones electorales corte 2015, identificadas por circunscripción, estado, municipio, distrito, cabecera distrital, tipo de casilla, total de ciudadanos en el padrón electoral, total de ciudadanos inscritos en la lista nominal, domicilio de la casilla correspondiente a la sección, además de los resultados electorales anteriores: 2015, 2012, 2009 y 2006.

3.2.3. Cómputos Distritales de la Elección a Presidente de la República (2018)

Con el objetivo de comprobar los resultados de la propuesta de este trabajo, se hará uso de los resultados oficiales contenidos en los cómputos distritales de la elección de 2018. A partir de estos resultados y usando un porcentaje de no respuesta teórico basado en los datos de las encuestas públicas hechas durante el proceso electoral 2018, se realizarán simulaciones de respuesta en una encuesta hipotética, probando los resultados del muestreo propuesto contra un muestreo conservador.

Los cómputos distritales son la suma de los resultados contenidos en las actas de escrutinio y cómputo de las casillas ubicadas en un distrito electoral; esta suma se realiza en los 300 distritos electorales y son los resultados definitivos del INE, la única manera de cambiar lo publicado en estas bases se deriva de alguna sentencia del Tribunal Electoral. Durante la creación y consolidación de esta base de datos se pueden realizar cotejos, en algunos casos el recuento de los paquetes electorales junto al conteo de los votos.

Capítulo 4

Segmentación a través del análisis de *clusters*

Uno de los análisis de mayor relevancia en los estudios de investigación de mercados y opinión pública, consiste en la segmentación de los individuos, con el objetivo de encontrar perfiles que permitan obtener resultados particulares a cada partición en la población de estudio, de esta manera se pueden generar estrategias y recomendaciones orientadas a un sector en particular, lo que beneficia al ahorro de los costos de implementación. Para segmentar a los elementos o variables, existen diversas técnicas que logran este objetivo, aunque la mayoría de ellas sin sustento estadístico.

El análisis de *clusters* o *cluster analysis* es un método multivariado, que permite obtener una clasificación de los individuos en la población de estudio, con base en las variables y su comportamiento para cada individuo. El objetivo es crear grupos que sean homogéneos al interior del grupo y heterogeneos entre sí, se crean grupos tan diferentes como sea posible; para lograr esto el análisis se basa en la similaridad de los casos que se estudian, es decir, la distancia que existe entre ellos y entre las variables, teniendo algunas que describen a los individuos de mejor manera.

El análisis de *clusters* se usa cuando se quiere realizar particiones en los informantes, el investigador obtiene evidencia para pensar que existen características con mediciones muy diferentes entre sí, por ejemplo, supongamos que tenemos información sobre el uso de las tecnologías de la información y la comunicación, hábitos de los consumidores sobre éstos, que incluyen el tiempo de uso, las redes sociales favoritas, etc. Además de información que caracteriza a los informantes como edad, sexo, ciudad que habita, por mencionar algunas; se puede ver que hay diferencias naturales y sustanciales en la información basada en la edad del informante y el lugar que habita, por lo que estas diferencias llevarían a pensar en particiones de interés. Cabe señalar, que para garantizar una segmentación adecuada cada elemento debe pertenecer a un solo grupo, todos los elementos son clasificados en alguno de los segmentos y cada grupo comparte características similares entre sus elementos.

Otra razón para usar este tipo de análisis, está en la necesidad del investigador por jerarquizar

a los elementos; es decir, crear particiones dentro de los individuos, de tal forma que los grupos superiores contienen a los inferiores; por ejemplo, si buscamos conocer las palabras más usadas en *facebook* se pueden jerarquizar y agrupar en un dendrograma.

En ocasiones no se busca clasificar a los individuos, sino a las variables; esto permite agruparlas y disminuir la dimensión de las variables, facilitando el análisis de la información, de esta manera se crean indicadores que resumen a un grupo de variables. Esta clasficación, es similar a la realizada en un análisis factorial e incluso tiene aplicaciones similares al discriminante.

Finalmente, este tipo de análisis genera particiones y clasifica tanto a individuos como a variables. Dentro de los estudios electorales, se suele partir de un muestreo estratificado, por lo que no resulta raro pensar que un análisis de este tipo puede generar una partición que pueda usarse como la estratificación dentro del diseño de muestreo; generalmente, los estudios electorales, usan una estratificación de las secciones electorales definida por el INE en tres niveles: urbano, mixto y rural; sin embargo, este tipo de selección genera sesgos a la información al destinar parte importante de la muestra a unidades que no necesariamente aportan información valiosa. Por esta razón la propuesta consiste en definir una segmentación con base en los estudios socioeconómicos y demográficos públicos, que se disponen.

Para realizar un análisis de *clusters*, se pueden usar métodos tanto jerárquicos como no jerárquicos; como mencionamos anteriormente, la clasificación de información se basa en las similitudes que existen en las variables y los individuos, por lo que resulta importante definir algunos conceptos respecto a las distancias y similitudes.

4.1. Medidas de proximidad

Cuando se realiza un estudio, particularmente en temas de investigación de mercados y opinión pública, la información contiene variables que usan diferentes escalas de medición, incluso se mezclan variables ordinales, nominales, cualitativas y cuantitativas, por lo que realizar mediciones entre ellas, implica usar distintos tipos de distancias y similitudes.

Para identificar las agrupaciones que se pueden realizar a partir de un análisis de clusters, necesitamos conocer cómo se puede medir la distancia entre dos unidades muestrales, además, saber cuando dos *clusters* pueden ser agrupados en uno solo; al segundo se le conoce como similitud o similaridad.

4.1.1. Distancias

Se conoce como **distancia** a una función $d : \mathbb{R}^p \times \mathbb{R}^p$ que mide qué tan lejos se encuentra un punto de otro (o un individuo de otro), ésta es una medida de proximidad que describe las disimilaridades entre dos puntos. Sea $x = (x_1, x_2, ..., x_p)$, $y = (y_1, y_2, ..., y_p)$ y $z = (z_1, z_2, ..., z_p)$ tres puntos en $\mathbb{R}^p$, se le conoce como distancia a una función d que cumple:

1. $d(x,y) \geq 0$
2. $d(x,y) = 0$ si y sólo si $x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)$ para todo x,y,z en $\mathbb{R}^p$

Las distancias, más usadas en estadística son: **Distancia euclidiana:** es la función de disimilitud más usada, se define como sigue:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$$
$$d^2(x,y) = (x-y)'(x-y)$$

Distancia de Mahalanobis: a partir de la distancia euclidiana, se introduce la matriz S^{-1}, que es la matriz de varianzas y covarianzas, esta matriz es simétrica y positiva; a esta distancia se le conoce como distancia de Mahalanobis:

$$d^2(x,y) = (x-y)'S^{-1}(x-y)$$

Distancia de Minkowski: se le llama así a la generalización de la distancia euclidiana, la función se describe a continuación:

$$d(x,y) = ((x_1 - y_1)^\alpha + (x_2 - y_2)^\alpha + ... + (x_p - y_p)^\alpha)^{\frac{1}{\alpha}}$$

donde α es un entero positivo.

Distancia *City block*: se trata de la función anterior cuando α es igual a 1:

$$d(x,y) = (x_1 - y_1) + (x_2 - y_2) + ... + (x_p - y_p)$$

Distancia de Chebychev: cuando α es igual a ∞:

$$d_\infty(x,y) = max_{i=1,...,p}|x_i - y_i|$$

Las distancias o dismilitudes se usan cuando se tienen variables cuantitativas.

4.1.2. Similitudes o similaridades

Al igual que las distancias, las similitudes son medidas de proximidad, a diferencia de las primeras, éstas definen qué tan cerca está un punto de otro, es decir, qué tan cercanos son dos puntos o individuos. Sea $x = (x_1, x_2, ..., x_p)$, $y = (y_1, y_2, ..., y_p)$ y $z = (z_1, z_2, ..., z_p)$ tres puntos en $\mathbb{R}^p$, se les llama similitudes o similaridades a las funciones de la forma $s : \mathbb{R}^p \times \mathbb{R}^p$ que cumplen:

1. $0 \leq s(x,y) \leq 1$
2. $s(x,y) = 1$ si y sólo si $x = y$

3. $s(x,y) = s(y,x)$

Una similaridad entre dos puntos, puede definirse a partir de su distancia como:

$$s(x,y) = \frac{1}{1+d(x,y)}$$

4.1.3. Similitudes con datos binarios

Para elementos con datos **binarios**, es decir, los valores de las variables son 0 ó 1, donde 1 significa éxito o presencia de alguna característica y 0 la ausencia; existen diferentes maneras para medir las similitudes entre individuos.

Sean x y y dos individuos, que pertenecen a la muestra, con m características medidas por variables dicotómicas:

a el número de variables que toman el valor 1 en cada característica.
b el número de variables que toman el valor 1 en la variable *i-ésima* y 0 en la *j-ésima* variable.
c el número de variables que toman el valor 0 en cada característica.
d el número de variables que toman el valor 0 en la variable *i-ésima* y 1 en la *j-ésima* variable.
$m = a+b+c+d$

Coeficiente de **Russell y Rao**, mide la probabilidad de que una variable sea 1 en ambos elementos de la muestra; sólo es válida en aquellos que contengan éxito o presencia de la característica en ambos elementos:

$$s(x,y) = \frac{a}{a+b+c+d} = \frac{a}{m}$$

Coeficiente de **parejas simples**, es la probabilidad de que una variable presente coincidencia en ambos elementos o unidades muestrales, aquí se incluye éxito-éxito y fracaso-fracaso:

$$s(x,y) = \frac{a+d}{a+b+c+d} = \frac{a+d}{m}$$

Coeficiente de **Jaccard**, es la probabilidad de que una variable sea 1 en los individuos; sin que se tomen en cuenta las coincidencias de fracaso o ausencia de la característica como parte del total:

$$s(x,y) = \frac{a}{a+b+c}$$

Coeficiente **Sorensen-Dice** o de **Czekanowski**, en este caso se trata de una medida donde se excluye por completo la coincidencia 0-0, donde la coincidencia 1-1 pesa el doble. Es una extensión del coeficiente de *Jaccard*, donde se compensa la ausencia de d:

$$s(x,y) = \frac{2a}{2a+b+c}$$

Otros coeficientes o medidas, usadas con menor frecuencia en este tipo de variables son:

- **Rogers-Tanimoto:** Es una extensión de las parejas simples, donde se pesan doblemente las diferencias, es decir, cuando se tiene 1-0 ó 0-1:

$$s(x,y) = \frac{a+d}{a+d+2(b+c)}$$

- **Kulcynski:** es el cociente entre coincidencias y no coincidencias:

$$s(x,y) = \frac{a}{b+c}$$

- **Ochiai:**

$$s(x,y) = \frac{a}{\sqrt{(a+b)(c+d)}}$$

Sólo por mencionar algunas, ya que existen distintas posibilidades considerando las combinaciones que se pueden dar.

4.1.4. Similitudes con datos cualitativos

Cuando se tienen variables cualitativas, hay varias posibilidades para obtener una medida de similitud, si se trata de variables ordinales pueden usarse las medidas de disimilitudes, es decir, las distancias entre individuos; si por el contrario, se tratan de variables nominales, se puede hacer uso de variables *dummy*, tantas como opciones de respuesta nominal se tenga por cada variable.

No obstante, la forma más simple de obtener una medida de similitud entre estas variables, es construir una tabla de contingencia y calcular una χ^2 para probar independencia; sin embargo, la prueba χ^2 depende del tamaño de la muestra (n), donde $\chi^2 \leq n$, ya que crece en la medida en que lo hace n, por lo tanto, se recurre a la contingencia cuadrática, definiendo la medida de similitud como:

$$s(x,y) = \sqrt{\frac{\chi^2}{n}}$$

4.1.5. Similitudes con datos en distintas escalas

Cuando se tienen datos en distintas escalas, se recurre al **coeficiente de similitud de Gower**. Esta medida, propuesta por Gower en 1971, tiene la particularidad de poder usar al mismo tiempo variables cuantitativas y cualitativas. Por tanto, si se tiene una similaridad alta, cercana a 1, querrá decir que los dos individuos evaluados tienen características homogéneas.

Se define $s_h(i,j)$ como el coeficiente de similaridad entre los individuos i y j en la variable *h-ésima*, donde s es una función simétrica, no negativa y con valores reales entre 0 y 1.

A partir de la similaridad entre individuos, se recurre una transformación para obtener una distancia entre ellos, la cual se define como: $d(x,y) = 1 - s(x,y)$; sin embargo, esta función no necesariamente cumple la propiedad de desigualdad triangular. Entonces, se define la distancia entre i y j como:

$$d(i,j) = \sqrt{a(1-s(i,j))}$$

donde a es un entero positivo.

El coeficiente de similitud de Gower se define como:

$$\delta^2(i,j) = 1 - s(i,j)$$

$s(i,j)$ se define como:

$$s(i,j) = \frac{\sum_{i=1}^{p}\left(1 - \frac{|x_{ih}-x_{jh}|}{G_h}\right) + a + \alpha}{p+q-d+r}$$

donde:
p es el número de variables cuantitativas.
q es el número de variables binarias o dicotómicas.
r es el número de variables cualitativas.
a es el número de coincidencias 1-1 en las q variables binarias.
d es el número de coincidencias 0-0 en las q variables binarias.
α es el número de las r variables cualitativas.
G_h es el rango de la *h-ésima* variable cuantitativa.

Con la transformación de las variables mixtas, a partir del uso de la similaridad $s(x,y)$ y la δ de Gower es posible realizar una clasificación de individuos en grupos heterogéneos entre sí y homogéneos en los elementos de cada grupo.

4.2. Métodos jerárquicos

Se conoce como método jerárquico al proceso de clasificación de *clusters* que tiene por objetivo formar grupos, a partir de uno nuevo o separar uno existente para originar dos más, de forma que se minimiza la distancia o se maximiza la similaridad. Dependiendo de si se unen grupos o se separan, los métodos pueden ser:

Aglomerativos, también llamados ascendentes. Este método consiste en crear grupos a partir de la unión de grupos, se comienza con n grupos, donde n es el número de unidades en la muestra, a partir de ahí se comparan uno a uno los individuos y según sus similitudes se unen o no a un individuo, formando un grupo conjunto o bien un grupo nuevo.

Disociativos o descendentes, en este caso se trata del proceso inverso al aglomerativo, es decir, se parte de 1 grupo de n elementos y se comparan los elementos del mismo, si se encuentran similaridades cercanas a 1 permanece en el grupo, en caso contrario se expulsa, iniciando un nuevo grupo.

En general, los métodos jerárquicos están basados en un proceso aglomerativo, ya que éste permite visualizar los elementos en un gráfico, este gráfico se llama **dendrograma** y es muy útil en este método, ya que de manera gráfica se identifica una clasificación óptima, o bien una clasificación adecuada para el investigador.

Dependiendo de la manera en que se consideran las distancias entre individuos, los *clusters* jerárquicos por procesos aglomerativos pueden seguir distintas estrategias:

- Distancia mínima o similitud máxima, también se le conoce a esta estrategia como **liga simple**. Para clasificar a los individuos, se toma la mínima de todas las distancias posibles entre elementos de cada *cluster*, de forma que se unen los dos *clusters* más cercanos, este proceso se repite k veces hasta formar $n-k$ *clusters*, la distancia entre dos *clusters* (C_i, C_j) con n_i y n_j elementos respectivamente, se define de la siguiente manera.

$$d(C_i, C_j) = min\{d(x_l, x_m)\} \quad l = 1, ..., n_i; \quad m = 1, ..., n_j$$

  donde x_l es un elemento de C_i y x_m es un elemento de C_j. De forma análoga se define la similitud máxima.

- Distancia máxima o similitud mínima, a este tipo de estrategia se le llama de **liga completa**, aquí se considera que la distancia entre *clusters* se define con la máxima distancia o mínima similitud entre dos elementos que pertenecen a *clusters* distintos, de forma que si la distancia es muy grande se considera que los *clusters* no se pueden unir, de la misma forma que el anterior, se unen los *clusters* con menor distancia. La distancia entre dos *clusters* (C_i, C_j) con n_i y n_j elementos respectivamente es:

$$d(C_i, C_j) = max\{d(x_l, x_m)\} \quad l = 1, ..., n_i; \quad m = 1, ..., n_j$$

  donde x_l es un elemento de C_i y x_m es un elemento de C_j. De forma análoga se define la similitud mínima.

- Distancia o similitud promedio, también llamada estrategia de la **liga promedio**; para ésta, se toman todas las distancias o similitudes posibles entre dos de los elementos de *clusters* distintos. A partir del promedio de estas distancias, se unen los dos *clusters* más cercanos, es decir, los *clusters* con menor promedio de distancias; la distancia entre dos *clusters* (C_i, C_j) con n_i y n_j elementos respectivamente, se define de la siguiente manera.

$$d(C_i, C_j) = \frac{1}{(n_i n_j)} = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} d(x_i, x_j)$$

- Método del **centroide**, para comenzar se identifican los centroides de cada *cluster*, un centroide es el vector de medias asociado a las variables en cada uno de los *clusters*. Supongamos que tenemos al *cluster* C_j con n_j elementos y al *cluster* C_i con n_i elementos, el *cluster* C_i está formado por k *clusters* (C_{ik}) cada uno con n_{ik} elementos respectivamente, si denotamos a m^j, m^{ik} con $k = 1, ..., n_{ik}$, como los centroides de cada cluster respectivamente, con

dimensión n (tamaño de la muestra), el centroide del *cluster* en notación vectorial se ve como sigue.

$$m^i = \frac{\sum_{k=1}^{n_k} n_{ik} m^{ik}}{\sum_{k=1}^{n_k} n_{ik}}$$

donde cada componente está definida por:

$$m_l^i = \frac{\sum_{k=1}^{n_k} n_{ik} m_l^{ik}}{\sum_{k=1}^{n_k} n_{ik}} \quad con \quad l = 1, ..., n$$

si $k = 2$ la distancia de C_i a C_j es la distancia euclidiana entre ambos *clusters*; sin embargo, puede usarse la distancia que convenga o corresponda.

Con esto, se toma la distancia entre los centroides y se repite el procedimiento, que en la liga sencilla, se toma la menor distancia entre centroides y se agrupan ambos *clusters* en uno, si no, se considera parte de un segundo *cluster*, y así sucesivamente.

- Método de la **mediana**, se trata de una variación del método de centroides, en este último se considera el número de individuos que forman cada *cluster*; en este método por el contrario, no se toma en cuenta esto, sino el número de *clusters*.

- Método **Ward**, en este caso se establece que la distancia entre dos *clusters* (C_i, C_j) es el incremento del valor total de la suma de los cuadrados de las diferencias dentro de cada *cluster*, es decir, de cada individuo al centroide del grupo.

$$\begin{aligned} d(C_i, C_j) &= \sum_{k \in C_i \cup C_j} ||x_k - m_{C_i \cup C_j}||^2 - \sum_{k \in C_i} ||x_k - m_{C_i}||^2 - \sum_{k \in C_j} ||x_k - m_{C_j}|| \\ &= \frac{n_i n_j}{(n_i + n_j)} ||m_{C_i} - \bar{x}_{C_j}||^2 \end{aligned}$$

En este método, se comienza con una suma de cuadrados igual a 0, ya que se parte de la hipótesis que cada elemento es un *cluster*, esta suma va creciendo a medida que se unen conjuntos.

Murtagh y Legendre mostraron que la suma de distancias al cuadrado de los grupos, puede calcularse a partir de los datos originales, en función de las observaciones. Por lo tanto, cuando se consideran los datos originales, se puede usar otra función para minimizar, como la suma de las trazas de las matrices o la suma de sus determinantes. Si dos grupos tienen la misma distancia (la mínima) a otro *cluster*, la primera en unirse será la de menor número de elementos.

A pesar de la nobleza de éstos métodos al permitir visualizar rápidamente la formación de *clusters*; la principal desventaja de los métodos jerárquicos, radica en su poca efectividad en muestras grandes, además, una vez que se realiza la clasificación es imposible separar a algún elemento del *cluster*, aún en indicaciones posteriores del algoritmo.

Por esta razón suele recurrirse a métodos no jerárquicos, donde se fija previamente el número de *clusters*.

4.3. Métodos no jerárquicos

La idea central de este tipo de métodos es realizar una clasificación, a partir de especificar el número de *clusters*; este tipo de análisis, se usa para clasificar elementos y no variables.

Es a partir de un número k fijo de *clusters*, que se realiza la clasificación de los elementos; los algoritmos comienzan con k puntos y se genera una serie de movimientos de estos puntos a otros, este paso se repite hasta encontrar el óptimo local de la función, con esta partición. Se suele usar la información adicional para determinar k; sin embargo cuando no se dispone de ella, existe una forma sencilla de conocer el número óptimo de *clusters*, aplicando el algoritmo seleccionado n veces e identificando el valor de k, donde se reduce significativamente la suma del total de las varianzas dentro de cada cluster, de tal forma que en $k+1$ la suma no es estadísticamente menor que en k; a éste forma se le conoce como el método del codo, debido a que se representan gráficamente las sumas totales de las varianzas entre clusters y se selecciona la k donde se forma un "codo".

Para clasificar a los elementos, se pueden intercambiar de un grupo a otro, o bien unir *clusters*, con base en las características que presentan, si la distancia de los centroides de los *clusters* es suficientemente pequeña, entonces pueden ser agrupados en una sola partición.

4.3.1. Método de *k-medias*

El más conocido y usado es el método de **k-medias**, el cual consiste en un algoritmo que comienza con un número fijo de conjuntos k, en cada paso siguiente, se reasignan los elementos minimizando la suma de distancias al cuadrado de éstos al centroide que les corresponde. Se parte de la definición tradicional de distancia:

$$d(x,m_c) = (x-m_c)'(x-m_c)$$

Donde $x = (x_1,...,x_p)$ y m_c es el centroide del *cluster* c. Se busca entonces, minimizar la suma de las distancias al cuadrado, es decir:

$$min\{d^2(x,m_c)\} = SCDG = min\{\sum_{c=1}^{k}\sum_{i\in c} d(x_i,m_c)^2\}$$

En este caso sólo se puede garantizar la existencia de un mínimo local, por lo que el método funciona mejor, cuando se realizan varias iteraciones.

En general, se puede resumir este método con el siguiente algoritmo:

1. Se fija un número k de *clusters*, y se seleccionan k puntos como centroides de los *clusters* iniciales. La manera de hacerlo puede variar, desde asignar aleatoriamente los elementos

a las particiones y tomar sus centros como los centroides iniciales; otra manera, es tomar los k puntos que sean más distantes entre sí, es decir, maximizar las distancias de todos los elementos, y escoger los primeros k de una lista ordenada de mayor a menor; por último, se pueden escoger k grupos, conforme a la experiencia del investigador o a la información que se tiene, de la misma manera se pueden seleccionar los k centros, que serán las unidades muestrales con características específicas que los convierten en perfiles *ad-hoc* para el estudio.

2. Una vez identificados los k centros, se calculan las distancias euclidianas de cada unidad a los centroides, y se les asigna el grupo de pertenencia donde ésta es menor, el paso se repite hasta por n veces y en cada paso se recalcula el valor del centro, según los elementos que contenga el *cluster* en ese paso. Se realiza tantas veces sea necesario, hasta que todos los elementos pertenezcan a alguna de las particiones.

3. Se define un criterio de efectividad del análisis (la varianza entre grupos es lo más grande posible, el vector de medias de cada grupo es significativamente diferente de los otros, lo que se puede realizar con una prueba z, etc.), y se realiza la reasignación de elementos, donde el criterio mejore secuencialmente.

4. Si no se puede mejorar el criterio, el proceso termina.

Para este reporte, se usará este método, pues se utiliza el total de secciones, además del total de municipios y el total de ciudadanos inscritos en la lista nominal. Pensar en un método jerárquico, a partir de un dendrograma no resulta de mucha utilidad; sin embargo, se realizará un análisis exploratorio jerárquico, usando las facilidades de algunos paquetes estadísticos.

4.3.2. Método de *k-medoides*

Es muy similar al método de *k-medias*; en ocasiones, se dispone únicamente de la matriz de distancias, por lo tanto, en lugar de usar los centroides se prefiere el uso de un elemento representativo, al que se le llama medoide, el objetivo es minimizar la función:

$$min\{g(x,m_c)\} = min\{\sum_{c=1}^{k}\sum_{i\in c} d(x_i,m_c)^2\}$$

Con m_c el medoide, en vez del centroide.

Estos dos métodos son los más recurrentes, pero además de ellos existen otros como el método de Forgy y variante de Jancey, donde se fija un conjunto de puntos semilla, como si funcionaran como los centroides o los medoides; otros incluyen un número semifijo k, a partir de este número, se realizan pruebas, variando este número hasta conseguir un *cluster* que no corresponde a los objetivos de la investigación, con lo que se selecciona el número de *clusters* anterior inmediato a esto; otro más es el llamado Isodata, es una combinación de varios métodos a partir de la asignación por centroides más cercanos.

4.4. Método de mezcla de normales

Se trata de una alternativa a los métodos anteriores; la clasificación parte del punto de vista bayesiano, y se obtiene la probabilidad de cada observación de pertenecer a cada uno de los *clusters*. Para conseguir esto se realiza una mezcla de distribuciones, siendo la mezcla de normales la más simple.

Sean x_j con $j \in 1,...k$ variables aleatorias independientes, con distribución normal y parámetros desconocidos μ_j, σ_j^2 con $j \in 1,...k$. La mezcla de las k distribuciones normales, está dada por la función de densidad conjunta:

$$f(x) = \sum_{j=1}^{k} w_j f_j(x_j; \mu_j, \sigma^2{}_j)$$

Donde: w_j es la proporción de la j-ésima variable en la mezcla.
$0 \leq w_j \leq 1, \sum w_j = 1$ La función de verosimilitud de f(x) está dada por:

$$\prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f_j(x_j; \mu_j, \sigma_j^2)$$

Si consideramos a $\phi = (\phi_1, \phi_2, ..., \phi_n)$ como el vector de pertenencia a cada *cluster*, es decir si $\phi_i = j$ entonces, $x_i \in C_j$. Si $\theta_j = \mu_j, \sigma_j^2$, dado un ϕ y además con $f_j(x; \theta_j)$, las funciones de densidad de C_j; y θ_j los parámetros desconocidos de las funciones, la verosimilitud está dada por:

$$L(\theta_1, ..., \theta_k, \phi) = \prod_{j=1}^{k} f_j(x_j; \theta_{\phi j})$$

El método, consiste en seleccionar los parámetros θ y ϕ de tal forma que se maximiza la función de verosimilitud; donde ϕ es el número de *clusters*.

La función logverosimilitud, es con una de las cuales se puede obtener la solución máxima, aquí se sustituyen las medias poblacionales por las medias estimadas y la matriz de covarianzas asociada. Finalmente, se usan distintas formas de covarianza y se selecciona la que obtenga un menor BIC (Criterio Bayesiano de Información)[1].

[1] Análisis de datos multivariantes. Daniel Peña (2002)
Capítulo XI Métodos de inferencia avanzada multivariante

Capítulo 5

Muestreo

El muestreo es el procedimiento que se realiza para seleccionar un conjunto de observaciones de una población de interés, o bien una parte del universo de estudio. Una muestra permite que se pueda generar conocimiento e información a un costo accesible, además de que los resultados son altamente confiables y su recopilación ocurre con rapidez, en comparación a otro tipo de recolección de datos.

Se entiende por población de interés al conjunto de elementos (finito) que se pueden reconocer por compartir características similares e inherentes a los objetivos de una investigación, estos elementos comparten, entre otras, características de identificación en tiempo y espacio.

Una muestra es evaluada positivamente si es representativa de la población, es decir, que las variables de interés tienen una distribución semejante a las de la población, para ello el muestreo identifica el tamaño adecuado respecto al presupuesto y la información disponible.

5.1. Conceptos importantes

- **Población:** Es el conjunto de elementos con algunas características similares que permiten su identificación como parte del conjunto. Estas características incluyen identificación de los elementos referenciada geográficamente; la población objetivo comprende el total de los elementos que se desean estudiar, mientras que la población muestreada es el conjunto de elementos de donde se selecciona la muestra. Una población objetivo puede ser finita o infinita según el número de elementos que pertenecen a ella.

- **Unidad de Muestreo:** Se trata del conjunto de elementos donde se realiza la muestra, estas unidades pueden ser seleccionadas o no, en caso positivo se conocen como unidades muestrales o unidades en muestra; derivado de la etapa de selección pueden ser unidades primarias, secundarias, terciarias, finales, etc.

- **Marco muestral o de muestreo:** Se conoce así a los recursos que recopilan información que permite identificar a todas las unidades de muestreo, en términos generales describe en

alguna medida a la población objetivo.

- **Distribución Muestral:** Es la función de distribución asociada a un estimador.
- **Error muestral o de muestreo:** Es la diferencia entre las estimaciones derivadas de la muestra con los valores reales de las variables en la población objetivo; este error es calculado a partir del diseño de muestreo, debido a que *a priori* no se conoce el resultado en la población objetivo.

Existen otro tipo de errores que no se derivan del diseño involucrado, como la no respuesta o la falsedad en las respuestas; además de la sustitución arbitraria de los elementos contenidos en la muestra. Para minimizar este tipo de errores, se deben considerar algunas exigencias de un buen estudio mediante encuestas, el diseño del instrumento por ejemplo, resulta de suma importancia, se debe tener un diseño apropiado del cuestionario, la redacción de las preguntas, el orden, definiciones concretas y administración de las preguntas abiertas; la realización de una prueba piloto, también puede ayudar a disminuir estos errores; logística del levantamiento que incluya a supervisores, encuestadores, horarios establecidos, así como revisión de cuestionarios y verificación de la entrevista; finalmente el correcto análisis e interpretación de la información.

5.2. Los estimadores derivados del muestreo

Los resultados de una investigación realizada por medio de encuestas derivada de un muestreo, se obtienen a partir de estimadores, que varían de muestra a muestra, definidos teóricamente por una distribución muestral. Como todos los estimadores estadísticos, se busca que cumplan con algunas características y propiedades.

Insesgado; se trata de una propiedad clave en los estimadores obtenidos del análisis y se define como la característica donde la esperanza del estimador es igual al parámetro. Además, se busca que el estimador tenga varianza mínima, es decir, la distancia de los valores del estimador sea próxima a su media.

Como consecuencia del Teorema Central del Límite[1] se sabe que para alcanzar una distribución similar a la normal con los posibles valores del promedio muestral, se necesita una n suficientemente grande; no obstante, la velocidad de convergencia depende de la distribución de la variable de interés.

[1] Sea $X_1, X_2, ..., X_n$, un conjunto de variables aleatorias idénticamente distribuidas, cada una con distribución F, $E(X_k) = \mu < \infty$ y $var(X_k) = \sigma^2 < \infty$ Sea:

$$S_n^* = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}}$$

Y sean a y b dos números cualesquiera, donde $a < b$, entonces:

$$\lim_{n\to\infty} P[a < S_n^* < b] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx$$

i.e. $\frac{\sum_{k=1}^{n}}{n} \sim N(\mu, \frac{\sigma^2}{n})$

En la población de estudio con una variable que tiene un parámetro θ y a partir del diseño de una muestra, se pueden obtener muchos valores estimados del parámetro, a los que llamamos $\hat{\theta}$. Por el Teorema Central del Límite:

$$E(\hat{\theta}) = \theta$$
$$Var(\hat{\theta}) = E\left[\hat{\theta} - E(\hat{\theta})\right]^2 = E\left[\hat{\theta} - \theta\right]^2$$

$$P\left[\theta - \delta \leq \hat{\theta} \leq \theta + \delta\right] = 1 - \alpha \equiv P\left[|\hat{\theta} - \theta| \leq \delta\right] = 1 - \alpha$$

Es decir, la probabilidad de una diferencia máxima δ entre θ y $\hat{\theta}$ es $1 - \alpha$.
δ se define como precisión muestral o de muestreo, también conocido como error de estimación.
$1 - \alpha$ es conocida como la confianza del muestreo.

5.3. Muestreo aleatorio simple

El muestreo aleatorio simple es el método que consiste en seleccionar n unidades (de muestreo) de una población de tamaño N, las N unidades poblacionales tienen la misma probabilidad de ser seleccionadas. Existen C_n^N muestras distintas en total y cada una tiene la misma probabilidad de ser seleccionada ($\frac{1}{C_n^N}$); en este caso se trata de un muestreo sin reemplazo, aunque puede ser también con reemplazo.

En un muestreo aleatorio simple (m.a.s) cualquier elemento de la población (U_j con $j = 1, ..., N$) tiene una probabilidad de selección de $\frac{1}{N}$ en alguna de las n extracciones. Entonces, la probabiidd de que una unidad U_j esté incluida en la muestra es $\frac{n}{N}$, esta probabilidad es definida como la probabilidad de inclusión de primer orden (π_j).

Usualmente, previo al análisis de resultados se lleva a cabo una ponderación de la muestra, usando factores de expansión o pesos muestrales, que es el inverso de la probabilidad de inclusión, es decir, $\frac{1}{\pi_j} = \frac{N}{n}$.

Bajo el esquema de muestreo aleatorio simple (m.a.s) algunos de los estimadores de mayor interés son los siguientes:

- **Estimador Insesgado de la Media:**

$$\hat{\overline{X}} = \sum_{i=1}^{n} \frac{x_i}{n} = \overline{x}$$

con varianza

$$Var(\overline{x}) = S_{\overline{x}}^2 = E(\overline{x} - \overline{X})^2 = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

donde

$$S^2 = \sum_{i=1}^{N} \frac{\left(X_i - \overline{X}\right)^2}{N-1}$$

a $(1-\frac{n}{N})$ se le conoce como el factor de corrección por finitud, es decir, por muestrear una población finita. Conforme a la teoría de inferencia, se conoce que un estimador de S^2 es:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Usando el teorema central del límite, pueden obtenerse intervalos de confianza para la media. El intervalo del $(1-\alpha)\times 100\,\%$ de confianza es

$$\left(\overline{x} - z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}, \overline{x} + z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}\right)$$

- **Estimador del Total:**

$$\hat{X} = N\overline{x} = \sum_{i=1}^{n} \frac{N}{n} x_i$$

notemos que $\frac{N}{n}$ es igual a $\frac{1}{\frac{n}{N}}$, es decir, ponderamos la muestra de acuerdo al tamaño de esta en comparación al tamaño de la población. con varianza estimada

$$s^2_{\hat{X}} = N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}$$

El intervalo de confianza del $100(1-\alpha)\,\%$ de confianza para X es:

$$\hat{X} \pm z_{1-\alpha/2}\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

- **Estimador de proporciones:**

$$\hat{P} = p = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

donde $x_i = \begin{Bmatrix} 0 & \text{si no es un acierto} \\ 1 & \text{si es un acierto} \end{Bmatrix}$. con varianza estimada

$$s^2_p = \left(1-\frac{n}{N}\right)\frac{p(1-p)}{n-1}$$

El intervalo del $100(1-\alpha)\,\%$ de confianza es:

$$p \pm z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{p(1-p)}{n-1}} + \frac{1}{2n}$$

a $\frac{1}{2n}$ se le conoce como factor de corrección por continuidad.

5.3.1. Cálculo del tamaño de una muestra

Para calcular el tamaño de muestra partiendo de un muestreo aleatorio simple, se comienza determinando el nivel de precisión δ (error absoluto), también puede definirse un error relativo ε, donde $\varepsilon = \frac{\delta}{\overline{X}}$.

Si el objetivo de la investigación es estimar la media, el tamaño de la muestra se obtiene iniciando con la definición de confianza, Definiendo un nivel de precisión δ y una confianza $1-\alpha$, a partir de las definiciones previas, tenemos lo siguiente:

$$P(|\bar{x}-\overline{X}|<\delta) = 1-\alpha$$
$$\Longrightarrow P(\bar{x}-\delta<\overline{X}<\bar{x}+\delta) = 1-\alpha$$

Un intervalo de confianza para la media poblacional, está definido por:

$$\left(\bar{x} - z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}} < \overline{X} < \bar{x} + z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}\right)$$

Por lo tanto:

$$\delta = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$
$$\Longrightarrow \delta = z_{1-\alpha/2}\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)s^2}$$

Despejando n, obtenemos

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}s}{\delta}\right)^2$$

Si no se conoce el tamaño de la población o bien, el tamaño de la población de estudio es infinito, el cálculo del tamaño de muestra se define de esta manera:

$$n_0 = \left(\frac{z_{1-\frac{\alpha}{2}}s}{\delta}\right)^2$$

Si por el contrario, se conoce el tamaño de la población o bien este es pequeño, la fórmula del cálculo del tamaño de muestra queda como sigue:

$$n = \frac{n_0}{1+\frac{n_0}{N}}$$

El valor de s que corresponde al estimador de S, se puede calcular por medio de estudios anteriores, o bien puede obtenerse a partir de una prueba piloto.

En general, en los estudios de opinión pública, particularmente estudios electorales, se busca estimar proporciones (los votos para cada candidato). Para comenzar, observemos que en este tipo de estudios, las variables de interés toman valores de 0 y 1, cada variable se distribuye como una función bernoulli, la función conjunta de todos los candidatos en consecuencia se distribuye multinomial; por lo tanto, una proporción tiene estas características.

Una variable aleatoria binomial, con parámetro p, tiene una media p y una varianza $p(1-p)$, por lo tanto se sustituye esta varianza en las fórmulas anteriores. Si no se conoce el tamaño de la población o bien, el tamaño de la población de estudio es infinito, el cálculo del tamaño de muestra se define de esta manera:

$$n_0 = \left(\frac{z_{1-\alpha/2} p(1-p)}{\delta} \right)^2$$

Si por el contrario, se conoce el tamaño de la población o bien este es pequeño, la fórmula del cálculo del tamaño de muestra queda de la siguiente manera:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Vale la pena mencionar que estas fórmulas parten de una aproximación a la normal, por lo tanto funcionan si $0.2 \leq p \leq 0.8$; partiendo del desconocimiento del parámetro, generalmente se asume una $p = 0.5$ porque la varianza se maximiza en este valor de p.

5.4. Muestreo estratificado

A veces existen ocasiones en las que se conoce que la población de estudio está dividida en conjuntos heterogéneos entre sí pero homogéneos en su interior. Cuando esto pasa, se recurre a un muestreo estratificado, la razón técnica es que con esta partición, se puede reducir la varianza de los estimadores, por lo que los resultados son más precisos.

Cuando ocurren estas particiones previas, se muestrea cada estrato por separado, además de poder hacer diferencias en las variables de interés en cada uno de ellos, siempre y cuando el tamaño de la muestra en cada estrato sea suficiente para realizar estimaciones. En el caso de los estudios electorales se suele realizar una estratificación de las secciones electorales, por el tipo definido por el INE (urbano, mixto, rural); sin embargo, en este trabajo se plantea una estratificación de la población a partir del municipio en el que habitan, de esta manera se puede diferenciar a los ciudadanos con mayor potencial de voto.

Un muestreo estratificado puede disminuir los costos, ya que administra la dispersión de la muestra a partir de la estratificación seleccionada. Otra ventaja de este tipo de muestreo está en la disponibilidad de los marcos muestrales, ya que éstos pueden ser independientes entre sí, es decir, las fuentes de información no tienen que ser iguales para todos los estratos. Se pueden usar distintos tipos de diseño muestral en cada estrato.

Consideremos L como el número de estratos.

N_h es el tamaño del estrato h, $h = 1,...L$.

$N = \sum_{h=1}^{L} N_h$ es el tamaño de la población.

Entonces en cada estrato definiremos X_{hi} como los valores observados en la *i-ésima* unidad muestral dentro del *h-ésimo* estrato, $i = 1,...,N_h$ $h = 1,...,L$. Siguiendo esta lógica tendremos una media poblacional por estrato $\overline{X}_h = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h}$.

$X_h = N_h \overline{X}_h$ es el total de la población en el estrato h.

$X = \sum_{h=1}^{L} X_h = \sum_{h=1}^{L} \sum_{i=1}^{Nh} X_{hi}$ es el total poblacional.

Por lo tanto:

$$\overline{X} = \frac{X}{N} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi}}{\sum_{h=1}^{L} N_h}$$

es la media poblacional.

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (X_{hi} - \overline{X_h})^2}{N_h - 1}$$

es la varianza poblacional, dentro del estrato h.

Cada estrato tiene diferente peso dentro de la muestra. Dependiendo del tamaño de la partición en la población de estudio, el peso de cada estrato en la muestra está dado por:

$$W_h = \frac{N_h}{N}, conh = 1...L$$

$$\Longrightarrow \sum_{h=1}^{L} W_h = 1$$

Cualquier tipo de muestreo se reduce en alguna de sus etapas a un muestreo aleatorio simple. La forma más sencilla del muestreo estratificado consiste en el uso:

n_h tamaño de la muestra en el estrato *h-ésimo*.

$n = \sum_{h=1}^{L} n_h$ es el tamaño total de la muestra.

$\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$ estimador de la media en el estrato *h-ésimo*.

$N_h \bar{x}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$ el estimador del total en el estrato *h-ésimo*.

El estimador del **total poblacional**, uno de los objetivos probables en los estudios electorales, está definido por:

$$\begin{aligned}\hat{X} &= \sum_{h=1}^{L} \hat{X}_h \\ &= \sum_{h=1}^{L} N_h \overline{x_h} \\ &= \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{N_h}{n_h} x_{hi}\end{aligned}$$

$\frac{N_h}{n_h}$ es el factor de expansión en cada estrato de la muestra.

Para calcular la varianza del estimador del total, tomaremos en cuenta que las muestras dentro de cada estrato en cuestión, son independientes entre sí. Por lo tanto la varianza del estimador está dado por:

$$\begin{aligned}S^2_{\hat{X}} &= \sum_{h=1}^{L} S^2_{\hat{X}_h} \\ &= \sum_{h=1}^{L} N_h^2 S^2_{\overline{x}_h}\end{aligned}$$

De tal manera (asumiendo un muestreo aleatorio simple en cada estrato), tenemos que la varianza del estimador del total está dado por:

$$S^2_{\hat{X}} = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

Por tanto el estimador de la varianza, para el estimador de total, es:

$$s^2_{\hat{X}} = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

el estimador de la varianza dentro de cada estrato está dado por:

$$s_h^2 = \frac{\sum_{i=1}^{n_h} (x_{hi} - \overline{x}_h)^2}{n_h - 1}$$

Si el tamaño de la muestra dentro de cada estrato es lo suficientemente grande, la distribución del estimador del total se distribuye como una variable normal, en caso contrario puede aproximarse a una t de *Student*; sin embargo, todos los estratos deben tener las mismas condiciones de aproximación para generar un intervalo de confianza; supongamos que la muestra es grande en cada estrato,

el intervalo de confianza al $(1-\alpha)\times 100\,\%$ para el estimador del total poblacional es:

$$\hat{X} \pm z_{1-\alpha/2}\sqrt{\sum_{h=1}^{L} N_h^2 \left(1-\frac{n_h}{N_h}\right)\frac{\frac{\sum_{i=1}^{n_h}(x_{hi}-\bar{x}_h)^2}{n_h-1}}{n_h}}$$

El estimador de la **media poblacional** en un muestreo estratificado está dado por:

$$\begin{aligned}\hat{\bar{X}} &= \frac{\hat{X}}{N} \\ &= \sum_{h=1}^{L}\frac{N_h}{N}\bar{x}_h\end{aligned}$$

Recordemos que a $\frac{N_h}{N}$ se le conoce como el peso del estrato h, también conocido como ponderador. $\hat{\bar{X}}$ es la suma ponderada de los promedios muestrales en cada estrato.

La varianza del estimador de la media poblacional está dado por:

$$\begin{aligned}S^2_{\hat{\bar{X}}} &= var\left(\sum_{h=1}^{L}\frac{N_h}{N}\bar{x}_h\right) \\ &= \sum_{h=1}^{L}\left(\frac{N_n}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}\end{aligned}$$

Para encontrar el estimador de la varianza del estimador de la media poblacional, usamos el estimador de la varianza en cada estrato de manera independiente, obteniendo:

$$s^2_{\hat{\bar{X}}} = \sum_{h=1}^{L}\frac{N_h}{N}^2\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

Asumimos que el tamaño de la muestra en cada estrato es lo suficientemente grande, para poder aproximar la función de distribución del estimador a una normal, entonces el intervalo de confianza al $(1-\alpha)x100\,\%$ está dado por:

$$\hat{\bar{X}} \pm z_{1-\alpha/2}\sqrt{s_{\hat{\mu}}}$$

Finalmente el estimador de una **proporción**, análogo a los resultados anteriores y del muestreo aleatorio simple es:

$$\hat{P} = p = \sum_{h=1}^{L}\frac{N_h}{N}\sum_{i=1}^{n_h}\frac{x_{hi}}{n_h} \text{ donde } x_{hi} = \{0,1\}$$

La varianza del estimador de proporciones, es por consiguiente:

$$S_p^2 = \sum_{h=1}^{L}\frac{N_h}{N}^2\left(1-\frac{n_h}{N_h}\right)\frac{p_h(1-p_h)}{n_h}$$

el estimador de esta varianza es:

$$s_p^2 = \sum_{h=1}^{L} \frac{N_h}{N}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{p_h(1-p_h)}{n_h}$$

El intervalo de confianza al $(1-\alpha)\times 100\,\%$, asumiendo normalidad del estimador, es:

$$p \pm z_{1-\alpha/2} \sqrt{\sum_{h=1}^{L} \frac{N_h}{N}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{p_h(1-p_h)}{n_h}}$$

5.4.1. Distribución de la muestra entre los estratos

Uno de los elementos a considerar en este tipo de muestreo, es el tamaño que se le debe asignar a cada estrato en la muestra. Supongamos que tenemos una muestra de tamaño n, definido inicialmente por el investigador, sea C_h el costo de acceder a la información de alguna unidad en el estrato h; la función de costo, está definida por:

$$costo = C_0 + \sum_{h=1}^{L} C_h n_h$$

Usando una de las fórmulas obtenidas por Cochran, la varianza del estimador de la media, se minimiza con un costo total fijo:

$$n_h = n \frac{N_h S_h}{\sqrt{C_h}} \frac{1}{\left[\sum_{k=1}^{L} \frac{N_k S_k}{\sqrt{C_k}}\right]}$$

$$n_h \propto \frac{N_h S_h}{\sqrt{C_h}}$$

Lo que significa que en el estrato la muestra es más grande si éste es más grande, el estrato es más variable y si el costo es menor.

A esta distribución se le conoce como la distribución óptima; además de ésta, existen dos distribuciones que pueden ser usadas.

La distribución de Neyman, considera que los costos C_h son constantes en todos los estratos, entonces:

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

La distribución proporcional considera que tanto los costos como las varianzas son constantes en todos los estratos, por lo tanto:

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

una de las ventajas de este tipo de distribución es que produce muestras autoponderadas, lo que quiere decir que:

$$\frac{n_h}{N_h} = \frac{n}{N}$$
$$\Longrightarrow \frac{N_h}{n_h} = \frac{N}{n}$$

$\frac{N}{n}$ es el factor de expansión asociado.

El tamaño de la muestra depende de la distribución que el investigador decida usar, generalmente se usa una distribución proporcional, por la ventaja de incidir en menor medida en los factores de expansión, obteniendo resultados más precisos

5.5. Muestreo por conglomerados

Se conoce como conglomerado a un conjunto de elementos de la población; en el caso del muestreo, son unidades que comparten características en tiempo y espacio, por lo que resulta más simple la recolección de los datos.

El muestreo por conglomerados usa el principio del muestreo aleatorio; las unidades a muestrear son precisamente los conglomerados. Este tipo de muestreo es usado debido a que no existen marcos muestrales de la población, o bien es mucho más cara la construcción de ellos que el estudio en sí, o simplemente es imposible construirlos; además, este tipo de muestreo permite que el proceso sea menos costoso que un muestreo aleatorio simple, sobre todo cuando la recolección de la información implica recorrer grandes distancias a lo largo del territorio donde se realiza el estudio.

Los conglomerados pueden existir previamente definidos por el marco muestral o pueden construirse con base en características definidas por el investigador. Una desventaja de este tipo de muestreo, es que los elementos dentro del conglomerado pueden estar correlacionados, priorizando la varianza entre conglomerados.

El estimador del **total poblacional** está dado por:

$$\hat{X} = N\hat{\bar{X}}$$
$$= \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i} x_{ij}$$

Donde:

$X =$ total poblacional.
$N =$ número de conglomerados en la población.
$n =$ número de conglomerados en la muestra.

$M_i =$ número de unidades en el *i-ésimo* conglomerado.
$x_{ij} =$ valor de la variable en la *j-ésima* unidad del *i-ésimo* conglomerado.

El estimador de la varianza del estimador del total poblacional es:

$$s^2_{\hat{X}} = N^2 \left(1 - \frac{n}{N}\right) \frac{\frac{1}{N-1} \sum_{i=1}^{N} (\sum_{j=1}^{M_i} x_{ij} - \overline{x})^2}{n}$$

Para la **media poblacional**, supongamos que conocemos el tamaño de la población de estudio, es decir, el número de unidades de la población, entonces se puede obtener un estimador de la poblacional por elemento:

$$\hat{\mu}_e = \overline{x}_e \frac{N}{M} \sum_{i=1}^{n} \frac{x_i}{n}$$

El estimador de la varianza del estimador de la media por elemento es:

$$s^2_{\hat{\mu}_e} = \frac{1}{M^2} N^2 \left(1 - \frac{n}{N}\right) \frac{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2}{n}$$

El estimador del total poblacional y el de la media por elemento, son estimadores insesgados; sin embargo, la varianza en ambos casos es grande, debido a que el número de elementos en cada conglomerado difiere entre sí, lo que provoca que exista variabilidad entre los totales de los conglomerados; por lo tanto, se usan estimadores de razón, siempre y cuando el tamaño del conglomerado *i-ésimo* esté relacionado con el total del conglomerado.

El estimador por elemento o de **razón** para la media es:

$$\begin{aligned} \hat{\mu}_r = \overline{x}_r &= \frac{\frac{N}{n} \sum_{i=1}^{n} x_i}{\frac{N}{n} \sum_{i=1}^{n} M_i} \\ &= \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \end{aligned}$$

El estimador de la varianza para el estimador de razón de la media es:

$$s^2_{\hat{\mu}_r} = s_{\overline{x}_r} = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{\left(\sum_{i=1}^{n} \frac{M_i}{n}\right)^2} \sum_{i=1}^{n} \frac{(x_i - \overline{x}_r Mi)^2}{n-1}$$

El estimador de **razón para el total poblacional** es:

$$\hat{X} = M\left(\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i}\right)$$

la varianza estimada asociada a este estimador es:

$$s^2_{\hat{X}} = M^2 s_{\bar{x}_r} = M^2 \left(\left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{\left(\sum_{i=1}^{n} \frac{M_i}{n}\right)^2} \sum_{i=1}^{n} \frac{(x_i - \bar{x}_r Mi)^2}{n-1} \right)$$

El estimador de **razón para proporciones** es:

$$\hat{P} = p = \frac{\sum_{1=1}^{n} x_i}{\sum_{i=1}^{n} M_i}$$

con varianza estimada:

$$s^2_p = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{\left(\sum_{i=1}^{n} \frac{M_i}{n}\right)^2} \sum_{i=1}^{n} \frac{(x_i - pM_i)^2}{n-1}$$

Para determinar el tamaño de la muestra, se realiza de forma análoga al muestreo aleatorio simple, tomando en consideración que lo que se muestrea son los conglomerados, por lo tanto lo que se obtendrá es el número de conglomerados en muestra. El tamaño de la población de estudio es el número de conglomerados en total que componen dicha población.

5.6. Muestreo sistemático

A diferencia de los diseños anteriores que dependen de una partición de la población, el muestreo sistemático se diferencia por la forma en la que se seleccionan las unidades de muestreo. En este tipo de muestreo, la primer unidad en muestra se selecciona aleatoriamente, a partir de ahí el resto se seleccionan sistemáticamente; es decir, se seleccionan con la determinación de un salto. A partir de la unidad seleccionada aleatoriamente, se cuentan k unidades y se selecciona como parte de la muestra la unidad $k+i$ y así sucesivamente.

Sea k el intervalo del muestreo sistemático (el intervalo de muestreo), entonces:

$$k = \frac{N}{n}$$

donde N corresponde al tamaño de la población y n corresponde al tamaño de la muestra.

Para comenzar, se selecciona un número aleatorio i, de tal manera que $1 \leq i \leq k$. La muestra está formada por unidades (U) que se ven de la siguiente manera:

$$U_i, U_{i+k}, U_{i+2k}, ..., U_{i+k(n-1)}$$

Podemos observar que se selecciona una unidad cada k unidades. Existen entonces $k = \frac{N}{n}$ muestras posibles, cada una con probabilidad $\frac{1}{k}$ de ser seleccionada. Entonces con probabilidad $\frac{n}{N}$ de que la

unidad *i-ésima* esté en muestra, definida como:

$$\pi_i = \frac{n}{N}$$

Por tanto

$$\pi_{ij} = P(U_i, U_j \text{estén en la muestra})$$

Esta probabilidad es 0 si tanto la unidad i y como la j no pertenecen al mismo conglomerado, $\frac{n}{N}$ si sí.

La ventaja de este tipo de muestreo es que su selección es más sencilla que en un m.a.s; sobre todo cuando la selección de la última unidad de muestreo se realiza durante el levantamiento de la información; además, se evita la concentración de la muestra en alguna zona en particular, por tanto la dispersión de la muestra permite tener mayor cobertura de forma uniforme. No obstante, si no se define una numeración homogénea de las unidades en el marco muestral, ocasiona que la muestra no sea correcta para su análisis; además, no se pueden calcular estimadores de la varianza con una sola muestra hecha por este método.

Si no se conoce el orden de las unidades, o bien no se realiza una numeración con criterios previamente estudiados, el muestreo es equivalente a un m.a.s; Cuando el orden corresponde a una relación con las variables de estudio, este tipo de muestreo genera varianzas menores en los estimadores que los que se obtienen con un m.a.s; si por el contrario las unidades tienen un orden con cambios periódicos, y el periodo coincide con k, se producen varianzas mayores de los estimadores que los obtenidos con el m.a.s.

El estimador de la media poblacional es:

$$\hat{\mu}_s = \bar{x}$$

La varianza es:

$$s_s{}^2 = \frac{k-1}{k} S_b^2$$

S_b^2 es la varianza muestral entre conglomerados. La varianza no se puede estimar, por lo que se usa la varianza estimada por m.a.s.

Para realizar un muestreo sistemático se siguen estos dos pasos:
1. Se selecciona aleatoriamente una unidad entre 1 y N.
2. Cada k-ésima + 1 unidades, se selecciona la que corresponde, hacia adelante hasta llegar a N y hacia atrás hasta llegar a 1.

5.7. Muestreo polietápico

Un muestreo polietápico es aquel que está formado por diferentes etapas de selección.

Cada etapa de selección puede tener su propio diseño de muestreo, por lo que el cálculo de sus

estimadores se realiza anidando las funciones, según el tipo de muestreo que se tiene en cada una de las etapas y tantas etapas de selección se tengan. Las unidades de la primera selección se denominan Unidades Primarias de Muestreo (UPM), en la siguiente selección se llaman Unidades Secundarias de Muestreo (USM) y así hasta la última selección, donde se nombran Unidades Últimas de Muestreo (UUM).

Un ejemplo de este tipo de muestreo, sería uno que comienza con la estratificación de la población; supongamos que tenemos 2 estratos, en el primer estrato, realizaremos un muestreo por conglomerados y en cada conglomerado realizaremos un muestreo sistemático con arranque aleatorio. En el segundo estrato, realizaremos un muestreo sistemático de conglomerados y en cada conglomerado aplicaremos un muestreo aleatorio simple.

En el ejemplo anterior, podemos ver que el esquema de muestreo tiene cuatro y cinco etapas de selección respectivamente en cada estrato para llegar a las últimas unidades de muestreo, que son aquellas unidades objetivo de la investigación.

Como se mencionaba anteriormente, los estimadores están definidos por funciones dependientes, tantas como etapas se tengan, el punto donde se evalúa la función corresponde a la etapa inmediata anterior. La varianza de estos estimadores corresponde a la suma de la varianza del estimador en cada etapa de selección. Para entender de mejor manera cómo se realiza un muestreo polietápico, usaremos el caso particular con 2 etapas (bietápico).

5.7.1. Muestreo bietápico

En un muestreo bietápico existen 2 momentos en los que se realiza una selección de las unidades muestrales; en los estudios electorales por ejemplo, existe una primera etapa de selección aleatoria de secciones electorales y en cada sección electoral se eligen aleatoriamente los informantes de la encuesta, que serían ciudadanos inscritos en la lista nominal del INE.

Sea:
N el número de unidades primarias de muestreo.
M_i el número de unidades secundarias en la unidad i-ésima primaria de muestreo.
M es el número de unidades secundarias de muestreo.
n el número de unidades primarias de muestreo en muestra.
m_i el número de unidades secundarias de muestreo.
y_{ij} la observación de la unidad j-ésima secundaria de muestreo en muestra de la unidad i-ésima primaria de muestreo en muestra.

El estimador del total poblacional es:

$$\begin{aligned}\hat{X} =& \frac{N}{n}\sum_{i=1}^{n} M_i \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij} \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{m_i}{M_i}\frac{n}{N} x_{ij}\end{aligned}$$

En este caso $\frac{m_i}{M_i}\frac{n}{N}$ es el factor de expansión. El estimador de la varianza para el estimador del total poblacional es:

$$\hat{S}^2_{\hat{X}} = N^2(\frac{1}{n}-\frac{1}{N})\frac{1}{n-1}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i} x_{ij} + \frac{N}{n}\sum_{i=1}^{N} M_i^2(\frac{1}{m_i}-\frac{1}{M_i})\frac{1}{m_i-1}\sum_{j=1}^{m_i}(x_{ij}-\overline{x_i})^2$$

La primer parte de la ecuación anterior corresponde a la varianza de las unidades primarias de muestreo, lo cual significa entre el 90 % y 95 %. Por lo tanto un intervalo del $(1-\alpha)X100\,\%$ para el total poblacional X es: $\hat{X} \pm z_{1-\frac{\alpha}{2}}\sqrt{\hat{S}^2_{\hat{X}}}$.

El estimador de razón para la media, está dado por:

$$\hat{\overline{X}}_r = \frac{\sum_{i=1}^{n} M_i \overline{x}_i}{\sum_{i=1}^{n} M_i}$$

El estimador de la varianza para este estimador es:

$$\hat{S}^2_{\hat{\overline{X}}_r} = (1-\frac{n}{N})\frac{1}{n\sum_{i=1}^{n}\frac{M_i}{n}^2}\sum_{i=1}^{n}\frac{M_i^2(\overline{x_i}-\frac{\sum_{i=1}^{n} M_i\overline{x}_i}{\sum_{i=1}^{n} M_i})^2}{n-1} + \frac{1}{nN\sum i=1^n\frac{M_i}{n}^2}\sum_{i=1}^{n} M_i^2(1-\frac{m_i}{M_i})(\frac{\frac{1}{m_i-1}\sum_{j=1}^{m_i}(x_{ij}-\overline{x}_i)^2}{m_i})$$

El estimador de una proporción para un muestreo bietápico es:

$$\hat{P} = \frac{\sum_{i=1}^{n} M_i \sum_{j=1}^{m_i}\frac{x_{ij}}{m_i}}{\sum i=1^n M_i}$$

x_{ij} es 1 para las observaciones donde se presenta la característica o el éxito en nuestra investigación. El estimador para la varianza del estimador de proporción está dado por:

$$\hat{S}^2_{\hat{P}} = (1-\frac{n}{N})\frac{1}{n\sum_{i=1}^{n}\frac{M_i}{n}}\frac{\sum_{i=1}^{n} M_i^2(\sum_{j=1}^{m_i}\frac{x_{ij}}{m_i}-\hat{P}}{n-1} + \frac{1}{nN\sum_{i=1}^{n}\frac{M_i}{n}}\sum_{i=1}^{n} M_i^2(1-\frac{m_i}{M_i})(\frac{\sum_{j=1}^{m_i}\frac{x_{ij}}{m_i}(1-\sum_{j=1}^{m_i}\frac{x_{ij}}{m_i})}{m_i-1})$$

Como podemos apreciar, los cálculos para los estimadores en un muestreo bietápico no son nada simples, por lo que al aumentar el número de etapas, aumenta la complicación de los cálculos de los estimadores. En este trabajo, como veremos más adelante, partimos de un diseño polietápico, por tanto para verificar nuestras hipótesis realizaremos distintas simulaciones, con el fin de obtener una densidad característica en cada estimado.

5.8. Propuesta de selección de la muestra

En general, para cualquier estudio electoral se usa el mismo diseño de muestreo, sin importar el tipo de estudio que se está realizando, la principal diferencia radica en las unidades últimas de muestreo, pues en los estudios previos a la jornada electoral se conoce el tamaño final de la muestra desde un inicio; por el contrario, durante la jornada electoral se desconoce con certeza el tamaño final de la muestra, por lo que el error asociado varía conforme se tienen informantes, mientras que en los conteos rápidos el conglomerado asociado a la sección electoral en muestra es censado.

El presente reporte se centra en los estudios electorales que ocurren durante la campaña electoral, es decir, encuestas electorales. Cabe señalar que una encuesta electoral puede ser longitudinal o transversal, se dice que la encuesta es longitudinal cuando se tienen varios ejercicios en periodos homogéneos de tiempo, a este tipo de estudio se le conoce como *tracking* electoral, y funciona como una serie de tiempo en la que es posible distinguir tendencias y ciclos en la intención de voto; a partir de estos estudios es posible crear un escenario probable para los resultados finales; generalmente se realiza una rotación parcial de la muestra, para emular promedios móviles y que el movimiento de las tendencias siga una trayectoria suave. En el caso de los estudios transversales, se realizan sin seguir una periodicidad señalada, estos estudios no tienen continuidad y por lo tanto suelen no ser comparables entre sí, ya que cada uno tiene objetivos particulares, inherentes a la coyuntura del momento en que se levanta la información.

Sin importar el tipo de encuesta electoral que se realiza, el diseño de muestreo es igual en ambos casos. Se realiza un muestreo estratificado, por conglomerados, y dentro de los conglomerados una selección sistemática con arranque aleatorio. El muestreo se realiza en dos etapas principales, ya que en la primera se seleccionan los conglomerados que formaran parte de la muestra y posteriormente una selección del informante. La estratificación se basa en el tipo de sección electoral, según la definición del INE.

Este tipo de diseño muestral suele tener grandes ventajas, sobre todo en costos de operación y administración de las varianzas muestrales. En algunos casos se incluye una etapa más de selección, donde se muestrean las manzanas contenidas en cada sección electoral, otras incluyen una más, seleccionando las viviendas en muestra previo al levantamiento, en cuyo caso se genera un procedimiento similar a los estudios del INEGI, donde se visita la vivienda en muestra hasta 4 veces y en caso de no tener éxito, no hay sustituciones y se captura como un rechazo.

En el presente trabajo, se propone una alternativa en la estratificación de la muestra. El diseño seguiría siendo un muestreo estratificado (doblemente estratificado, una primera partición se hará según la circunscripción y se dividirá en partes iguales, la segunda será proporcional al tamaño del estrato), por conglomerados, con selección sistemática de los informantes dentro del conglomerado, arranque aleatorio para la primer unidad seleccionada. La estratificación se basará en la clasificación de los informantes según la sección electoral en la que estén inscritos, además del municipio en el que habitan o están registrados ante la lista nominal del INE.

Para clasificar sociodemográficamente a los informantes potenciales (ciudadanos inscritos en la lista nominal) usaremos la información disponible en el INEGI, donde se tienen indicadores a nivel municipal[2]. Dicha clasificación se realizará a partir de una segmentación por medio de un análisis de *clusters* de k-medias; se usarán variables socioeconómicas y demográficas que se encuentran hipotéticamente relacionadas con la manera de votar (sexo, edad, educación, situación de empleo, nivel de ingresos y acceso a internet).

En las encuestas electorales buscamos estimar la proporción de votos para cada candidato, se parte de un número fijo de entrevistas acotado por el presupuesto disponible; supondremos un tamaño de muestra de 1,200 entrevistas distribuidas en 150 secciones (conglomerados) seleccionadas, de dimensión 8 en cada caso. Con este diseño se tiene un margen de error asociado de $\pm 2.8\,\%$ al $95\,\%$ de confianza. Para definir el margen de error, se usa la fórmula para calcular el tamaño de la muestra en un diseño aleatorio simple, considerando la estimación de proporciones; la razón de usar m.a.s es porque en un muestreo estratificado el margen es menor o igual que el del aleatorio simple, más adelante realizaremos distintas simulaciones para verificarlo.

En cada circunscripción se tendrán 240 entrevistas (5 circunscripciones electorales, que tienen tamaño similar, según la división territorial hecha por el INE) distribuidas proporcionalmente a la segmentación definida.

[2]Los municipios son la partición más pequeña del país, que coincide con la división territorial del INE.

Capítulo 6

Resultados

6.1. Preámbulo

Uno de los principales problemas que enfrenta un investigador al realizar estudios electorales, radica en las limitaciones presupuestales; en general, el fondo asignado a cada estudio que pretende conocer la intención de voto ciudadana, solo permite entrevistar a un número muy limitado de personas, por lo que optimizar los recursos disponibles se vuelve prioritario. Con este trabajo buscamos aumentar la calidad de los informantes, disminuyendo la probabilidad de selección entre los ciudadanos con menor potencial de acudir a votar.

En 1960, en el libro *The American Voter*, A. Campbell, P. Converse, W. Miller y D. Stokes, desarrollaron el llamado modelo de Michigan, que habla acerca de la identificación partidista de los ciudadanos; en él, se asegura que la forma en la que un ciudadano decide su voto se basa en el *embudo de causalidad*, como resultado de estos estudios se pudo concluir que el voto no es más que un resumen, resultado de factores previos.

Los factores inherentes en el voto de cada ciudadano pueden ser de corto o largo plazo; dentro de los factores de corto plazo se encuentra el posicionamiento personal sobre temas coyunturales, como las propuestas, ideas o conductas de los candidatos, entre otras; estos factores influyen en la tendencia de la intención de voto. Por otro lado, se encuentran los factores de largo plazo, los cuales influyen en buena medida en la ideología personal; dichos factores son determinados por el lugar en el que vive cada persona, las características culturales y sociales de las que forma parte el ciudadano, como etnia o religión, también características demográficas, sociales y económicas como el sexo, edad, escolaridad, ocupación, nivel económico, etc. En este trabajo, nos centraremos en los factores de largo plazo, ya que previo al levantamiento de la información, durante el diseño muestral, se dispone de indicadores de alto valor informativo y análitico que describen las características mencionadas.

Con los factores de largo plazo se puede determinar la forma más probable en la que cada ciudadano votaría el día de la elección; pero, también se puede estimar la probabilidad de acudir a votar. Es decir, podemos obtener particiones de la población, que permitan perfilar a los ciudadanos con

mayor potencial de acudir a votar y diseñar una muestra que facilite encontrar precisamente a ese perfil (la probabilidad de encontrar a un informante óptimo aumenta al referenciar geográficamente las unidades muestrales según el número de ciudadanos en cada perfil).

La influencia de los factores de largo plazo, que se pueden expresar como indicadores sociodemográficos, es determinante. En los estudios realizados en 2012 para el Proyecto Comparativo de Elecciones Nacionales (CNEP por sus siglas en inglés) y el Centro de Estudios Sociales y de Opinión Pública de la Cámara de Diputados (CESOP), se determinó que el sexo del votante, influyó en el resultado final, las mujeres reflejaron una tendencia favorable al candidato del *PRI* y ligeramente positiva hacia la candidata del *PAN*, el estudio aseguraba que a mayor escolaridad, mayor era la afinidad por el candidato opositor de izquierda; además, la clase media mostró ser más empática (ligeramente mayor) a este mismo candidato, lo mismo que usuarios de internet y consumidores frecuentes de información. Ni qué decir de la regionalización del voto, pues en los estados del norte, el voto por la izquierda era significativamente menor que en el resto del país.

En el Estudio Censal sobre la Participación ciudadana de 2012 y 2015 del INE, se demostró que la participación de las mujeres es mayor a la de los hombres; en 2012, el porcentaje de participación ciudadana general alcanzó el 62.08 %, en mujeres fue de 66.08 % mientras que en hombres fue del 57.77 %. En 2015, la participación ciudadana fue de 47.07 %, la participación femenina alcanzó un 50.89 % superando por 8 puntos porcentuales a la participación masculina que fue de 42.95 %; por lo tanto, el sexo del ciudadano influye en la probabilidad de acudir a votar. Para fines de este trabajo, consideraremos el género como una variable dicotómica, donde toma el valor de 1 si es mujer, 0 si es hombre; la media de esta variable (porcentaje de mujeres por unidad) funciona como un indicador de género.

Respecto a la edad de los electores, en el estudio del INE en 2012, se muestra una mayor participación en los ciudadanos de 40 a 79 años, quienes superaron el 62.08 % en cada grupo decenal, lo mismo sucedió en 2015; sin embargo, por las características de la población (el porcentaje de ciudadanos menores de 40 años es mayor a 55 %, tanto en 2012 como en 2015) hay un mayor número de votos de personas menores a 40 años que de personas mayores, por lo tanto, buscamos contar con información que permita matizar esta información; usaremos un indicador de edad, en grupos quinquenales y buscaremos que se encuentren cercanos a la media poblacional de la lista nominal, o bien, que el grupo objetivo sea más joven al resto.

A partir de esta información, es posible comenzar a delinear el perfil ciudadano con mayor interés dentro de un estudio electoral; además la experiencia adquirida en la elaboración de este tipo de estudios, nos permite identificar algunas características importantes en el grupo objetivo, Por lo tanto, nuestro primer paso se centra en identificar geográficamente las zonas con un mayor número de personas que cumplan con las siguientes características:

- Mayor proporción de mujeres que de hombres.
- Jóvenes y adultos jóvenes, mayor proporción de ciudadanos entre 18 y 40 años.
- Nivel de escolaridad alto, o al menos por encima de la media nacional.

- Zonas con mayor tasa de empleo, respecto a la media nacional.
- Ciudadanos con un nivel de ingresos similar a la media nacional.
- Mayor proporción de hogares conectados a internet; lo que se traduce en mayor consumo de información (sin perdida de generalidad, información político-electoral).
- Por último, buscamos zonas con una densidad de población de media a alta.

Usando los resultados de la Encuesta Intercensal del INEGI 2015, generaremos indicadores simples (proporciones ponderadas) de sexo, edad, escolaridad, situación de empleo, nivel de ingreso y acceso a internet, con lo que obtendremos una segmentación que permita identificar a la población objetivo en estudios electorales. Usaremos a los municipios como unidades de nuestro estudio, se trata de la partición que genera mayor cantidad de conjuntos con información representativa.

6.2. Resultados de la estratificación sociodemográfica

De acuerdo a la encuesta intercensal, México estaba habitado en 2015 por 119,530,753 personas (actualmente se calcula que viven 133,086,426 aproximadamente), de los cuales 48.57 % son hombres y 51.43 % son mujeres; alrededor del 65 % por ciento de la población de 2015 tenía 15 años o más, lo que los convertiría en potenciales votantes para 2018.

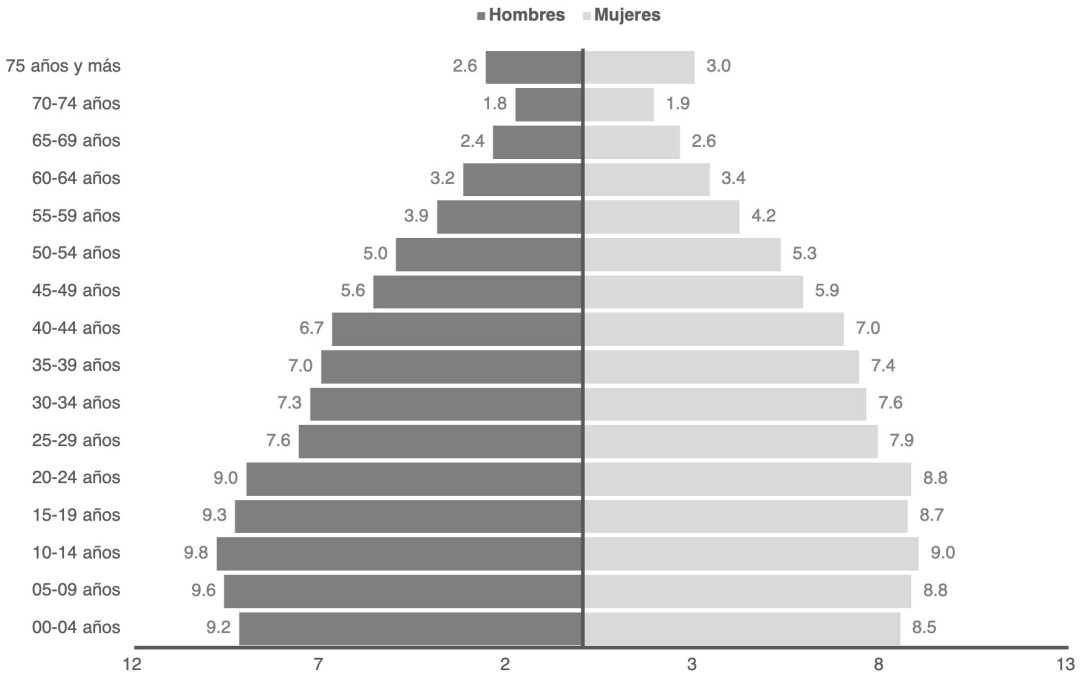

Figura 6.1: Pirámide poblacional de México, Encuesta Intercensal 2015 INEGI

La figura 6.1 señala una población predominantemente joven; nuestro interés se centra en la población de 15 años o más. De acuerdo a las estimaciones del INEGI, alrededor de 86,692,424 personas serían mayores de 18 años en 2018, obteniendo así posibilidad de participar en las elecciones federales (el INE registró un total de 89,123,355 ciudadanos inscritos en la lista nominal previo al 2 de julio, en proporción, el resultado con las estimaciones del INEGI es muy preciso).

En la Encuesta Intercensal del INEGI 2015, los ciudadanos y potenciales votantes en México tienen las siguientes características:

	Valor
Mujeres %	52.2
Hombres %	47.8
Índice de edad	4.2
Índice de escolaridad (años)	7.2
Tasa de empleo %	95.9
Nivel de ingresos	2.53
Acceso a internet %	33

Cuadro 6.1: Características sociodemográficas prospectivas de los ciudadanos en 2018

Con los indicadores de la encuesta intercensal, encontramos información relevante respecto a la ubicación de los ciudadanos con el perfil de la población objetivo; en estos nos describen al Estado de México como el estado con mayor población en edad para votar, con cerca de 12 millones de posibles votantes, la Ciudad de México, Veracruz, Jalisco y Puebla, completan la lista de los 5 estados con mayor número de potenciales votantes; estos resultados son coherentes con la información del INE, pues son estos 5 estados los que encabezan el número de ciudadanos inscritos en la lista nominal.

Oaxaca, es el estado con mayor proporción de mujeres, seguido de Puebla, la Ciudad de México, Hidalgo y Guerrero; vale la pena mencionar que existe una relación en cuanto a la ubicación de éstos, ya que los estados en cuestión, están en la zona centro y centro-sur del país. Por otro lado, los estados fronterizos de Baja California, Baja California Sur y Quintana Roo cuentan con una proporción menor de mujeres que el resto.

Hablando de la edad, existe cierta monotonía en el país, pues la mayoría de los estados presenta un nivel similar, la Ciudad de México es quien tiene una población de mayor edad en comparación al resto, Chiapas y Oaxaca son los estados con la población más joven. En cuanto a la educación se refiere, en el país hay una media de 7.2 años de estudio, es decir que la mayoría de la población mayor a 15 años tiene una educación básica, comprendida por preescolar y primaria, principalmente. En la Ciudad de México existe una media de 11.1 años de estudio, esto comprende preescolar, primaria y secundaria, con un porcentaje significativo de bachillerato. Los estados de Guerrero, Oaxaca y Chiapas son los que presentan mayor rezago educativo.

En general, la tasa de empleo en el país rebasa el 95 %; los estados de Yucatán, Campeche y Quintana Roo, presentan los mejores niveles, cabe resaltar que se trata de Estados que se dedican al turismo principalmente y se encuentran en la zona conocida como la Riviera Maya, otro centro turístico que destaca es el estado de Baja California Sur, al que pertenecen municipios como Los Cabos o La Paz, considerados como puntos de interés para el turismo estadounidense, principalmente. Tabasco, presenta la tasa de empleo con menor nivel. No obstante, no existe una correlación con el nivel de ingreso, pues Nuevo León lidera el *ranking*; Oaxaca, Chiapas y Guerrero son los estados con el peor ingreso del país, este resultado sí se relaciona con el nivel de escolaridad, pues hay una correlación positiva entre educación e ingreso.

Uno de los factores que recientemente ha tomado relevancia es el acceso a internet, pues es precisamente a través de las redes sociales, que se da inicio al debate ciudadano, se fijan posturas y se comparten noticias referentes a los candidatos y a la elección en general; la Ciudad de México y Nuevo León son los estados más conectados, al rebasar el 50 % de hogares con acceso a internet, nuevamente Guerrero, Oaxaca y Chiapas, ni siquiera alcanzan un 20 % de hogares conectados, demostrando así porque es que estos Estados se encuentran con el mayor rezago tanto, escolar, científico y tecnológico del país.

La información anterior, nos permite comenzar a identificar geográficamente los estados que deberían tener mayor presencia en la muestra; sin olvidar el número de habitantes, que nos señalaría las proporciones muestrales. Siguiendo con las características de nuestra población objetivo comparadas con los indicadores del INEGI, definiremos una simbología que nos permita entender de forma visual lo que estamos buscando:

	Valor
Mujeres %	$+$
Hombres %	$-$
Índice de edad	$\div -$
Índice de escolaridad (años)	$+$
Tasa de empleo %	$+$
Nivel de ingresos	$\div$
Acceso a internet %	$+$

Cuadro 6.2: Perfil de votantes potenciales: $+$ mayor en proporción, $-$ menor en proporción, $\div -$ en la media o menor, $\div +$ en la media o mayor, $\div$ en la media

6.2.1. Segmentación municipal

Como mencionábamos anteriormente, clasificaremos los municipios según sus indicadores sociodemográficas, buscando encontrar en alguno de los segmentos, características afines al perfil de la población objetivo .

Usaremos el método *k-means*, que como definimos anteriormente, es un método no jerárquico

de clasificación, basado en el mínimo de las distancias, en este caso euclideas, de las observaciones con los centros de los *clusters*; o sea, de las observaciones con características más diferentes entre sí. Una manera, de identificar el número óptimo de conglomerados en la clasificación es a partir del uso de una gráfica de codo.

Una gráfica de codo, nos permite identificar de forma sencilla el número *k* óptimo de *clusters*, aplicando el algoritmo de *k-means* para un rango de valores, y así identificar aquel valor donde se encuentra el mínimo posible de la suma total de las varianzas dentro de los conglomerados (después de este valor, la suma es prácticamente igual a este mínimo); o lo que es lo mismo, minimiza la suma de los cuadrados de las distancias de las observaciones al centro del *cluster* asociado. A continuación, presentamos la gráfica de codo asociada a nuestro análisis de segmentación:

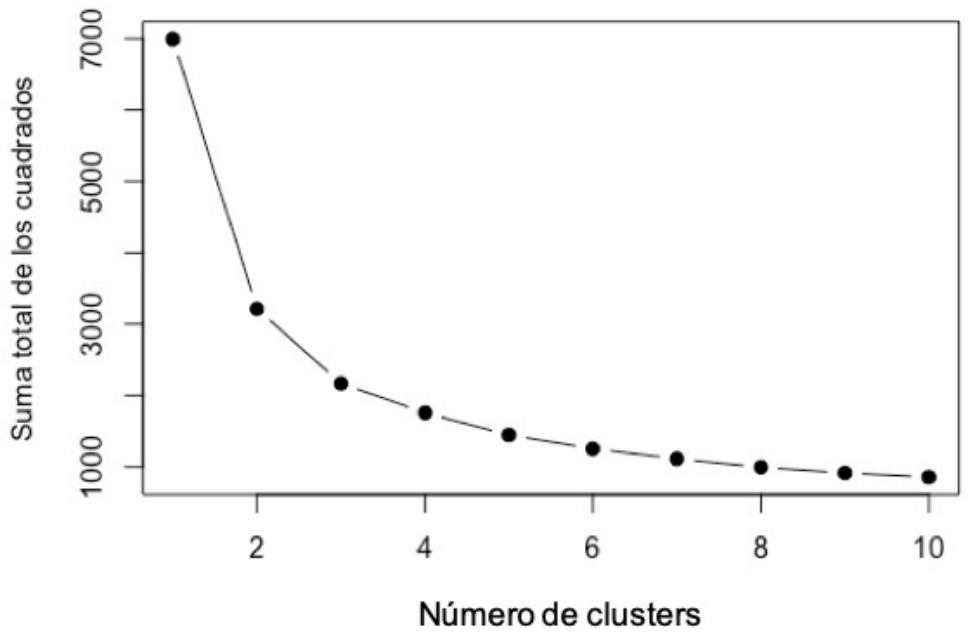

Figura 6.2: Gráfica de codo para el análisis de segmentación

Como podemos observar en la figura 6.2, el número óptimo de conglomerados sería 3; sin embargo, nuestra intención es encontrar dentro de las segmentaciones un grupo que coincida con las características del perfil definido; además, nos gustaría que la segmentación elegida se divida en partes de similar tamaño, para así evitar en este paso un sesgo no intencional en la muestra.

Con la segmentación que divide en 3 grupos a la población, la suma de la varianza entre segmentos es de 69 %; mientras que con la segmentación de tamaño 4 es de 75 %. En un primer acercamiento, la segmentación con 4 grupos, sería una buena alternativa, a partir de 5 grupos no hay mayor ganancia en este indicador. Para definir la segmentación que usaremos, compararemos los resultados del análisis para 3, 4 y 5 *clusters* y contrastaremos los centros con las características del perfil de votantes buscado.

El análisis de conglomerados por métodos no jerárquicos, sugiere variar el número de *clusters* para llegar a la mejor clasificación posible, por lo que el contraste entre las 3 opciones nos ayudará a elegir la segmentación adecuada.

- Los resultados de la clasificación en **3 conglomerados** muestran lo siguiente:

	Centros de *clusters*		
	1	2	3
Proporción de mujeres	0.522	0.517	0.528
Índice de edad	4.17	4.41	4.43
Índice de escolaridad	9.41	7.34	5.65
Tasa de empleo %	96.1	95.8	94.6
Nivel de ingreso	2.48	2.15	1.66
Acceso a Internet %	29.6	10.9	2.7

Cuadro 6.3: Resultados de la clasificasión en 3 grupos

El tamaño de cada conglomerado se muestra en el cuadrosiguiente

	Unidades por *cluster*	
Cluster	1	525
	2	1,099
	3	833

Cuadro 6.4: Tamaño de cada *cluster*

En una primera instancia, los resultados de la segmentación con 3 conglomerados no cumplen con las exigencias que buscamos; como se muestra en el cuadro6.4, el conglomerado 2 está formado por cerca de la mitad de las unidades poblacionales, por tal motivo, la muestra estaría cargada a este grupo. De acuerdo a las características del centroide del conglomerado 1, este sería el grupo afín al perfil de votantes potenciales, para fines prácticos lo llamaremos *cluster* afín.

- Para el análisis con **4 conglomerados** estos son los resultados más relevantes obtenidos:

	Centros de *clusters*			
	1	2	3	4
Proporción de mujeres	0.523	0.520	0.517	0.528
Índice de edad	4.14	4.56	4.30	4.30
Índice de escolaridad	9.76	6.63	7.94	5.29
Tasa de empleo %	96.1	95.5	95.9	94.1
Nivel de ingreso	2.52	1.95	2.29	1.58
Acceso a Internet %	33.7	6.5	15.0	1.9

Cuadro 6.5: Resultados de la clasificasión en 4 grupos

El tamaño de cada conglomerado es:

	Unidades por *cluster*	
Cluster	1	363
	2	851
	3	731
	4	512

Cuadro 6.6: Tamaño de cada *cluster*

En el caso de este análisis, las particiones son más homogéneas en cuanto al número de unidades que pertenecen a cada una, el conglomerado 1 cuenta con características afines al perfil buscado, nivel de escolaridad elevado, alto acceso a internet, nivel de ingreso de medio a medio alto y el índice de edad muestra una población más joven respecto a la media; este grupo será nuestro *cluster* afín.

El conglomerado 3, como se ve en el cuadro6.5, también cumple con características interesantes, pues a diferencia del conglomerado 1, agrupa a personas con acceso a internet medio y un nivel de ingresos menor, ambas características convertirían a este segmento en un conjunto complementario al perfil que buscamos.

- Con **5 conglomerados** los resultados fueron de la siguiente manera:

	Centros de *clusters*				
	1	2	3	4	5
Proporción de mujeres	0.525	0.517	0.528	0.526	0.518
Índice de edad	4.17	4.33	5.07	4.08	4.26
Índice de escolaridad	10.18	7.07	6.06	5.38	8.47
Tasa de empleo %	96.2	95.7	95.5	94.0	95.9
Nivel de ingreso	2.57	2.10	1.74	1.60	2.37
Acceso a Internet %	39.3	9.3	3.7	2.1	19.2

Cuadro 6.7: Resultados de la clasificación en 5 grupos

El número de unidades por conglomerado se describe a continuación:

	Unidades por *cluster*	
Cluster	1	226
	2	841
	3	377
	4	458
	5	555

Cuadro 6.8: Tamaño de cada *cluster*

El análisis con 5 conglomerados, nos deja al *cluster* 1 como el *cluster* afín; los resultados son muy similares al anterior análisis; sin embargo, el número de elementos en el grupo afín disminuyó drásticamente a 226, siendo el grupo más pequeño y en consecuencia, desequilibrando el tamaño de las particiones.

Comparando las características del grupo afín en cada análisis y contrastándolos con el perfil de la población objetivo obtenemos lo siguiente:

	Centros de *clusters*			
	Cl. Análisis 3	Cl. Análisis 4	Cl. Análisis 5	Perfil buscado
Proporción de mujeres	0.522	0.523	0.525	(+) 0.522
Índice de edad	4.17	4.14	4.17	(÷−) 4.2
Índice de escolaridad	9.41	9.76	10.18	(+) 7.2
Tasa de empleo %	96.1	96.1	96.2	(+) 95.9
Nivel de ingreso	2.48	2.52	2.57	(÷) 2.53
Acceso a Internet %	29.6	33.7	39.3	(+) 33

Cuadro 6.9: Contraste de los análisis con el perfil buscado

La mayor diferencia entre los tres posibles *clusters* afines, se encuentra en el acceso a internet, en el análisis de 3 conglomerados, el grupo afín presenta un indicador por debajo de la media, lo que aumenta la probabilidad de encuestar a personas con poco interés o bien con menor consumo de información de lo que nos interesaría tuvieran, el análisis con 4 conglomerados sí presenta un grupo afín con un nivel de acceso a internet por encima de la media nacional, lo mismo que el análisis con 5 conglomerados.

El nivel de ingreso es un indicador con rango de 1 a 3, nos interesa que el *cluster* afín, tenga un nivel cercano a la media nacional, mientras más cerca se encuentre del 3, sesgaríamos la muestra hacia la población con mayores ingresos (lo cual podría beneficiar el voto por algún candidato en particular) y lo mismo en el sentido contrario. El análisis con 5 conglomerados presenta al grupo con más ingresos, pues es el único que se encuentra por encima de la media nacional; además el índice de escolaridad es significativamente mayor al resto, esto resultaría en un mayor número de encuestados con nivel de educación alto, lo que podría cargar la muestra a ciertas zonas específicas del país, perdiendo dispersión y representatividad a nivel nacional; sumado a esto, el *cluster* afín en este análisis es el de menor tamaño por lo que el número de encuestados dentro del perfil no sería suficiente para las estimaciones que buscamos. Es un grupo que podría describirse como la parte de la población considerada como el *círculo rojo*, esto es un sector poblacional con alta preparación académica y consumo de información, tal que fungen como líderes de opinión tanto locales (entre amigos y familiares) como globales (en la población general), este no es un perfil accesible al momento de levantar la información, lo que echa por la borda parte de los objetivos de la estratificación que buscamos.

Dicho lo anterior, descartaremos el análisis con 5 conglomerados; en el caso del análisis con 3 grupos, la proporción de mujeres no es mayor a la media nacional como nos gustaría, además de tener el menor nivel de ingresos y como lo habíamos mencionado, el menor nivel de acceso a internet, esto aumentaría el número de encuestados muy probablemente en zonas rurales, lo cual no resuelve uno de los problemas iniciales del muestreo tradicionalmente usado en estudios electorales. Por estas razones y debido a ser el análisis con las particiones de tamaño más homógeneo, optaremos por el análisis con 4 conglomerados.

6.2.2. Caracterización de la segmentación seleccionada

Con el fin de facilitar la interpretación durante la metodología, en términos de una investigación social, nombraremos y caracterizaremos cada segmento del análisis de clasificación que usaremos.

- Grupo 1: **Electores potenciales**: la proporción promedio de mujeres (52.4 %) es ligeramente mayor a la media nacional (52.2 %); el índice edad (4.14) nos habla de un grupo con ciudadanos de menor edad que en el promedio nacional, su nivel de ingresos (2.52) muestra que se trata de ciudadanos de clase media a media alta, con acceso a internet en el 33.7 % de las viviendas. Se trata del grupo considerado como el de mayor probabilidad de acudir a votar, por lo tanto, la información proporcionada por este segmento es fundamental, pues de acuerdo al perfil del votante mexicano, es este grupo quien dicta, en mayor medida, las tendencias sobre la intención de voto.

- Grupo 2: **Veteranos propensos**: es el grupo de mayor edad (5.9 en el índice), el porcentaje de mujeres es ligeramente menor a la media nacional (52.0 %), el número promedio de años de estudio es 6.6, por debajo de los 7.2 años en la media del país; la tasa de empleo (95.5 %) es baja, respecto al resto de los grupos, por lo anterior, consideramos que se trata de un grupo donde destacan los jubilados y pensionados, así como trabajadores de edad avanzada; apenas el 6.5 % de las viviendas cuenta con acceso a internet. Este grupo no es de particular interés, pues su perfil no corresponde con el de los votantes tradicionales; sin embargo, se puede tratar de votantes duros, es decir, que forman parte de las bases de los partidos tradicionales; por lo tanto, explorar la información obtenida en el segmento, puede darnos indicios de pérdida o ganancia en el voto duro.

- Grupo 3: **Ciudadanos promedio**, se trata del segundo perfil en importancia dentro de las características del votante mexicano, pues presenta similitudes con la media nacional, a excepción del acceso a internet; por lo tanto, es el segmento que describe mejor a la población mayor de edad, pero no necesariamente a los electores. Este segmento es, al menos hipotéticamente, el más accesible; sin embargo, hay que tener cuidado con la información proporcionada, pues podemos esperar altos niveles de no respuesta, o bien cambios en la intención de voto que afecten la estimación final; dentro de este conjunto, esperamos encontrar informantes con baja probabilidad de acudir a votar, ya que sus características son poco afines a las del votante tradicional.

- Grupo 4: **Ciudadanos en rezago**, el nivel de ingreso (1.58), es el menor de todos, así como su acceso a internet, que alcanza apenas 1.9 % de viviendas con internet; hay en promedio un porcentaje mayor de mujeres (52.8 %) que de hombres, proporción incluso mayor al de la media nacional; el índice de edad (4.3) es muy similar a la media; presentan la escolaridad más baja con 5.3 años de estudio, lo que corresponde a nivel primaria o primaria incompleta; la tasa de empleo también es la más baja con 94.1 %. Dadas las características mencionadas, nos lleva a concluir que se trata de madres solteras principalmente, muy probablemente dependientes de programas sociales; en términos de votantes, se trata de un segmento con muy poca probabilidad de acudir a votar, no obstante, es un grupo de fácil movilización para los partidos políticos (los llamados *acarreados*, personas con tendencia a vender su voto).

Es importante señalar que con esta segmentación buscamos caracterizar a los ciudadanos, por lo tanto la distribución de la muestra será proporcional al número de ciudadanos que habitan en los municipios pertenecientes a cada conglomerado. La siguiente tabla muestra la distribución de habitantes en cada grupo y la proporción relativa.

		Número de ciudadanos	%
Cluster/ Estrato	Electores potenciales	55,189,746	63.7
	Veteranos propensos	8,691,581	10
	Ciudadanos promedio	18,869,411	21.8
	Ciudadanos en rezago	3,941,686	4.5

Cuadro 6.10: Distribución de conglomerados/ estratos

Según la información de los cómputos distritales del INE en las elecciones de 2018, participaron 56,610,985 personas; de acuerdo a la segmentación obtenida, tenemos 55,189,746 ciudadanos con alta probabilidad de acudir a votar, este número es muy cercano a la participación final; a pesar de que el objetivo no corresponde a una estimación del número de votantes, nos habla de un acercamiento interesante a los electores con mayor potencial de acudir a votar.

6.3. Diseño de la muestra

En un estudio electoral, nuestro objetivo consiste en la estimación de la intención de voto en un momento en particular, para así conocer la tendencia que nos permita estimar el resultado final de la elección.

El tamaño de la muestra está limitado por el presupuesto asignado, tradicionalmente se solicitan estudios de 1,200 encuestas aproximadamente; una de las razones principales radica en la rapidez del levantamiento, que se traduce en anticipación para la planeación de una campaña o bien para la publicación de una nota periodística. Por este motivo usaremos este tamaño de muestra dentro de nuestro diseño metodológico; el error teórico asociado a una muestra de 1,200 entrevistas (usando una varianza máxima de 0.25, al tratarse de una estimación de proporciones) es de 2.8 % al 95 % de confianza.

El margen de error usado de forma teórica, surge de la aplicación de la fórmula $e = z_{1-\frac{\alpha}{2}}\sqrt{(\frac{1}{n}-\frac{1}{N})s^2}$, con una confianza definida al 95 % se tiene un estadístico z de 1.96; al tener N muy grande o infinita, la fórmula se reduce a: $e = 1.96\sqrt{\frac{s^2}{n}}$ y tomando en cuenta que estimamos proporciones, la varianza máxima es de 0.25 con lo que se obtiene el márgen de error de 2.8 % al 95 % de confianza. El muestreo estratificado produce un margen de error más pequeño que el de un muestreo aleatorio simple; por lo que el error obtenido de este último, es usado como una cota superior del error esperado.

Para comprobar que el margen de error que se plantea en un muestreo aleatorio simple de tamaño 1,200, a un 95 % de confianza es de 2.8 %, realizaremos 1,000 simulaciones de muestras aleatorias, las gráficas de densidad de los estimadores en un muestreo aleatorio simple son:

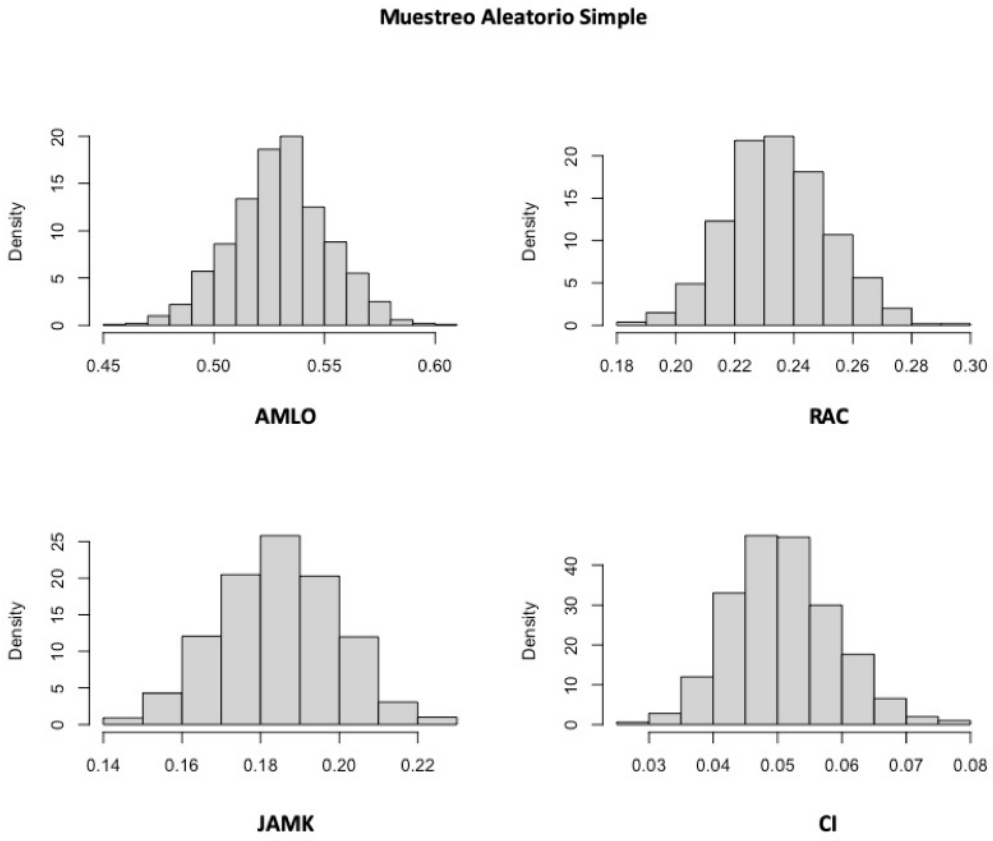

Figura 6.3: Densidades del voto para cada candidato usando m.a.s.

El error máximo corresponde a la estimación del voto por AMLO, obtenido de la distancia entre los límites del intervalo de confianza, que es de 2.82. Por lo tanto el margen de error al 95 % de confianza corresponde efectivamente al $\pm 2.8\,\%$, lo que coincide con el valor teórico usado en los estudios electorales de tamaño 1,200, su uso entonces, puede ser planteado como el máximo error teórico esperado.

En los estudios electorales al disponer de recursos muy limitados, se vuelve fundamental optimizarlos de la mejor manera, por ello, una forma de asegurar una mejor distribución geográfica es agregar una segunda estratificación en la muestra de acuerdo a las circunscripciones electorales, posterior a la segementación sociodemográfica; las circunscripciones electorales dividen a la población uniformemente en 5 zonas alrededor del país. Además de la distribución uniforme de ciudadanos, también se tiene una distribución uniforme de los votos emitidos en cada elección. La siguiente tabla muestra el número de votos por circunscripción de acuerdo a los cómputos distritales del INE en 2018.

		Número de votos
Circunscripción	1	10,340,187
	2	11,117,926
	3	11,712,348
	4	11,683,397
	5	11,757,127

Cuadro 6.11: Votos emitidos por circunscripción, cómputos distritales del INE 2018

Usaremos esta particularidad, como parte del diseño de la muestra. En resumen, el diseño metodológico corresponderá a un Muestreo Estratificado Previa Segmentación Sociodemográfica (MEPSS), la selección de la muestra será aleatoria, polietápica, estratificada y por conglomerados, con una selección sistemática del informante. De tamaño 1,200 y un margen de error *teórico* asociado de $\pm 2.8\,\%$ al $95\,\%$ de confianza. Las etapas de la selección se describen a continuación:

1. Selección aleatoria de 150 secciones electorales distribuidas proporcionalmente al número de habitantes por sección, estratificadas de acuerdo a la segmentación y estratificadas por circunscripción electoral, la distribución de la muestra quedaría de la siguiente manera:

		Circunscripción				
		1	2	3	4	5
Segmentos	Electores potenciales	23	21	12	21	19
	Ciudadanos promedio	5	7	8	5	7
	Ciudadanos en rezago	2	2	6	2	3
	Veteranos propensos	0	0	4	2	1

Cuadro 6.12: Distribución proporcional de la muestra por cada grupo original

Como podemos ver el último segmento posee poca representación en la muestra; de acuerdo al cuadro6.10 tiene una presencia de 4.5 % en la muestra; por este motivo y debido a que se trata de uno de los dos grupos junto con el de ciudadanos en rezago, con menor afinidad al perfil de votantes, los compactaremos en un solo grupo, evitando también la sobrerepresentación de alguno de los 2; a este segmento le llamaremos "En rezago y veteranos", con lo cual la distribución de las 150 secciones queda de la siguiente manera:

		Circunscripción				
		1	2	3	4	5
Grupos	Electores potenciales	23	21	12	21	19
	Ciudadanos promedio	5	7	8	5	7
	En rezago y veteranos	2	2	10	4	4

Cuadro 6.13: Distribución proporcional de la muestra por cada grupo

2. Dentro de cada sección en muestra se realizarán 8 entrevistas efectivas, se seleccionarían al menos 3 manzanas (que fungirían como los conglomerados en muestra) a partir de la representación cartográfica de la sección electoral y de las manzanas con viviendas habitadas disponibles en la sección electoral muestreada; vale la pena señalar que este paso se realizaría al momento de levantar la encuesta y sería responsabilidad del equipo de campo.

3. En cada manzana se seleccionarían sistemáticamente un mínimo de 2 viviendas y un máximo de 3, a partir del mapeo de la zona que de igual manera sería realizado al momento del levantamiento por el equipo de campo.

4. En cada vivienda se seleccionaría a 1 persona de acuerdo a un plan de cuotas por sexo y edad, diseñado con antelación al levantamiento; la persona seleccionada será la primer persona en abrir la puerta, con credencial para votar, vigente y domicilio en la vivienda, en caso de ser menor de edad o de no cumplir con la cuota, se seleccionaría a la persona (mayor de edad, con credencial vigente domiciliada en esa vivienda) con el cumpleaños más próximo.

Las unidades primarias de muestreo son las secciones electorales y las últimas unidades de muestreo serían los ciudadanos inscritos en la lista nominal.

6.4. Resultados obtenidos en estudios electorales, usando un muestreo estratificado previa segmentación sociodemográfica vs muestreos tradicionales

Para fines prácticos nos referiremos como **MEPSS** a la propuesta desarrollada en el presente trabajo, por las iniciales de Muestreo Estratificado Previa Segmentación Sociodemográfica.

Para comprobar nuestra hipótesis que de usar un MEPSS para estudios electorales se obtienen resultados más precisos en la estimación de la intención de voto que en el muestreo tradicionalmente usado, realizaremos distintas simulaciones que nos permitan demostrarlo. Para ello usaremos los cómputos distritales del INE en la elección presidencial del proceso electoral 2017-2018; omitiremos las etapas 2 y 3 del diseño metodólogico (pues se trata de una selección *in-situ*) para seleccionar directamente 1 conglomerado con 8 encuestas en cada sección, para ello realizaremos una simulación lo más cercana a la realidad; es decir, en cada sección en muestra seleccionaremos 8 ensayos de una distribución multinomial, con parámetros iguales a la proporción de voto para cada candidato en cuestión; en cada ensayo solo habrá un éxito correspondiente a la intención de voto del ciudadano encuestado.

Un último punto a considerar es el porcentaje de no respuesta; si bien en cada simulación obtendremos 1,200 encuestas, no podemos dejar de lado el hecho que existe una pérdida de información por aquellas personas que se niegan a responder a la pregunta de intención de voto específiamente, además de aquellos votos en blanco. Fijaremos pues, una tasa de no respuesta teórica; de acuerdo a los modelos de *Bloomberg* y *Oraculus*, que se encargaban de recopilar todas las encuestas públicas, con metodología probada ante el Instituto Nacional Electoral, la no respuesta oscila entre

8 % y 30 %, teóricamente disminuye conforme avanza el proceso electoral, aunque en las dos últimas elecciones esto no sucedió, por lo cual es importante usar una tasa de no respuesta con un escenario pesimista. Por lo anterior y de acuerdo a nuestra experiencia en la realización de estas investigaciones, optaremos por una tasa de no respuesta teórica de 20 %.

Como consecuencia de la tasa de no respuesta, obtendremos estimadores de razón; es decir, el número de ciudadanos que piensan votar por el candidato n, entre el número total de personas que respondieron la pregunta con alguna de las opciones de candidatos.

Nuestro punto de comparación serán los resultados finales de la elección, de acuerdo a los cómputos distritales del INE, la elección presidencial de 2018 quedó de la siguiente manera:

		Porcentaje de votos (%)
Candidatos	Ricardo Anaya Cortés (RAC)	22.28
	José Antonio Meade Kuribreña (JAMK)	16.44
	Andrés Manuel López Obrador (AMLO)	53.29
	Candidatos Independientes (CI)	5.30
	Nulos y Candidatos No Registrados	2.69

Cuadro 6.14: Resultados de la elección para Presidente de la República, cómputos distritales del INE 2018, por candidato

Usaremos únicamente los votos válidos, obteniendo como punto de referencia lo siguiente:

		Porcentaje de votos (%)
Candidatos	Ricardo Anaya Cortés (RAC)	22.9
	José Antonio Meade Kuribreña (JAMK)	16.89
	Andrés Manuel López Obrador (AMLO)	54.76
	Candidatos Independientes (CI)	5.45

Cuadro 6.15: Porcentaje de votos efectivos, con base en los cómputos distritales del INE 2018

Realizaremos 1,000 simulaciones de muestras usando el MEPSS, además de 1,000 para un muestreo tradicional donde la estratificación es por tipo de sección rural, mixta o urbana y 1,000 para un muestreo tradicional con estratos por secciones urbanas y no urbanas.

Para cada caso, obtendremos la densidad de los estimadores para la intención de voto de cada candidato, así como el intervalo de confianza al 95 % y el error de estimación (margen de error) asociado a cada tipo de muestreo.

6.4.1. Resultados obtenidos con la propuesta de MEPSS

La intención de voto estimada, a partir de la media de las intenciones de voto en los 1,000 ensayos es:

		Porcentaje de votos (%)
Candidatos	Ricardo Anaya Cortés (RAC)	23.08
	José Antonio Meade Kuribreña (JAMK)	17.62
	Andrés Manuel López Obrador (AMLO)	53.99
	Candidatos Independientes (CI)	5.31

Cuadro 6.16: Media de las simulaciones de muestras con MEPSS

Las gráficas de las densidades asociadas a las simulaciones con MEPSS son:

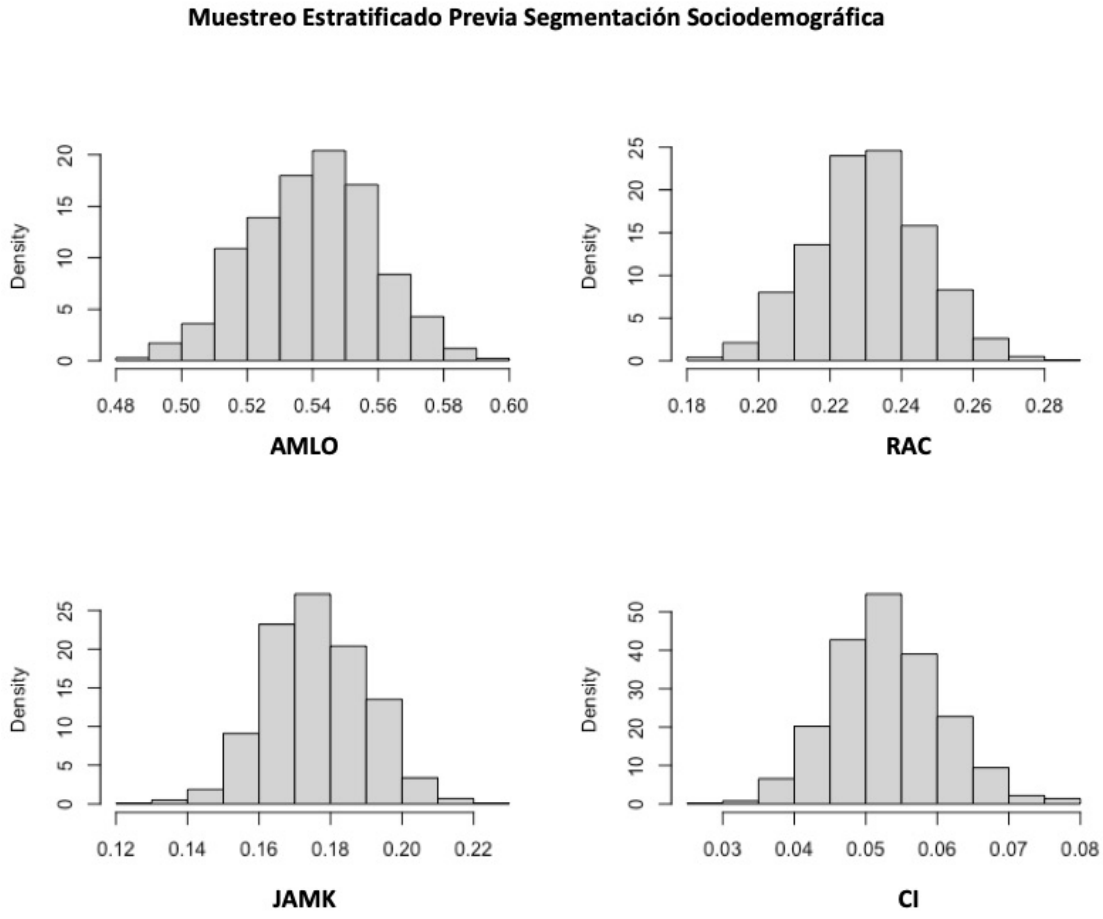

Figura 6.4: Densidades del voto para cada candidato usando MEPSS

Los intervalos de confianza al 95 % correspondientes:

Intervalo de Confianza		
Candidatos	RAC	$(22.0, 26.2)$
	JAMK	$(16.6, 20.3)$
	AMLO	$(52.6, 57.7)$
	CI	$(4.8, 6.9)$

Cuadro 6.17: Intervalos de confianza usando MEPSS

El margen de error asociado en este tipo de muestreo es de magnitud 2.53; sin embargo el error de estimación absoluto (máxima diferencia entre los valores reales y los valores estimados) fue de 0.77. Por lo tanto consideraremos un margen de error teórico de ± 3.3 % al 95 % de confianza.

Podemos apreciar que las estimaciones obtenidas a partir de nuestra propuesta son muy cercanas al resultado final de las elecciones presidenciales de 2018, esto pasa debido a la estratificación sociodemográfica que potencializa la probabilidad de encontrar informantes afines al perfil de votantes. Pensando en la aplicación real del MEPSS y viendo los intervalos obtenidos, nos permite concluir que el resultado de las elecciones estaba prácticamente decidido, el primer lugar llevaba una clara ventaja como lo mostraban la mayoría de las encuestas; sin embargo, en cuanto al 2o y 3er lugar siempre se hablaba de una competencia cerrada, pero con el MEPSS comprobamos que no había posibilidad alguna de empate y las posiciones estaban claras. Los siguientes ejercicios evidencian precisamente esta problematica.

6.4.2. Resultados obtenidos con un muestreo estratificado por tipo de sección: rural, mixta y urbana

La intención de voto estimada correspondiente a los 1,000 ensayos es:

		Porcentaje de votos (%)
Candidatos	Ricardo Anaya Cortés (RAC)	23.24
	José Antonio Meade Kuribreña (JAMK)	18.53
	Andrés Manuel López Obrador (AMLO)	53.21
	Candidatos Independientes (CI)	5.02

Cuadro 6.18: Media de las simulaciones de las muestras estratificadas con 3 tipos de secciones

Las densidades asociadas al muestreo estratificado por tipo de secciones electorales mixta, rural y urbana, se muestran en la siguiente gráfica:

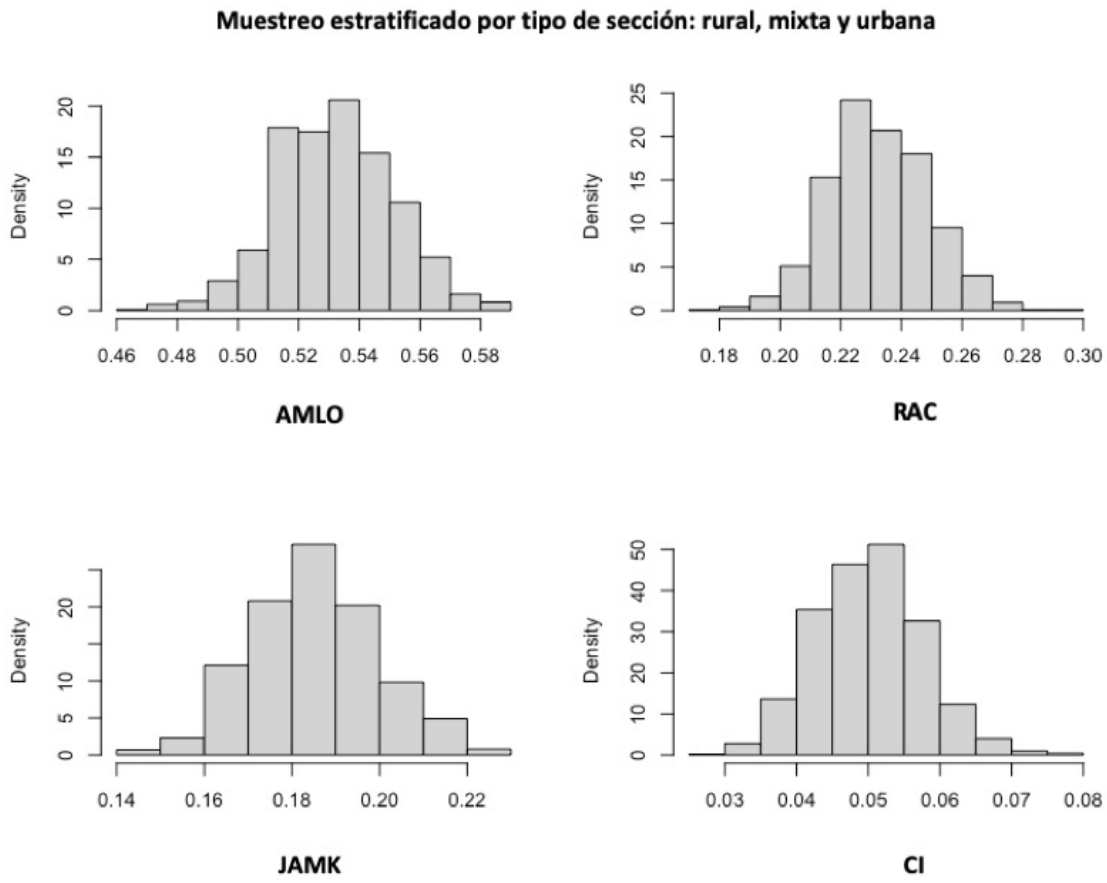

Figura 6.5: Densidades del voto para cada candidato usando muestreo estratificado por 3 tipos de secciones electorales

Los intervalos al 95 % de confianza son:

Intervalo de Confianza		
Candidatos	RAC	$(22.1, 26.5)$
	JAMK	$(17.6, 22.0)$
	AMLO	$(51.8, 56.8)$
	CI	$(4.5, 6.6)$

Cuadro 6.19: Intervalos de confianza usando muestreo estratificado por 3 tipos de secciones electorales

En este caso el máximo margen de error teórico es del orden de 2.56; con una diferencia de estimación entre los valores reales y los valores calculados de 1.64, por lo que el margen de error asociado sería de ± 4.2 % al 95 % de confianza.

6.4.3. Resultados obtenidos con un muestreo estratificado por tipo de sección: urbana, no urbana

La intención de voto estimada por la media en los 1,000 ensayos fue:

		Porcentaje de votos (%)
Candidatos	Ricardo Anaya Cortés (RAC)	23.33
	José Antonio Meade Kuribreña (JAMK)	18.44
	Andrés Manuel López Obrador (AMLO)	53.22
	Candidatos Independientes (CI)	5.01

Cuadro 6.20: Media de las simulaciones de las muestras estratificadas con 2 tipos de secciones

Los histogramas de las densidades en este tipo de muestreo:

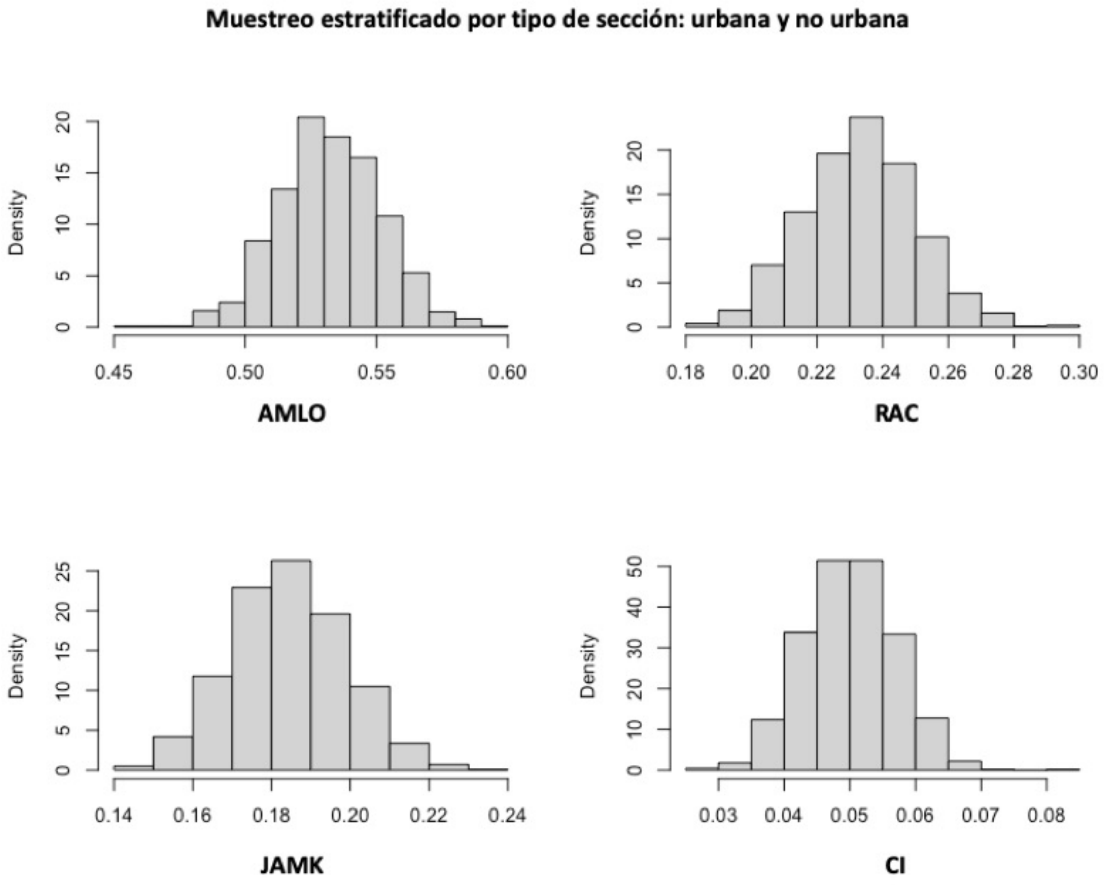

Figura 6.6: Densidades del voto para cada candidato usando muestreo estratificado por 2 tipos de secciones electorales

Los intervalos al 95 % de confianza fueron:

		Intervalo de Confianza
Candidatos	RAC	$(22.2, 26.6)$
	JAMK	$(17.4, 21.3)$
	AMLO	$(51.2, 57)$
	CI	$(4.5, 6.4)$

Cuadro 6.21: Intervalos de confianza usando muestreo estratificado por 2 tipos de secciones electorales

Para el caso del muestreo estratificado en 2 tipos de secciones, el máximo error muestral fue de magnitud 2.51; con una diferencia entre los valores reales y los valores estimados de 1.55, por lo que el margen de error asociado es de ± 4.6 % al 95 % de confianza.

Capítulo 7

Conclusiones

En los procesos electorales actuales, las encuestas se han convertido en un protagonista más de la contienda; debido a que a partir de ellas se define la estrategia a seguir para los candidatos, también son publicadas en los distintos medios de comunicación, con el fin de dar a conocer el escenario más probable para el día de la elección y así saber que candidato o candidatos tienen posibilidades de salir victoriosos en la elección.

La demanda de información en esta época ha aumentado considerablemente, en consecuencia surgen cientos de opciones que ofrecen precisión en la estimación a costos cada vez más bajos; sin embargo, en la mayoría de los casos no hay rigor estadístico en las investigaciones. Algunos ofrecen encuestas a través de internet, teléfonos celulares, aplicaciones móviles e incluso redes sociales; basan sus afirmaciones y análisis a partir de conceptos novedosos o en tendencia como el *big data*, la minería y la ciencia de datos, Estas técnicas que parten del análisis de grandes volúmenes de información, han influenciado a los estudios electorales, asegurando que al tener una muestra de decenas de miles de entrevistados obtienen estimaciones con menor margen de error y en consecuencia mayor precisión; sin embargo, la disponibilidad de recursos tecnológicos y los filtros de confirmación de datos personales (como sexo y edad) en México, imposibilitan la viabilidad de estudios electorales confiables.

Las encuestas cara a cara continúan dando mayor certeza de la calidad de la información obtenida previa al análisis estadístico. Para su realización, se usan metodologías estándar que dieron buenos resultados en el pasado; no obstante, las estimaciones obtenidas han fallado gravemente en los años más recientes.

La metodología de selección de muestra, usa tradicionalmente una estratificación por tipo de sección, ya sea rural, mixta y urbana, o bien urbana y no urbana; esta separación obedece a la idea de que la forma de votar varía entre las comunidades rurales y urbanas. La selección de la muestra con estas hipótesis sigue una distribución proporcional al número de secciones electorales en cada estrato, pero no toma en cuenta que la división del territorio nacional en secciones electorales definidas por el INE depende de la densidad de población, la extensión territorial y las cracterísticas geográficas; es decir, en una zona altamente poblada habrá más secciones electorales y lo mismo

en zonas con áreas geográficas muy grandes, o lo que es lo mismo, un estado muy poblado tendrá más secciones y también un estado muy grande. La distribución en circunscripciones y distritos federales sí divide a la población en particiones homógeneas, en cuanto al número de habitantes; no así las secciones electorales.

Por lo anterior, al usar el muestreo tradicional habrá un sesgo hacia el número de secciones rurales y mixtas o no urbanas; en los resultados con el muestreo estratificado con 3 tipos de secciones, vemos que el intervalo de confianza del voto por JAMK (17.6, 22.0) prácticamente se junta en el límite inferior del intervalo del voto por RAC (22.1,26.5), por lo tanto existe la posibilidad de estimar un resultado muy cerrado entre estos dos candidatos e incluso plantear escenarios de empate; más grave aún es que el intervalo de JAMK está por encima del resultado real, lo que quiere decir que en este tipo de muestreos se tenderá a sobreestimar el voto por el candidato del PRI; históricamente los habitantes de comunidades rurales tienden a votar por el PRI en mayor proporción, por lo que tener más secciones rurales en muestra de las necesarias, favorece el resultado para los candidatos de este partido. Esto mismo ocurre con un muestreo estratificado por dos tipos de secciones electorales, el intervalo al 95 % de confianza para el voto por JAMK es de 17.4 a 21.3, lo que deja al resultado real fuera del intervalo y en consecuencia sobreestimando la proporción de votos por el candidato; este intervalo no es tan cercano al intervalo de RAC, por lo que podríamos asumir que se disminuye la probabilidad de estimar un resultado muy cerrado entre 2o y 3er lugar.

Uno de los supuestos estadísticos en muestreo es que a mayor número de estratos mejor será el resultado de las estimaciones; sin embargo, en el caso de los estudios electorales parece que esto no necesariamente se cumple, si observamos los estimados usando un muestreo estratificado con 3 tipos de secciones electorales, podemos apreciar que son ligeramente menos precisos que los estimados con el muestreo estratificado por 2 tipos de secciones, esta pequeña diferencia se debe a que asignamos un número de encuestas fijo en cada estrato rural y mixto, lo que aumenta ligeramente la representatividad de los votantes principalmente de zonas rurales; si consideramos que en ambos casos estamos usando las circunscripciones como un segundo estratificado, tenemos 15 y 10 estratos respectivamente. En el muestreo con 15 estratos el margen de error total es de $\pm 4.2\,\%$ mientras que en el de 10 es de $\pm 4.06\,\%$, en términos generales no existiría una diferencia significativa en la precisión de los estimados obtenidos.

Hablando del vencedor de las elecciones, en ambos casos obtuvimos estimaciones que subestiman la proporción de votos obtenida; a pesar de obtener intervalos que incluyen el resultado final, al tener una media subestimada, los estimados obtenidos con muestreos estratificados por tipo de secciones electorales, tienen un alto riesgo de estar por debajo del resultado real.

Cuando usamos el MEPSS obtenemos resultados más precisos que cualquier estratificación territorial basada en el tipo de secciones electorales, las medias de las proporciones estimadas de voto distan en menos de 1 % al resultado real para todos los candidatos. Otro factor a considerar es que acota mejor el intervalo de JAMK (16.6,20.3), alejándolo más que en los casos anteriores, del intervalo de RAC (22.0,26.2), por lo que se obtiene claridad en el 2o lugar de la contienda.

Si hablamos del voto por AMLO, con el MEPSS se mejora la estimación, la siguiente figura muestra las densidades obtenidas con los muestreos estratificados por tipo de secciones electorales y el MEPSS:

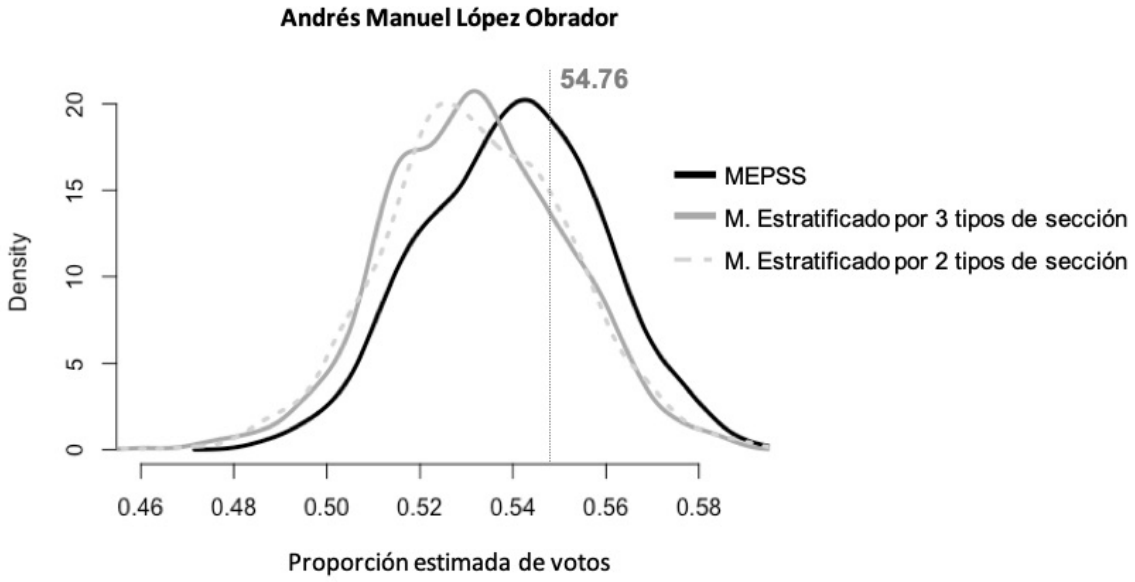

Figura 7.1: Densidades comparativas del voto para AMLO

Como podemos ver en la figura anterior, el voto por AMLO estimado con muestras estratificadas por tipo de sección, están sesgadas a la izquierda, por lo que la probabilidad de subestimar la proporción de votos por el candidato ganador es mayor que al usar MEPSS.

Hemos mencionado que parte fundamental de las fallas al estimar usando los muestreos tradicionales, radica en las secciones rurales y que es precisamente ahí donde el candidato priísta suele ser más respaldado. Si vemos la siguiente gráfica que presenta las densidades del voto por JAMK con los distintos muestreos, podemos ver que efectivamente al usar estratos por tipo de secciones electorales, aumenta la probabilidad de sobreestimar el voto por este candidato.

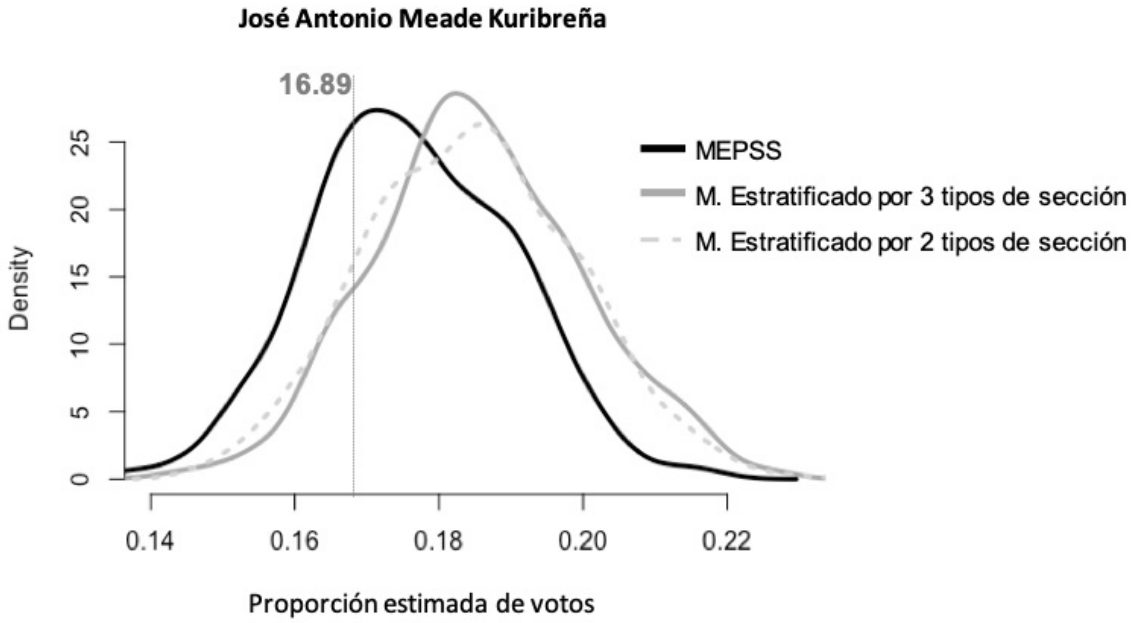

Figura 7.2: Densidades comparativas del voto para JAMK

En cuanto al margen de error asociado al MEPSS es de $\pm 3.3\,\%$ al $95\,\%$ de confianza, mejorando la precisión de las estimaciones para el tipo de estudios planteados. La ventaja radica en mejorar la calidad de los estratos, porque los 15 (3 segmentos por 5 circunscripciones) grupos que usamos matizan mejor las características de las unidades últimas de muestreo; nos basamos en la hipótesis de que el voto está influenciado por una serie de características sociodemográficas, en lugar de suponer una influencia por el tipo de sección (rural, mixto y urbano); la premisa con el MEPSS es que el resultado de las elecciones depende de quienes acudirían a votar, por el contrario en los muestreos tradicionales se parte de suposiciones sobre la forma de votar, que finalmente es el objetivo de la investigación. Es así como optimizamos la calidad del informante, pues con el MEPSS aumentamos la probabilidad de encuestar a un votante potencial; mientras que en el muestreo tradicional, cualquier ciudadano inscrito en la lista nominal tiene la misma probabilidad de ser encuestado.

A pesar de que los errores absolutos en los 3 tipos de muestreo usados, están por debajo del error en una muestra aleatoria simple, les sumamos el máximo error de estimación para obtener un margen de error asociado a 2 tipos de error: de muestreo y de estimación, obteniendo un margen más realista de lo que se esperaría en cada tipo de muestreo.

Desde el punto de vista estadístico, al usar la estratificación natural (rural, mixta y urbana o urbana y no ubana), aumentamos la varianza de los estimadores, pues se considera a cada estrato como un conjunto independiente, por lo que en las estimaciones obtenemos rangos mayores; en

consecuencia aumentamos el riesgo de sobreestimar el voto de él/los candidatos con mejor posición en alguno de los estratos frente al otro (estrato). Todo esto se deriva de hipótesis subjetivas sobre la opinión pública por parte de los investigadores; la seguridad para realizarlas, muchas veces es provocada por la experiencia misma. Con el MEPSS, partimos de una sola hipótesis: el resultado de la elección depende de la decisión de los votantes y no de las preferencias de la ciudadanía; dicho lo anterior, se usa la información disponible para caracterizar el perfil del votante, obteniendo una muestra proporcional al número de electores potenciales, evitando así suposiciones en la forma de votar y por ende un aumento en la varianza de los estimadores.

Si comparamos la distribución de la muestra entre las estratificaciones tradicionales y el MEPSS, tenemos lo siguiente:

		Tipo de sección		
		Rural	Mixto	Urbano
Grupos	Electores potenciales	4	6	85
	Ciudadanos promedio	4	14	14
	En rezago y veteranos	4	15	4

Cuadro 7.1: Distribución de la muestra por tipo de sección y segmentos

Con la tabla anterior podemos ver que en el MEPSS hay 103 secciones urbanas en muestra, por encima de las 95 que se requieren en un muestreo con la estratificación natural, esto nos habla de como en el MEPSS eliminamos parte del sesgo producido por la sobrerepresentación de la muestra en el estrato rural. En el MEPSS requerimos de 35 estratos rurales por debajo de los 41 que se piden en el muestreo usado típicamente; todo esto es porque existe un mayor número de votantes potenciales en zonas urbanas que en el resto. Finalmente esta característica, facilitaría que el levantamiento se realizara más rápido y por ende agiliza la presentación de resultados.

Una de las desventajas de la propuesta, es la debilidad para sustituir elementos de la muestra; por ejemplo, si se decide que algunas unidades en la muestra no son viables para visitar y resulta que son unidades pertenecientes al grupo de "Electores Potenciales", habría que volver a realizar la selección, pues se compromete el supuesto de aleatoriedad, ya que el segmento domina la muestra:

		Proporción de la muestra (%)
Segmentos	Electores Potenciales	64
	Ciudadanos Promedio	21.33
	En Rezago y Veteranos	14.67

Cuadro 7.2: Proporción de la muestra por segmento

Como se puede ver en la tabla 7.2, habría que sustituir un máximo de 64 % de la muestra; en el caso de usar el muestreo estratificado tradicional, se sustituiría un máximo de 30 unidades muestrales (considerando una partición 100 % Urbana), lo que significa un 20 % de la muestra.

Otra desventaja es la poca flexibilidad de la técnica, ya que la segmentación que se obtiene es única; en ejercicios posteriores, habría que volver a realizarla a partir de la actualización de los datos sociodemógraficos oficiales; si se trata de elecciones intermedias o bien locales, habría que realizar una segmentación específica. Mientras que en el muestreo típico la estratificación es universal; incluso, se puede usar en cualquier estudio de opinión pública, incluyendo evaluación de personajes e instituciones, de programas sociales o gestión gubernamental, por mencionar algunos ejemplos; agregando que se podría usar en estudios de mercado.

Para terminar, con este trabajo nos gustaría transmitir la importancia de proponer mejoras metodológicas en los estudios no solo de opinión pública sino de *marketing*; en la actualidad, las ventajas tecnológicas facilitan el trabajo estadístico, por lo que el desarrollo del sector depende en mayor medida de la incorporación de técnicas novedosas (algunas que se usaban tradicionalmente en otras áreas) que de la agilidad en la obtención de resultados; deberíamos enfocarnos en mejores estimaciones y no en la prontitud de la publicación de resultados. Si la credibilidad en las encuestas se ha puesto en entredicho por malas prácticas de algunos investigadores, deberíamos recuperarla a partir de métodos disruptivos que incorporen cada vez más la información disponible, además de empezar a incluir aquella información obtenida a partir de redes sociales, internet o móviles. Sin pasar por alto la necesidad de instrumentos de medición (cuestionarios) más dinámicos, que permitan enriquecer la información y estimación de resultados, sin comprometer la privacidad y la confianza de las personas.

Apéndice A

Código usado en R

```
#### Datos de los computos distritales en la eleccion para
presidente de la Republica##########################
setwd("/Users/taniasarmiento/Documents/Reporte␣Titulacion/base_18/")
datos <- read.csv("base_candidatos_2018.csv",header = TRUE,
                  stringsAsFactors = FALSE)
datos<- subset(datos ,TOTAL_VOTOS_CALCULADOS-nulo > 0,
                  header=TRUE, stringsAsFactors = FALSE)

################## Numero optimo de clusters #########
library(dplyr)

########### Base de datos del INEGI, sociodemograficos##
dat <- read.csv("base_sociod.csv",header = TRUE)
dat_clust <- read.csv("base_sociod_CL.csv",header = TRUE)
attach(dat_clust)

borrar <- c("Pob")
dat_clust<- dat_clust[ , !(names(dat_clust) %in% borrar)]
dat_clust<-dat_clust[ dat_clust$X...folio!="#N/A", ]

rownames(dat_clust)=dat_clust$X...folio
borrar2 <- c("X...folio")
dat_clust<- dat_clust[ , !(names(dat_clust) %in% borrar2)]
str(dat_clust)
summary(dat_clust)

############## Grafica de codo ########################
library(cluster)
```

```
wss <- sapply(1:10, function(k){kmeans(na.omit(dat_clust),
                k,iter.max = 15, nstart=50 )$tot.withinss})
plot(1:10, wss, type="b", pch = 19,  xlab="Numero de clusters",
                ylab="Suma total de los cuadrados")

clust3<- kmeans(na.omit(dat_clust),3,nstart = 50)
clust4<- kmeans(na.omit(dat_clust),4, nstart = 50)


############################################################
################## Muestreo ################################
############################################################

mat_muestra <- matrix(c(23,5,2,21,7,2,12,8,10,21,5,4,19,7,4), nrow = 3)
colnames(mat_muestra)=c("c1","c2","c3","c4","c5")
rownames(mat_muestra)=c("Electores Potenciales","Ciudadanos Propensos",
                                        "En rezago y veteranos")

K=1000
cand=4
encu=8

amlo=rep(0,K)
jamk=rep(0,K)
rac=rep(0,K)
ci=rep(0,K)

############################## MEPSS ########################
for (k in 1:K){
  t=0
  MUestJ=data.frame()
  MUESTRA=data.frame()
for (j in 1:nrow(mat_muestra)){
  estJ<- datos[datos$clust_muest==j,]

  for (i in 1:ncol(mat_muestra)){
    estJ_i<- arrange(na.omit(estJ[estJ$circ== i,]),LISTA_NOMINAL_CASILLA)
    t=length(estJ_i$seccion)
    estJ_MUESTi<-sample(t,mat_muestra[j,i])

        estJI_M<-estJ_i[estJ_MUESTi,]
        MUestJ<-na.omit(rbind(MUestJ,estJI_M))
    }
```

```
MUESTRA<-MUestJ
}

secc=length(MUESTRA$seccion)
matSECC_rsim <- matrix(rep(0,secc*encu), nrow=encu, ncol=secc)
muest<-MUESTRA

  for (l in 1:secc){
    simuest_i<-matrix(rep(0,encu*cand),nrow = cand,encu)
    RAC_nri=0.8*((muest[l,]$RAC_pan_prd_mc)/(muest[l,]$
    TOTAL_VOTOS_CALCULADOS-muest[l,]$nulo))
    JAMK_nri=0.8*((muest[l,]$JAMK_pri_pvem_panal)/(muest[l,]$
    TOTAL_VOTOS_CALCULADOS-muest[l,]$nulo))
    AMLO_nri=0.8*((muest[l,]$AMLO_mor_pt_pes)/(muest[l,]$
    TOTAL_VOTOS_CALCULADOS-muest[l,]$nulo))
    CI_nri=0.8*((muest[l,]$CI)/(muest[l,]$
    TOTAL_VOTOS_CALCULADOS-muest[l,]$nulo))
    NR_nri=0.2

    simuest_i<-rmultinom(encu,1,prob = c(RAC_nri,JAMK_nri,AMLO_nri,
                         CI_nri,NR_nri))

    a<-rep(1,sum(simuest_i[1,]))
    b<-rep(2,sum(simuest_i[2,]))
    c<-rep(3,sum(simuest_i[3,]))
    d<-rep(4,sum(simuest_i[4,]))
    e<-rep(9,sum(simuest_i[5,]))

    voto_i=c(a,b,c,d,e)
    matSECC_rsim[,l]=voto_i
  }

votoSECC=as.vector(matSECC_rsim)
votoSECC<-factor(votoSECC, levels=c(1,2,3,4,9),
                 labels=c("RAC","JAMK","AMLO","CI","NR"))

votoSECC_efec=votoSECC[votoSECC!="NR"]

rac_v<-length(votoSECC_efec[votoSECC_efec=="RAC"])
                 /length(votoSECC_efec)
jamk_v<-length(votoSECC_efec[votoSECC_efec=="JAMK"])
                 /length(votoSECC_efec)
amlo_v<-length(votoSECC_efec[votoSECC_efec=="AMLO"])
```

```
                    /length(votoSECC_efec)
ci_v<-length(votoSECC_efec[votoSECC_efec=="CI"])
                    /length(votoSECC_efec)

amlo[k]=amlo_v
jamk[k]=jamk_v
rac[k]=rac_v
ci[k]=ci_v
}

par(mfrow=c(2,2))

hist(amlo,freq=FALSE,col="lightgray",main="AMLO",sub="")
hist(rac,freq=FALSE,col="lightgray",main="RAC",sub="")
hist(jamk,freq=FALSE,col="lightgray",main="JAMK",sub="")
hist(ci,freq=FALSE,col="lightgray",main="CI",sub="")

ER_amlo= (quantile(amlo,probs = 0.975)-quantile(amlo,probs = 0.25))/2
ER_rac= (quantile(rac,probs = 0.975)-quantile(rac,probs = 0.25))/2
ER_jamk= (quantile(jamk,probs = 0.975)-quantile(jamk,probs = 0.25))/2
ER_ci= (quantile(ci,probs = 0.975)-quantile(ci,probs = 0.25))/2

mat=matrix(c(mean(amlo),mean(rac),mean(jamk),mean(ci),ER_amlo,ER_rac,
                    ER_jamk,ER_ci),nrow=4)
dimnames(mat)<-list(c("AMLO", "RAC", "JAMK", "CI"), c("Intencion de voto",
                    "Error de estimacion"))

################### Estratificado 3 estrat ########################
mat_muestra2 <- matrix(c(2,7,21,2,8,20,4,11,15,2,6,22,2,9,19), nrow = 3)
colnames(mat_muestra2)=c("c1","c2","c3","c4","c5")
rownames(mat_muestra2)=c("MIXTO","RURAL", "URBANO")

amlo2=rep(0,K)
jamk2=rep(0,K)
rac2=rep(0,K)
ci2=rep(0,K)

for (k in 1:K){
  t2=0
  MUestJ2=data.frame()
  MUESTRA2=data.frame()
  for (j in 1:nrow(mat_muestra2)){
    estJ2<- datos[datos$tipo_secc3_cod==j,]
```

```
    for (i in 1:ncol(mat_muestra2)){
      estJ_i2<- arrange(na.omit(estJ2[estJ2$circ== i,]),
                        LISTA_NOMINAL_CASILLA)
      t2=length(estJ_i2$seccion)
      estJ_MUESTi2<-sample(t2,mat_muestra2[j,i])

      estJI_M2<-estJ_i2[estJ_MUESTi2,]
      MUestJ2<-na.omit(rbind(MUestJ2,estJI_M2))
    }
    MUESTRA2<-MUestJ2
  }

  secc=length(MUESTRA2$seccion)
  matSECC_rsim2 <- matrix(rep(0,secc*encu), nrow=encu, ncol=secc)
  muest2<-MUESTRA2

  for (l in 1:secc){      [...]
    }

################ Estratificado 2 estrat ################################
mat_muestra3 <- matrix(c(9,21,10,20,15,15,8,22,11,19), nrow = 2)
colnames(mat_muestra3)=c("c1","c2","c3","c4","c5")
rownames(mat_muestra3)=c("NO_URBANO", "URBANO")

amlo3=rep(0,K)
jamk3=rep(0,K)
rac3=rep(0,K)
ci3=rep(0,K)

for (k in 1:K){
  t3_r=0
  t3_u=0
  MUestJ3=data.frame()
  MUestJ3_NU=data.frame()
  MUestJ3_U=data.frame()
  MUESTRA3=data.frame()

    estJ3_NU<- datos[datos$tipo_secc3_cod!=3,]
    for (i in 1:ncol(mat_muestra3)){
      estJ_i3_r<- arrange(na.omit(estJ3_NU[estJ3_NU$circ== i,]),
      LISTA_NOMINAL_CASILLA)
      t3_r=length(estJ_i3_r$seccion)
```

```
      estJ_MUESTi3_r<-sample(t3_r,mat_muestra3[1,i])
      estJ_M3_nu<-estJ_i3_r[estJ_MUESTi3_r,]

      estJ_M3_nu<-estJ_i3_r[estJ_MUESTi3_r,]
      MUestJ3_NU<-na.omit(rbind(MUestJ3_NU,estJ_M3_nu))
    }

      estJ3_U<- datos[datos$tipo_secc3_cod==3,]
      for (i in 1:ncol(mat_muestra3)){
        estJ_i3_u<- arrange(na.omit(estJ3_U[estJ3_U$circ== i,]),
        LISTA_NOMINAL_CASILLA)
        t3_u=length(estJ_i3_u$seccion)
        estJ_MUESTi3_u<-sample(t3_u,mat_muestra3[2,i])
        estJ_M3_u<-estJ_i3_u[estJ_MUESTi3_u,]

        estJ_M3_u<-estJ_i3_u[estJ_MUESTi3_u,]
        MUestJ3_U<-na.omit(rbind(MUestJ3_U,estJ_M3_u))
      }

       MUestJ3<-na.omit(rbind(MUestJ3_NU,MUestJ3_U))

        MUESTRA3<-MUestJ3


  secc=length(MUESTRA3$seccion)
  matSECC_rsim3 <- matrix(rep(0,secc*encu), nrow=encu, ncol=secc)
  muest3<-MUESTRA3

   for (l in 1:secc){      [...]
    }

###################### mas ####################################

amlo4=rep(0,K)
jamk4=rep(0,K)
rac4=rep(0,K)
ci4=rep(0,K)

for (k in 1:K){
  N_k<-length(na.omit(datos)$seccion)
  n_k<-150
  mas<-sample(seq(N_k),n_k)
  muestra_mas<-datos[mas,]
```

```
  MUESTRA4=na.omit(muestra_mas)

  secc=length(MUESTRA4$seccion)
  matSECC_rsim4 <- matrix(rep(0,secc*encu), nrow=encu, ncol=secc)
  muest4<-MUESTRA4

  for (l in 1:secc){      [...]
    }

############## Resultados ###################################
mat
mat2
mat3
mat4

par(mfrow=c(1,1))
hist(amlo2,freq=FALSE,col="white",main="AMLO",sub="",lty=0)


lines(density(amlo),type="l",col="black",lwd=3,main="AMLO",
                xlab="",ylab="",las=1)
lines(density(amlo2),type="l",col="darkgray",lwd=3,main="AMLO",
                xlab="",ylab="",las=1)
lines(density(amlo3),type="l",col="lightgray",lwd=3,lty=3, main="AMLO",
                xlab="",ylab="",las=1)

hist(jamk2,freq=FALSE,col="white",main="JAMK",sub="",lty=0)
lines(density(jamk),type="l",col="black",lwd=3,xlab="",ylab="",las=1)
lines(density(jamk2),type="l",col="darkgray",lwd=3,xlab="",ylab="",las=1)
lines(density(jamk3),type="l",col="lightgray",lwd=3, lty=3,xlab="",
                ylab="",las=1)
legend("bottomleft",col=c("blue","green"),legend =c("Coseno","Seno"),
                lwd=3, bty = "n")
lines(x,sin(x),col="green",lwd=3)
legend("bottomleft",col=c("blue","green"),legend =c("Coseno","Seno"), lwd=3, bty =

quantile(amlo,probs = c(0.25,0.975))
quantile(rac,probs = c(0.25,0.975))
quantile(jamk,probs = c(0.25,0.975))
quantile(ci,probs = c(0.25,0.975))
[...]
```

Apéndice B

Código usado en SPSS

```
########## Segmentacion ############

QUICK CLUSTER Mujeres Hombres Edad_ind
Escola_ind Tasa_empleo Tasa_desempleo
Nivel_Ingreso Internet
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(3) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /PRINT INITIAL.
########## cluster, varia hasta 6 ###########

########## Distribucion de la muestra #######
########## segun segmentos ##############

DO IF    Circunscripcion =       1       .
         RECODE  clustiET
                 (       1       =       23      )
                 (       2       =       5       )
                 (       3       =       2       )
         INTO            cuotas_muestra  .
END IF.
EXECUTE.

DO IF    Circunscripcion =       2       .
         RECODE  clustiET
                 (       1       =       21      )
                 (       2       =       7       )
                 (       3       =       2       )
```

```
        INTO            cuotas_muestra  .
END IF.
EXECUTE.

DO IF   Circunscripcion =       3       .
        RECODE  clustiET
                (       1       =       12      )
                (       2       =       8       )
                (       3       =       10      )
        INTO            cuotas_muestra  .
END IF.
EXECUTE.

DO IF   Circunscripcion =       4       .
        RECODE  clustiET
                (       1       =       21      )
                (       2       =       5       )
                (       3       =       4       )
        INTO            cuotas_muestra  .
END IF.
EXECUTE.

DO IF   Circunscripcion =       5       .
        RECODE  clustiET
                (       1       =       19      )
                (       2       =       7       )
                (       3       =       4       )
        INTO            cuotas_muestra  .
END IF.
EXECUTE.


RECODE DOMICILIO (MISSING=0) INTO cuotas_muestra.
EXECUTE.

RECODE Cluster_pertenencia (1=1) (2=3) (3=2) (4=3) INTO clust.
EXECUTE.

COMPUTE filter_$=(cuotas_muestra > 0).
VARIABLE LABELS filter_$ 'cuotas_muestra > 0 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
```

```
EXECUTE.

############### Muestra ############################

CSPLAN SAMPLE
   /PLAN FILE='C:xxxxx/muestreo_RT.csplan'
   /PLANVARS SAMPLEWEIGHT=SampleWeight_Final_
   /PRINT PLAN
   /DESIGN STRATA=Circunscripcion clustiET CLUSTER=UNICO
   /METHOD TYPE=SIMPLE_SYSTEMATIC ESTIMATION=DEFAULT
   /SIZE VARIABLE=cuotas_muestra
   /STAGEVARS INCLPROB(InclusionProbability_1_)
  CUMWEIGHT(SampleWeightCumulative_1_).
DATASET DECLARE  muestra1.
CSSELECT
   /PLAN FILE='C:xxxxxx/muestreo_RT.csplan'
   /CRITERIA STAGES=1 SEED=RANDOM
   /CLASSMISSING EXCLUDE
   /SAMPLEFILE OUTFILE='muestra1'
   /PRINT SELECTION.
```

Bibliografía

[1] Antonio Caporal. ¿quién vota en méxico? *Vértigo Político*, 2015.

[2] Compilación CEO. Historia de las encuestas en el mundo. *La Sociología en sus Escenarios*, 2008.

[3] Willian G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

[4] México cómo Vamos: Ana Gutiérrez, Valeria Mendiola, y Valeria Moy. Así votamos. *Animal Político*, 2018.

[5] Eugenia Coppel. ¿cómo son los mexicanos que votarán en este 2018? *Verne. El País*, 2018.

[6] Carles M. Cuadras. Distancias estadísticas. *Estadística Española*, 30(119):295–378, 1989.

[7] Dirección Ejecutiva de Capacitación Electoral y Educación Cívica. Estudio censal sobre la participación ciudadana en las elecciones federales de 2012. Inf. téc., Instituto Federal Electoral (IFE), 2013.

[8] Dirección Ejecutiva de Capacitación Electoral y Educación Cívica. Estudio censal sobre la participación ciudadana en las elecciones federales de 2015. Inf. téc., Instituto Nacional Electoral (INE), 2016.

[9] Instituto Nacional de Estadística y Geografía. *Encuesta Intercensal 2015. Síntesis metodológica y conceptual*. INEGI, 2015.

[10] Santiago de la Fuente Fernández. *Análisis de Conglomerados*. Universidad Autónoma de Madrid, 2011.

[11] Dirección Ejecutiva del Registro Federal de Electores. Informe sobre el estado de padrón electoral y la lista nominal de electores en respuesta a la solicitud formulada por el partido revolucionario institucional (atención a las observaciones y resultados de los programas de revisión y verificación). Inf. téc., Instituto Federal Electoral (IFE), 2013.

[12] H. Eulau. The american voter. by a. campbell, p. converse, w. miller, and d. stokes. *American Political Science Review*, 54(4):993–994, 1960.

[13] Wolfang Härdle y Léopold Simar. *Applied Multivariate Statistical Analysis*. TECH: Method & Data Technologies, 2003.

[14] INE. Catálogo de secciones electorales + ubicación. 2015. BDD en formato .csv, obtenida por Solicitud a través de la Plataforma Nacional de Transparencia.

[15] INE. Estadísticas y resultados electorales: cómputos distritales. 2018. BDD en formato .csv, www.ine.mx.

[16] INE. Normatividad del ine, glosario. 2018. Www.ine.mx.

[17] INEGI. Base de datos correspondiente a los resultados de la encuesta intercensal 2015. 2015. BDD en formato .csv.

[18] Gallup International. *Polling Around the World*. Gallup International Association, 2017.

[19] Joseph F. Hair Jr., William C. Black, Barry J. Babin, y Rolph E. Anderson. *Multivariate Data Analysis*. Pearson Education Limited, 2014.

[20] Arturo Barraza Macías. La encuesta: ¿método o técnica? *Apuntes sobre metodología de la investigación*, 2006.

[21] Rachel Macreadie. Public opinion rolls. Inf. téc., Parliament of Victoria, 2011.

[22] J.M. Marín y M.T Rodríguez-Bernal. *Multiple hypothesis testing and clustering with mixtures of non-central t-distributions applied in microarray data analysis*. Universidad Carlos III de Madrid, 2012. Statistics and Data Analysis.

[23] Gustavo Meixueiro y Alejandro Moreno. *El comportamiento electoral mexicano en las elecciones de 2012*. Centro de Estudios Sociales y de Opinión Pública (CESOP) & Instituto Tecnológico Autónomo de México (ITAM), 2014.

[24] Nelcy Rocío Escobar Moreno. *Análisis de conglomerados para la segmentación de mercados*. Universidad Autónoma de Madrid, 2012. ISBN 2011-6306. Documentos FCE-CID, Universidad Nacional de Colombia, Facultad de Ciencias Económicas.

[25] Luis Fernando Sánchez Murillo y Francisco de Jesús Aceves González. Campañas políticas y configuración del voto en 2006. encuestas electorales y publicidad política. *Revista Nacional de Ciencias Políticas*, 2008.

[26] Yamil Nares. Breve historia de las encuestas: El arte de observar a la democracia. *Letras Libres*, 2018.

[27] International Chamber of Commerce y ESOMAR World Research. International code on market, opinion and social research and data analytics. Inf. téc., ICC and ESOMAR, 2016.

[28] Oraculus. Poll of polls, elecciones presidenciales en méxico. 2018. Www.oraculus.mx.

[29] Francisco Abundis: Parametría. ¿quiénes eligieron a amlo como presidente? *Milenio*, 2018.

[30] Daniel Peña. *Análisis de datos multivariantes*. –, 2002.

[31] János Podani. *Introduction to the Exploration of Multivariate Biological Data*, cap. 1–3. Backhuys Publishers, 2015.

[32] Pedro López Roldán y Sandra Fachelli. *Metodología de la investigación cuantitativa*, cap. 2–3. Universidad Autónoma de Barcelona, 2015.

[33] Thomas Rusch, Ilro Lee, Kurt Hornik, Wolfgang Jank, y Achim Zeileis. Influencing elections with statistics: targeting voters with logistic regression trees. *The Annals of Applied Stastistics*, 7(3), 2013.

[34] F. Tusell. Análisis multivariante, 2016. Notas, para el curso de Estadística: Análisis Multivariante.

[35] William W.S. Wei. *Multivariate Time Series Analysis and Applications*. Wiley, 2019.