UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y
DE LA ESPECIALIZACÍON EN ESTADÍSTICA APLICADA

ESTIMACIÓN DE LA CONFORMACIÓN DE LA CÁMARA DE DIPUTADOS EN EL CONTEO RÁPIDO 2021

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

PRESENTA:
EDGAR GERARDO ALARCÓN GONZÁLEZ

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

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO, MARZO, 2022.

«Un modelo matemático, estadístico, actuarial, o de cualquier naturaleza es una representación simplificada de algún fenómeno real... modelar tiene algo de técnica y mucho de arte.»

José Salvador Zamora Muñoz

Agradecimientos

El camino que he recorrido para lograr escribir estas palabras no ha sido sencillo. Me considero una persona sumamente afortunada al haber encontrado en mi vida a personas que me han impulsado hasta llegar a convertirme en lo que soy hoy en día.

Primero y sin dudas, debo agradecer a mis padres: **Martha Fernanda González Soto y José de Jesús Alarcón García**. Hoy y siempre son mi ejemplo de personas de bien, seres humanos ejemplares que, como siempre le he dicho a todos los que me preguntan por ustedes, nunca me dieron lo que podían, siempre me dieron MÁS de lo que podían. He de decir que mis padres siempre vieron por mi hermano y por mi, cada vez que nos faltaba algo, que necesitamos ayuda, o hasta en el detalle más mínimo como darnos un consejo, mis padres han estado ahí de manera incondicional. Gracias por tanto y creo que nunca tendré palabras suficientes para agradecerles por todo lo que nos han dado. Gracias por cada pequeño esfuerzo que han hecho por nosotros.

A mi hermano, **Jesús Fernando Alarcón González**, que la verdad me enseñó que el amor incondicional no viene únicamente de los padres, sino también de los hermanos. En las buenas y en las malas.

Posteriormente a mi **familia**, a cada integrante de la misma, al igual que mis padres, debo agradecerles que son personas que no han hecho más que apoyarse los unos a los otros. Me han enseñado un ejemplo de cómo debe ser una familia de bien, una familia que apoya, que da cobijo, cariño, y principalmente, que está ahí en las buenas y en las malas siempre de manera incondicional. No muchas familias como la nuestra y por eso, doy gracias a ustedes y a la vida por haber nacido en una familia que siempre busca el bien común y que me ha apoyado en el transcurso de mi vida para llegar hasta aquí. Gracias.

A **Miriam Janeth Padilla Martínez**, gracias porque me apoyaste muchísimo en mi periplo de la maestría. Sin dudas agradezco el apoyo que me diste, las palabras de aliento y motivación siempre presentes. De igual manera, creo que jamás encontraré las palabras adecuadas para agradecer que estuviste en los momentos más difíciles de esta etapa.

A mis grandes amigos que he hecho a lo largo de mi vida: **Fernando Bautista, Daniela Espinosa, Jessica Castañeda, Jennifer Méndez y Corina Cerezo**.

Gracias por ser un apoyo en mi vida emocional, académica y profesional. Ustedes han sido una base sólida en mi crecimiento y sé que ustedes forman parte de una familia que yo mismo elegí.

A mi estimado M. en C. **José Salvador Zamora Muñoz** que fue el primero en darme una oportunidad para desarrollarme en el ámbito docente a nivel licenciatura. Te guardo una gran admiración y motivación para crecer en el ámbito actuarial, estadístico y en general profesional. Gracias por ser más que un maestro, un amigo.

A **Erick Mier Moreno** que me brindó la oportunidad de crecer aún más en el ámbito actuarial, amigo, eres una persona que admiro muchísimo y sin dudas eres un grande entre los mejores. Gracias por tanto y esperemos seguir viendo crecer nuestros proyectos para formar actuarios élite y difundir nuestra pasión que es esta bella carrera y en general la ciencia.

A mi tutora de maestría, Dra. **Lizbeth Naranjo Albarrán** porque fue contigo donde encontré que, aunque amo la actuaría, mi área favorita siempre será la estadística. Gracias a ti encontré una gran pasión y eso es algo que siempre voy a agradecerte, pues lo llevaré conmigo el resto de mi vida.

A la M. en C. **Guadalupe Eunice Campirán García** porque gracias a ti me metí más en el lenguaje de programación R que me ha dado tantas oportunidades, me ha dado bastante para comer y tener una buena vida. Eres también una gran admiración y por lo mismo, llevo tus enseñanzas conmigo.

A mi tutor de tesis, Dr. **Carlos Erwin Rodríguez Hernández-Vela** por darme la oportunidad de mi vida y apoyarlo en este proyecto que es el Conteo Rápido. Recuerdo y siempre llevaré conmigo el día que recibí el correo donde me invitabas a colaborar contigo, para mi ha sido un gran honor y orgullo trabajar apoyado de una persona tan brillante. Gracias por tu tiempo, por tu paciencia, tus regaños y enseñanzas. Me hiciste crecer y amar aún más lo que hago, gracias por marcar este camino y servir como inspiración para que este sea solo un paso más.

Y nuevamente a estas personas principalmente, y también a aquellas que quizás no menciono pero me han estimado y apoyado en mi camino, les agradezco mucho su tiempo y valoro demasiado el apoyo que me han dado. Incluyendo la gente que no está pero formó parte de esta trayectoria. Así como **a mis alumnos** que he visto crecer desde abajo y ahora veo que han comenzado a volar, es para mi un orgullo y una gran fuente de inspiración, vamos siempre por más.

Finalmente, **a ti**, por leer al menos algo de esta tesis que créeme me he esforzado mucho para que pueda aportar aunque sea un poco en tu camino o mínimo saques un dato curioso. Espero puedas sacarle provecho y no olvides de visitar mi GitHub[1]. :)

[1] https://github.com/A1arcon/R_Actuarial

Índice general

Capítulo 1

Motivación

El 6 de noviembre de 2020 mediante Acuerdo INE/CG558/2020, el Consejo General del Instituto Nacional Electoral determinó la realización del Conteo Rápido para la elección ordinaria de Diputaciones Federales por el principio de mayoría relativa, a fin de conocer las tendencias de los resultados de la votación el día de la Jornada Electoral del Proceso Electoral Federal 2020-2021. En donde se fundamenta que:

> ..."*Este Consejo General del INE es competente para determinar la realización del Conteo Rápido para la elección ordinaria de Diputaciones Federales por el principio de mayoría relativa, a fin de conocer las tendencias de los resultados de la votación el día de la Jornada Electoral del PEF 2020-2021*"...

Asimismo, mediante el acuerdo INE/CG559/2020, el Consejo General determinó ejercer la facultad de asunción parcial e implementar el Conteo Rápido en las elecciones de Gubernatura en los estados de Baja California, Baja California Sur, Campeche, Chihuahua, Colima, Guerrero, Michoacán, Nayarit, Nuevo León, Querétaro, San Luis Potosí, Sinaloa, Sonora, Tlaxcala y Zacatecas, durante sus Procesos Electorales Locales 2020-2021.[1] Es en este acuerdo donde se comenta que el interés de realizar el Conteo Rápido nace de una motivación y se menciona que:

> ..."*La realización de los conteos rápidos implica complejidades logísticas que deben atenderse mediante esquemas operativos uniformes para la recolección, transmisión y captura de datos.*
>
> *Dentro de los aspectos complejos, se destaca, determinar la metodología y el diseño de la muestra estadística de las casillas que generarán la información para construir los rangos de votación; tales actividades deben ser realizadas por un comité técnico de especialistas, por lo que se propone que sea el INE quien determine la integración del COTECORA, con base en*

[1] https://portal.ine.mx/conteos-rapidos-procesos-electorales-federal-y-locales-2020-2021/

> *las experiencias exitosas adquiridas en los pasados procesos electorales.*
>
> *Se debe resaltar que el INE cuenta con especialistas que pueden coordinar que en las entidades federativas se mantenga la misma metodología, garantizando con ello la certeza en la ejecución de esta actividad, así como en los propios resultados.*
>
> *Además, consideramos necesario y oportuno que este Instituto coadyuve con los Organismos Públicos Electorales de las entidades federativas para fortalecer la certeza en sus comicios, mediante acciones que aseguren los resultados electorales y generen mecanismos de colaboración y coordinación operativa entre las autoridades electorales locales y la nacional, a fin de reducir los costos de esta función comicial.*"...

Como parte de la motivación de la realización de esta tesis, fragmentos de ésta fueron presentados por el autor, Edgar Gerardo Alarcón González, en vivo en el *Actuarial Summit* 2021 organizado por el Colegio Actuarial Mexicano A.C. en colaboración con "AxMéxico" (Figura 1.1).

Los códigos que fueron utilizados para llevar a la práctica toda la teoría vista en esta tesis fueron escritos en lenguaje R y se pueden encontrar en el GitHub[2] del autor, así como videos referentes al congreso antes mencionado, evento de la selección de la muestra para el Conteo Rápido 2021 y resultados del Conteo Rápido 2021 que fueron presentados en la noche del mismo día de la Jornada Electoral.

[2] https://github.com/A1arcon/R_Actuarial

(a) Invitación a la plática.

(b) Reconocimiento.

Figura 1.1: Ponencia "Estimando la conformación de la Cámara de Diputados en el Conteo Rápido" por autor: Edgar Gerardo Alarcón González en el *Actuarial Summit* 2021 organizado por el Colegio Actuarial Mexicano A.C. en colaboración con "AxMéxico". Este evento fue en vivo de manera virtual y la repetición puede verse visitando el GitHub del autor o bien dando clic aquí.

Capítulo 2

Introducción

2.1. ¿Qué es el Conteo Rápido que realiza el INE?

El *Conteo Rápido* es un ejercicio estadístico que a partir de una muestra probabilística de casillas y mediante el empleo de procedimientos probabilísticos validados científicamente produce estimaciones de los resultados electorales en un tiempo corto con altos niveles de precisión. Es decir, es una tendencia de los resultados de las elecciones en cuestión, obtenida de una *muestra representativa* de la votación a nivel nacional. Las estimaciones se comunican a la población en la misma noche de la jornada electoral, en forma de intervalos de confianza, además se incluye la estimación del porcentaje de participación ciudadana en la elección.

Para la realización del conteo rápido, existe un procedimiento para obtener la muestra con la cual se harán las estimaciones de la conformación de la Cámara de Diputados, este proceso es realizado ante notario público días antes de la jornada electoral y es conocido como *Selección de la Muestra* (Subsección 5.6.1). Las tablas, llamadas *remesas*, que contienen los insumos necesarios para el cálculo de la Conformación de la Cámara se van generando y actualizando durante la jornada electoral, de acuerdo al conteo en tiempo real de los votos en las casillas seleccionadas. El proceso operativo principal del día de la jornada electoral es el siguiente:

1. El *Capacitador Asistente Electoral* (CAE), que tenga casillas en muestra, se encargará de recopilar los resultados de la votación de la elección en las casillas de su área de responsabilidad, directamente de las actas de escrutinio y cómputo, y la informará a la junta distrital que le corresponde.

2. En la junta distrital un capturista introducirá los resultados en el sistema de información del conteo rápido para su transmisión vía red INE al grupo de expertos del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA).

3. Los integrantes del COTECORA procesarán la información y realizarán las estimaciones estadísticas respectivas a fin de entregar un informe con los resultados del ejercicio al consejero presidente y a los miembros del consejo general (Subsección 7.3.3).

Los pasos 1 y 2 suceden tantas veces como la cantidad de casillas seleccionadas que hay, y son éstos los resultados que se van presentado con una latencia dada por la velocidad humana necesaria para llevar a cabo esta tarea. De tal manera que el COTECORA recibe la información de manera gradual, considerando obstáculos como los posibles contratiempos que enfrenten las casillas seleccionadas, que México tiene distintos husos horarios, entre otros del tipo geográfico, político y social. Motivo por el cual, los tiempos en los que la información es recibida y procesada en el paso 3 suelen alargarse incluso hasta un horario que podría considerarse nocturno en la Ciudad de México, que es la sede donde el INE implementa el Conteo Rápido.

Para que el Conteo Rápido sea relevante, es necesario que los resultados de las estimaciones sean presentados a más tardar al final del día de la jornada electoral. Debido a este tipo de obstáculos, se tiene a consideración un *tamaño de muestra* (Sección 5.3) apropiado que permita, que aún ante las posibles adversidades, se pueda contar con un tamaño de muestra observado lo suficientemente grande como para que las estimaciones sean lo más precisas posibles.

2.1.1. Breve historia de los Conteos Rápidos del INE

Pese a que los Conteos Rápidos estiman, con un margen de error muy pequeño, los resultados finales de una elección, su aceptación ha enfrentado una infinidad de retos a través de los años. Basta mencionar que en sus inicios fue objeto de denuncias, quejas y controversias por parte de partidos, analistas, medios de comunicación, redes sociales y ciudadanos. Por este motivo, los Conteos Rápidos han ido evolucionando a través del aprendizaje en cada proceso electoral. Esto y lo que mencionaremos a continuación puede ser consultado en [1].

2.1.1.1. Elección Presidencial de 1994

El primer conteo rápido del que se tiene registro, fue implementado en 1994 por el entonces *Instituto Federal Electoral* (IFE), en la elección presidencial de ese año. Los resultados del conteo fueron altamente coincidentes con los del cómputo final, probando así su valía para generar certidumbre sobre los resultados de una elección. Por tanto, se estableció su implementación en cada elección federal. En la Tabla 2.1, se muestran las estimaciones del CR, los resultados del *Programa de Resultados Electorales Preliminares* (PREP), así como los resultados definitivos de la elección.

2.1.1.2. Elección Presidencial de 2000

Para la elección federal del año 2000, los resultados del conteo rápido fueron esenciales la noche de la jornada electoral, ya que el PREP no tuvo resultados hasta 23 horas después del cierre de las casillas. En ese entonces, el *Comité Técnico Asesor de los Conteos Rápidos*, ahora COTECORA, se conformó con tres de los integrantes del comité de 1994 más el coordinador del PREP. Sin embargo, y al igual que en 1994 los ejercicios de estimación no los realizó directamente el IFE, si no que se contrató

	Estimaciones (%)			PREP(%)	Resultado (%)
	Mínimo	Máximo	Precisión		
PRI	49.3	50.7	0.7	50.1	50.2
PAN	26.8	28.2	0.7	28.8	26.7
PRD	15.8	17.1	0.65	17.1	17.1

Tabla 2.1: Comparación de resultados entre el conteo rápido, *Programa de Resultados Electorales Preliminares* (PREP) y cómputo final,1994.

a tres empresas encuestadoras para realizar este proceso.

Al dar a conocer las estimaciones, el factor primordial para la pronta aceptación de los resultados que arrojó el conteo rápido fue que el presidente de la República los aceptó públicamente, lo que redujo la tensión y contribuyó a que los demás actores políticos también aceptaran las estimaciones. La comparación entre las estimaciones del CR, PREP y resultados definitivos se muestran en la Tabla 2.2.

	Berumen (%) 622 casillas = 43.65 % de su muestra				Gallup México (%) 1511 casillas = 97.61 % de su muestra				Integrado por el Instituto (%)			PREP (%)	Resultado (%)
	Estimación	Mínimo	Máximo	Precisión	Estimación	Mínimo	Máximo	Precisión	Mínimo	Máximo	Precisión		
Alianza por el Cambio	43.2	41.2	45.2	2.0	42.1	40.8	43.3	1.3	39.0	45.0	3.0	42.7	42.5
Partido Revolucionario Institucional	34.7	33.3	36.2	1.5	36.6	35.5	37.6	1.1	35.0	38.0	1.5	35.8	36.1
Alianza por México	16.8	15.5	18.0	1.3	16.4	15.5	17.2	0.8	15.1	18.0	1.5	16.5	16.6

Tabla 2.2: Comparación de resultados entre el conteo rápido, *Programa de Resultados Electorales Preliminares* (PREP) y cómputo final, 2000.

Una vez más, los resultados del conteo rápido fueron coincidentes con los resultados finales de la elección. **Entonces, para las elecciones intermedias de 2003, se decidió implementar el conteo rápido para estimar la conformación de la Cámara de Diputados.** Destacando que en 2003 el entonces IFE, por primera vez se encargó de realizar el trabajo de campo a diferencia de elecciones pasadas donde se habían contratado empresas privadas.

2.1.1.3. Elecciones Locales de 2003

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

1. 300 diputados por elección directa (*Mayoría Relativa* (MR)), en cada uno de los 300 distritos electorales en los que se divide el país.

2. 200 diputados mediante *Representación Proporcional* (RP), pero aplicando algunas reglas de no sobre-representación descritas en el Reglamento de Elecciones.

Entonces, primero, era necesario que la muestra tuviera información de cada uno de los 300 distritos electorales. Ya que en cada distrito se tendría que hacer la estimación del partido ganador. Segundo, no sería posible utilizar métodos convencionales para realizar la estimación de la conformación debido a las reglas de no sobre-representación. En virtud de lo anterior, el IFE creó un nuevo Comité de Conteo Rápido, conformado por científicos que estuvieron en el ejercicio de 2000. Adicionalmente, en esta ocasión no se convocó a ninguna empresa; el IFE decidió emplear sus propios recursos, los Capacitadores y Asistentes Electorales (CAE), fueron los encargados de transmitir la información de las diferentes casillas en muestra.

Los resultados de esta elección se presentan en la Tabla 2.3. Se observa que la conformación de la cámara definitiva es estimada con gran precisión por el CR.

	Estimación del Conteo Rápido (%)				Número de Diputados		PREP (%)	Cómputos Distritales (%)	Número de Diputados	
	Puntual	Mínimo	Máximo	Precisión	Mínimo	Máximo			Asignación con base en PREP	Real
PAN	30.5	30.0	31.0	0.5	148	158	30.6	30.8	155 (154)*	151
PRI	34.4	33.9	34.9	0.5	222	227	37.4	34.6	2	222 (224)
PRD	17.1	16.6	17.6	0.5	93	100	17.7	16.6	1	96
PT	2.4	1.9	2.9	0.5	5	8	2.4	2.4	6 (5)*	5
PVEM	6.2	5.9	6.5	0.3	14	16	6.1	6.1	15 (17)*	17**
CONV	2.3	2.1	2.5	0.2	5	6	2.3	2.3	5	5
PSN	0.3	0.2	0.4	0.1			0.3	0.3		
PAS	0.7	0.6	0.8	0.1			0.7	0.7		
MP	1.0	0.9	1.1	0.1			0.9	0.9		
PLM	0.4	0.3	0.5	0.1			0.4	0.5		
FC	0.5	0.4	0.6	0.1			0.5	0.5		
** En dos de los distritos que ganó "Alianza para todos" de acuerdo al convenio, al momento de registro se cambiaron por dos de PVEM. Eso ocasionó que se redujera en uno de los plurinominales del PAN y del PT.*										
*** En la tabla de asignación final faltaron cuatro diputados en razón de la nulidad de dos elecciones.*										

Tabla 2.3: Comparación de resultados entre el conteo rápido, *Programa de Resultados Electorales Preliminares* (PREP) y cómputo final, 2003. **Los números en la tabla indican porcentajes de votación**.

Nota: En este año se hicieron 3 diferentes estimaciones utilizando los enfoques estadísticos clásico, bayesiano y fiducial. En la Tabla 2.3 se muestra el compulsado de los resultados.

2.1.1.4. Elección Presidencial de 2006

Esta elección presidencial ha sido una de las más competidas de la historia de nuestro país, con un margen de diferencia muy pequeño en los porcentajes de votación a favor del primer y segundo lugar. En esta elección el IFE decidió hacer el CR con sus propios recursos, es decir, empleó a los CAEs. Además, se determinó utilizar un

tamaño de muestra de 7,636 casillas, distribuidas en 481 estratos. Con este tamaño, se estimarían las proporciones de votos con un margen de error de 0.5 % y con una confianza mayor o igual a 95 %. Para ello, se utilizaron tres métodos estadísticos: clásico, bayesiano y robusto (en este último no se consideraba el diseño de muestra y era completamente al azar). Sin embargo, a pesar de que al momento del corte se había recibido el 95.12 % de la muestra total, se observó un traslape en los intervalos de confianza para las estimaciones de los dos candidatos punteros.

Bajo esta situación, el entonces Consejero Presidente del IFE (Luis Carlos Ugalde), anunció, acordando con el Consejo General (IFE 2006), que los resultados del procedimiento de conteo rápido no permitían anunciar a un ganador y por lo tanto no se darían a conocer las estimaciones del conteo rápido a la población. Por supuesto, esto creó incertidumbre acerca de la transparencia del CR y aumentó las dudas sobre la certeza de la elección, pues dejó al PREP como único instrumento para conocer el resultado de la elección antes de los cómputos distritales.

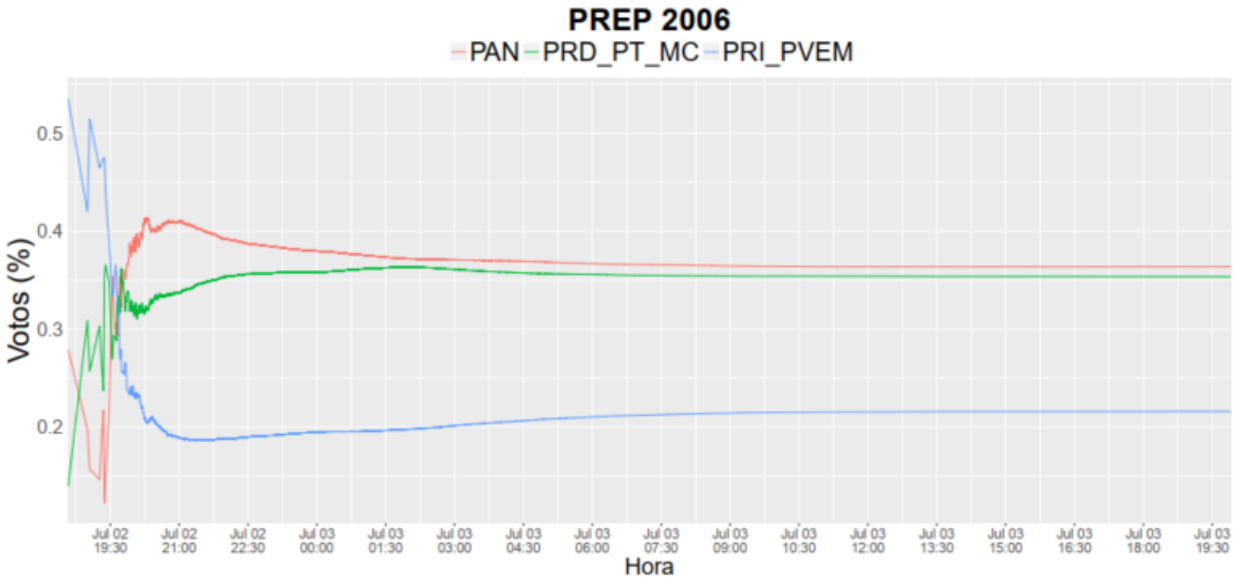


Figura 2.1: Captura de votos por el *Programa de Resultados Electorales Preliminares* (PREP) 2006.

La captura de votos realizada por el PREP comenzó a las 18:00 horas del 2 de julio de 2006, el porcentaje de avance en la recepción de votos para los tres partidos con mayor votación se presenta en la Figura 2.1[1], donde se observa que en las primeras 2 horas, el candidato de la colación `PRI_PVEM` Roberto Madrazo Pintado estaba en primer lugar, seguido por Andrés Manuel López Obrador de la coalición `PRD_PT_MC`. Por su parte, el candidato del PAN Felipe Calderón Hinojosa estaba en tercer lugar. Posterior a estas dos horas, los lugares entre el primero y el tercero se invierten co-

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

locando a Felipe Calderón en primer lugar, posición que mantendría al 3 de julio al cierre del PREP a las 20:00 horas con el 92.16 % de casillas computadas. No obstante, la diferencia entre los dos primeros lugares fue de apenas el 1.03 % de votos, esto es, el 36.37 % de votos para Felipe Calderón y 35.34 % para Andrés Manuel[2].

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

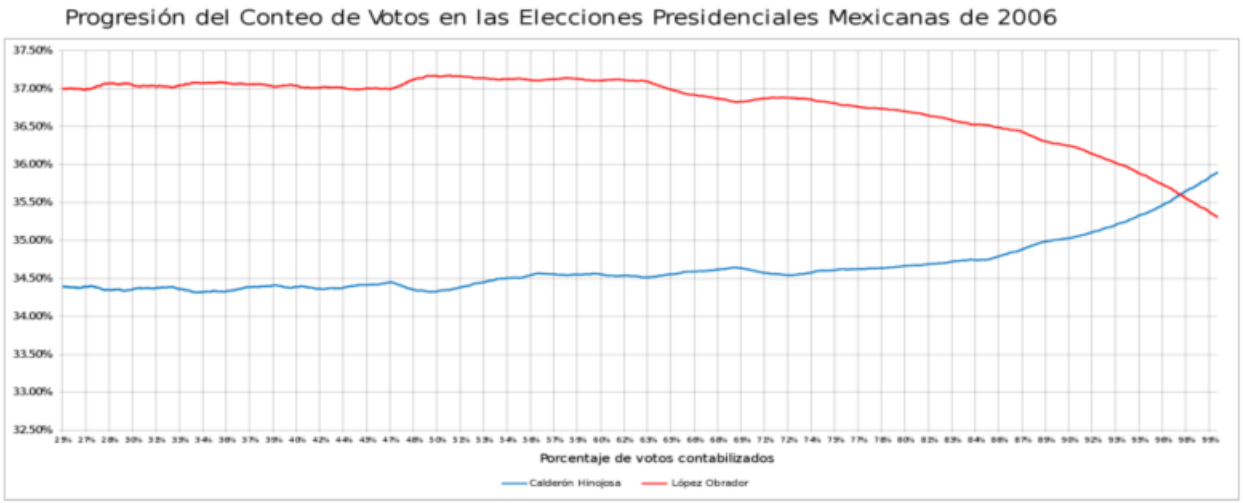


Figura 2.2: Resultados de casillas computadas por el, en ese entonces, *Instituto Federal Electoral* (IFE).

Finalmente, el jueves 6 de julio a las 3:56 horas con un 97.70 % de las casillas computadas, López Obrador pasó al segundo lugar, siendo aventajado por Felipe Calderón (ver Figura 2.2). El conteo concluyó a las 15:20 horas con el 35.89 % de los votos para Felipe Calderón, y el 35.31 % para López Obrador[3].

	Robusto (%)			Clásico (%)			Bayesiano (%)			PREP (%)	Cómputos
	Mínimo	Máximo	Precisión	Mínimo	Máximo	Precisión	Mínimo	Máximo	Precisión		Distritales (%)
PAN	35.25	37.4	1.08	35.68	36.53	0.43	35.77	36.4	0.31	36.4	35.89
APMx	20.85	22.7	0.93	21.66	22.26	0.3	21.72	22.24	0.26	21.48	22.26
CPBT	34.24	36.38	1.07	34.97	35.7	0.36	35.07	35.63	0.28	35.41	35.31
NA	0.75	1.19	0.22	0.93	1.03	0.05	0.94	1.05	0.06	0.99	0.96
ASDC	2.4	3.13	0.37	2.6	2.8	0.1	2.6	2.8	0.1	2.82	2.7

Tabla 2.4: Comparación de resultados entre el conteo rápido, *Programa de Resultados Electorales Preliminares* (PREP) y cómputo final, 2006.

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

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

En la Tabla 2.4 se presentan los resultados de esta elección, en los intervalos de confianza se observa una intersección entre primer y segundo lugar (método robusto y clásico). No obstante, obsérvese de nueva cuenta la coincidencia de resultados entre el conteo rápido, el PREP y los cómputos distritales.

2.1.1.5. Elecciones Locales de 2009

En las elecciones intermedias del 2009, para la renovación de la Cámara de Diputados, **no se realizó un conteo rápido**. Las razones de esta decisión no se conocen en su totalidad.

2.1.1.6. Elección Presidencial de 2012

El conteo rápido en la elección presidencial de 2012 fue posible gracias a que el Consejo General, de aquel entonces, especificó que los resultados del Conteo Rápido, así como los rangos de votación por candidato, cualesquiera que sean las diferencias entre ellos, se difundirían a la opinión pública[4]. En esta ocasión, el número de casillas en muestra fue de 7,597 con una estratificación semejante a la de 2006 (con 483 estratos). A pesar de que al momento del corte sólo se disponía del 82.4 % de las casillas en muestra[5], no hubo traslape alguno entre las estimaciones. Mas aún, los intervalos de confianza resultaron en una coincidencia total con los resultados del PREP y el cómputo final de la elección (ver Tabla 2.5).

Candidato	Mínimo (%)	Máximo (%)	Precisión (%)	PREP (%)	Cómputos Distritales (%)
Josefina Vázquez Mota	25.10	26.03	0.46	25.4	25.41
Enrique Peña Nieto	37.93	38.55	0.31	38.15	38.21
Andrés Manuel López Obrador	30.9	31.86	0.48	31.64	31.59
Gabriel R. Quadri de la Torre	2.27	2.57	0.15	2.3	2.29

Tabla 2.5: Comparación de resultados entre el conteo rápido, *Programa de Resultados Electorales Preliminares* (PREP) y cómputo final, 2012.

2.1.1.7. Ley General de Instituciones y Procedimientos Electorales 2014

Para el año 2014, se promulga la nueva *Ley General de Instituciones y Procedimientos Electorales* (LEGIPE), en donde se definen y describen con mayor rigor los programas

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

de resultados preliminares. Destaca el énfasis en su carácter meramente informativo y que no son definitivos. Con referencia a los conteos rápidos, la ley faculta al *Instituto Nacional Electoral* (INE), antes IFE, y a los Organismos Públicos Locales Electorales (OPLE) para ordenar su realización. Finalmente, en el reglamento de elecciones emitido por el Consejo General en 2016 se establece la obligación de realizar conteos rápidos para las elecciones de gobernador y de jefe de gobierno.

2.1.1.8. Elecciones Locales de 2015-2017

Debido a este cambio en la legislación electoral, desde 2015, se han realizado conteos rápidos para estimar las tendencias de la votación en diferentes elecciones. Particularmente, en la elección intermedia de 2015, el INE organizó un conteo rápido para conocer la composición de la Cámara de Diputados. La muestra se diseñó aleatoriamente a partir de un número fijo de casillas por cada distrito (treinta) con un ligero incremento en aquellos distritos que, por su diferencia horaria, pudieran tener problemas en el acopio de los resultados. Al final, la muestra fue de 9,450 casillas de las cuales sólo se tuvo información del 74.24 %, al momento de realizar las estimaciones. La Tabla 2.6, muestra que los intervalos de confianza del conteo rápido arrojaron una coincidencia casi total con el PREP.

	PREP (%) Corte de las 20:10 hrs. del día siguiente a la jornada electoral Avance del 98.63 %			Conteo Rápido (Estimación del COTECORA)		
	MR	RP	CONF	Mínimo	Máximo	Contiene el valor del PREP
PAN	56	53	109	105	116	SÍ
PRI	156	41	197	196	203	SÍ
PRD	28	27	55	51	60	SÍ
PVEM	27	18	45	41	48	SÍ
PT	6	7	13	3	12	NO
MC	11	15	26	24	29	SÍ
PANAL	1	10	11	9	12	SÍ
MORENA	14	21	35	34	40	SÍ
PH	0	0	0	0	1	SÍ
ES	0	8	8	8	10	SÍ
CI	1	0	1	1	1	SÍ
TOTAL	300	200	500			

Tabla 2.6: Comparación de resultados entre el conteo rápido y *Programa de Resultados Electorales Preliminares* (PREP), 2015.

Como se mencionó anteriormente, desde 2015, se empezaron a implementar conteos rápidos para estimar los resultados de varias elecciones para gobernador y para Jefe de Gobierno (CDMX 2015). El comparativo entre las estimaciones del CR y los cómputos finales de cada elección se resumen en la Tabla 2.7 (únicamente se presenta el comparativo para los tres candidatos punteros en cada elección).

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

Tabla 2.7: Conteo Rápido en elecciones de gobernador, 2015 - 2017.

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

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

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

2.1.1.9. Elecciones Locales y Presidencial de 2018

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

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

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

1. Generar problemas en las casillas a cargo de CAEs con sobrecarga de casillas para el CR.
2. Que los CAEs decidan no reportar las casillas del CR, aumentando la no-respuesta (`NA`) (ver Subsubsección 6.1.1.1).

Este segundo punto debe analizarse con cuidado, ya que en los estados en los que se ha realizado CR para estimar los resultados de la elección local, se han observado los porcentajes de no respuesta[7] mostrados en la Tabla 2.8.

Estado	No Respuesta	Hora de Corte
Sonora (2015)	**29.53 %**	**01:22 horas, siguiente día.**
Veracruz (2016)	**9.71 %**	**23:30 horas.**
Zacatecas (2016)	**10.66 %**	**22:40 horas.**
Colima (2016)	**27.8 %**	**20:31 horas.**
Oaxaca (2016)	**36.74 %**	**01:30 horas, siguiente día.**
Nayarit (2017)	**50 %**	**00:41 horas, siguiente día.**
Coahuila (2017)	**45.39 %**	**02:05 horas, siguiente día.**
México (2017)	**25.9 %**	**21:10 horas.**

Tabla 2.8: Porcentaje de no respuesta en los CR de 2015, 2016 y 2017

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

Teniendo en cuenta estos porcentajes de no respuesta, el COTECORA, para las elecciones de 2018, puso una restricción para el tamaño de muestra. Se buscó garantizar

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

que al menos el 80 % de los CAEs que participarían en el CR, les correspondiera reportar información de una sola casilla. Entonces, **se abordó sólo parcialmente el problema de sobrecarga de los CAEs**, ya que equivalió a utilizar un diseño estratificado común (como siempre se ha hecho), simplemente se buscó la mejor estratificación hasta satisfacer la restricción. Otro acuerdo fue que el margen de error en ningún caso fuera mayor al 1 %.

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

- En un gran número de casillas se tenían elecciones concurrentes (locales y federales).
- La existencia de coaliciones en varios estados dificultaban el cómputo de los votos.
- Se puso mucho énfasis en que se comunicaran estimaciones a la población antes de las 12 de la noche.

Todo lo anterior, obligando al COTECORA a realizar la estimación final con muestras incompletas.

La Tabla 2.9 muestra las estimaciones de los Conteos Rápidos para el proceso electoral 2018. Se presenta información únicamente para el primer y segundo lugar de la elección, los resultados del cómputo final y los porcentajes de no respuesta de cada elección[8]. Además, se incluye la hora de corte para las estimaciones y se ordena en orden descendente según el porcentaje de no respuesta.

	Primer lugar (%)			Segundo lugar (%)			No respuesta (%)	Corte
	Mínimo	Máximo	Cómputo Final	Mínimo	Máximo	Cómputo Final		
Morelos	51	53	52.6	13.4	16.1	14.1	20.0	22:10
Jalisco	37.7	40	39	23	25.3	24.7	26.3	22:15
Puebla	36.4	38.9	38.1	33.9	36.8	34.1	26.9	23:45
Guanajuato	49.5	51.5	49.9	23.2	25.2	24.2	28.6	21:45
Veracruz	43.9	45.9	44	37	38.7	38.4	29.9	23:10
Presidencial	50	53.8	53.2	22.1	22.8	22.3	32.5	22:30
Yucatán	38.4	40.8	39.6	34	36.5	36.1	37.3	23:55
Chiapas	40.2	44.2	39.3	19.1	22.6	22.5	41.2	00:30 (2 de Julio)
CDMX	46.6	47.7	47.1	30.4	31.2	31	41.6	22:15
Tabasco	62.1	64.3	61.4	16.8	18.4	19.6	41.6	23:50

Tabla 2.9: Conteos rápidos en elecciones de gobernador y presidencial 2018

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

Como se puede observar, hubo una coincidencia total entre los resultados del CR y el cómputo final para el primer lugar a excepción de Chiapas y Tabasco donde las estimaciones sobrestimaron los resultados finales. Adicionalmente, para el segundo lugar, ésta coincidencia es casi total, la excepción fue en el estado de Tabasco donde se subestimó por 1.2 % al porcentaje real. Adicionalmente, también se observa un traslape entre los intervalos, que corresponden al estado de Puebla. **Esto, haciendo énfasis en que el traslape no es necesariamente un problema técnico, simplemente la información recabada no permite dar una estimación final que permita concluir un resultado final certero**. A pesar de estos dos detalles, en general las estimaciones del CR de 2018 reflejaron correctamente las tendencias de las votaciones. Finalmente, nótese que los porcentajes de no respuesta son altos. Por ejemplo, para CDMX y Tabasco, esta fue del 40.6 % de las casillas en muestra con el último corte de las 22:15 y 23:50 horas, respectivamente. Por su parte, el estado donde se observó una afluencia menor de casillas fue Chiapas, pues con el corte de las 00:30 horas del día siguiente a la jornada electoral tenía una no respuesta del 40.1 % de las casillas en muestra.

2.2. Objetivos para lograr el Conteo Rápido 2021

De acuerdo con [2], la operación logística para el Conteo Rápido consideró la definición de los recursos necesarios para planear el operativo de campo, así como de las acciones que se implementaron para asegurar el adecuado flujo de la información de las casillas electorales de la muestra, para las elecciones Federal y locales, al COTECORA el día de la Jornada Electoral. En razón de lo anterior, **se buscó cumplir con los siguientes objetivos**:

- General

  Proveer, de manera confiable y oportuna al COTECORA, la información de los resultados de las votaciones asentados en los cuadernillos de las casillas electorales de las muestras correspondientes, con la finalidad de que realice las estimaciones estadísticas **para conocer las tendencias de las votaciones de la Elección Federal y de las 15 elecciones locales de Gubernatura, el día de la Jornada Electoral**.

- Específicos

  - Determinar los requerimientos para la etapa de planeación de la operación logística del Conteo Rápido.
  - Precisar las funciones que desarrollará el personal involucrado en la ejecución de la operación logística del Conteo Rápido.
  - Definir los procedimientos para la recopilación, reporte y captura de los datos de la votación emitida en cada una de las casillas electorales de la muestra.

- Definir el esquema de seguimiento para asegurar la oportunidad en la transmisión de los datos de las votaciones.
- Definir un esquema de contingencia que contemple soluciones ante complicaciones en el reporte de los resultados de las votaciones emitidas en cada una de las casillas electorales de la muestra.

Capítulo 3

La Cámara de Diputados

3.1. Conformación e Importancia de la Cámara de Diputados

De acuerdo a la Ley Orgánica del Congreso General de los Estados Unidos Mexicanos (LOCGEUM, [3]) en referencia al artículo 1o. "El Poder Legislativo de los Estados Unidos Mexicanos se deposita en un Congreso General, que se divide en dos Cámaras, una de Diputados y otra de Senadores.". Mismas que, con fundamento en el artículo 2o. de la LOCGEUM serán conformadas en aplicación de los artículos 52 y 56 de la *Constitución Política de los Estados Unidos Mexicanos* (CPEUM, [4]). En particular, el artículo 52 de la CPEUM establece que "La Cámara de Diputados estará integrada por 300 diputadas y diputados electos según el principio de votación mayoritaria relativa, mediante el sistema de distritos electorales uninominales, así como por 200 diputadas y diputados que serán electos según el principio de representación proporcional, mediante el Sistema de Listas Regionales, votadas en circunscripciones plurinominales.", mostrando así cómo debe ser la conformación de la cámara de diputados. Las candidaturas a estos puestos políticos vendrán dadas por lo establecido por cada *Fuerza Política*. Este término engloba a los partidos políticos con registro válido[1] y las candidaturas independientes que participarán en la elección en cuestión. Por otra parte, en el artículo 3o. de la LOCGEUM se menciona que la organización y el funcionamiento de la cámara de diputados estará establecido por la CPEUM, la misma LOCGEUM, las reglas de funcionamiento del Congreso General y de la Comisión Permanente, así como los reglamentos y acuerdos que cada una de ellas expida sin la intervención de la otra. La importancia de la cámara de diputados radica en sus facultades exclusivas en el congreso de la unión, de acuerdo a los artículos 220 al 229 del *Reglamento de la Cámara de Diputados del H. Congreso de la Unión* (RCD, [5]), entre las cuales destacan: 1) aprobación anual del Presupuesto de Egresos de la Federación; 2) revisión de la Cuenta Pública del año anterior; 3) aprobación del Plan Nacional de Desarrollo; y, 4) la ratificación de los funcionarios federales

[1]Para que un partido político tenga un registro válido debe apegarse a las condiciones dadas por el Instituto Nacional Electoral, la Ley General de Partidos Políticos, y La *Constitución Política de los Estados Unidos Mexicanos* (CPEUM).

establecidos en la Constitución.

3.2. Cálculo de la Conformación de la Cámara

3.2.1. Definiciones y notaciones matemáticas

Para describir el cálculo de la conformación de la Cámara, se asumirá la existencia de $k + r$ fuerzas políticas compuestas por k partidos políticos con registro válido ($\{P_j\}_{j=1}^{k}$) y r candidatos independientes ($\{I_j\}_{j=1}^{r}$) que han sido aprobados por el INE en una determinada elección de diputados a nivel nacional. Entonces,

$$\text{Fuerzas Políticas} = \{P_1, \ldots, P_k, I_1, \ldots, I_r\}.$$

Nota: Se denotará como $P_{i,j}$ a la participación del partido P_j en el distrito i con $i \in \{1, \ldots, 300\}$.

El objetivo de esta sección es describir los cálculos para determinar la *Conformación* de la Cámara de Diputados con fundamento en la *Ley General de Instituciones y Procedimientos Electorales* (LEGIPE, [6]) y poniendo de frente la antinomia[2] de la CPEUM. Es decir, buscamos determinar

$$\text{Conformación} = CONF = \{NP_1, \ldots, NP_k, NI\}. \tag{3.1}$$

En donde $NP_j \geq 0$ es el número de escaños/lugares/curules en la Cámara de Diputados que obtiene el partido P_j, y $NI \geq 0$ es el número de lugares totales que ocupan los r candidatos independientes **obtenidos por los cómputos distritales**. Es importante destacar que, de acuerdo a lo establecido en el artículo 54 de la CPEUM, los partidos políticos podrán participar en la repartición de curules por el principio de *Representación Proporcional* (RP), dejando fuera de este principio a los candidatos independientes.

De esta manera, los candidatos independientes únicamente podrán obtener escaños vía el principio de *Mayoría Relativa* (MR) mientras que los partidos políticos son los que podrán ser partícipes en ambos principios. Así, denotando como MP_j y RP_j como los lugares asignados por MR y RP respectivamente en favor del partido político P_j, tendremos que

$$NP_j = MP_j + RP_j, \qquad \forall j \in \{1, \ldots, k\}.$$

Mientras que los candidatos independientes serán agregados para efectos de la presentación del Comité Técnico Asesor de los Conteos Rápidos (COTECORA) durante la jornada electoral. De tal manera que si denotamos como $MI_j \in \{0, 1\}$, con

[2]En México, cuando existe una laguna en una ley o artículo siempre se pondrá de frente lo establecido en la Constitución Política de los Estados Unidos Mexicanos.

$j \in \{1, \ldots, r\}$, a los lugares asignados[3] por MR en favor del candidato independiente I_j. Entonces,

$$NI = \sum_{j=1}^{r} MI_j$$

Por consistencia y de acuerdo a lo mencionado en la sección anterior, se deben cumplir las siguientes igualdades:

1. La suma de curules otorgados por el principio de MR es igual a 300, *ie*,

$$\sum_{j=1}^{k} MP_j + NI = 300.$$

2. La suma de curules otorgados por el principio de RP es igual a 200, *ie*,

$$\sum_{j=1}^{k} RP_j = 200.$$

3. La suma de curules otorgados por ambos principios es igual a 500, *ie*,

$$\sum_{j=1}^{k} (MP_j + RP_j) + NI = \sum_{j=1}^{k} NP_j + NI = 500.$$

Esto, sujeto a que, con fundamento en el artículo 54 de la CPEUM, base IV, ningún partido podrá contar con más de 300 diputados por ambos principios. Es decir, que $NP_j \leq 300$.

3.2.2. Clases de votaciones contempladas en la Ley

Debido a la naturaleza de los principios de MR y RP, cada uno de ellos tiene diferentes consideraciones para otorgar escaños a las fuerzas políticas participantes en la elección en cuestión. De tal manera que antes de describir cada uno de estos procesos vamos a definir las clases de votaciones que son consideradas para su implementación. Denotaremos como:

$$\text{“Votos Totales Asignados al partido } P_j \text{ en el distrito } i\text{”} = VTP_{i,j},$$

con $j \in \{1, \ldots, k\}$ e $i \in \{1, \ldots, 300\}$ para cada uno de los 300 distritos federales. A continuación y previo a definir formalmente las clases de votaciones contempladas en la ley se explicará en la siguiente subsección lo que sucede con los votos aportados hacia las *coaliciones* entre partidos políticos.

[3] Al contrario de un partido político, el cual tiene diferentes candidatos en los Distritos Federales en los que participa, a un candidato independiente solo se le puede asignar la diputación por la que se está postulando.

3.2.2.1. Coaliciones

De acuerdo al portal web de "Central Electoral - INE"[4] una coalición puede ser formada por dos o más partidos para postular candidaturas en común por el principio de MR y a la presidencia, registrando su *convenio de coalición* ante el INE, en caso de elecciones Federales y ante el *Organismo Público Local Electoral* (OPLE), en casos de elecciones locales. Una nota importante es que independientemente de la coalición, cada partido debe registrar listas propias de candidaturas por el principio de RP. Asimismo, al terminar la etapa de resultados y declaraciones de validez, la coalición se da por concluida automáticamente, de tal manera que las y los candidatos electos quedarán comprendidos en el partido o grupo parlamentario señalado en el *convenio de coalición.*

En cada una de las 300 elecciones por el principio de MR puede haber coaliciones entre diferentes partidos. En caso de coaliciones, su impacto residirá en términos de los $VTP_{i,j}$ tomando en cuenta las siguientes consideraciones.

- En la boleta electoral los votantes pueden elegir de forma válida cualquier combinación de partidos en la coalición. Por ejemplo y sin pérdida de generalidad, si en el distrito i-ésimo hay una coalición de dos partidos $P_{i,1}$ y $P_{i,2}$, los votos válidos en la boleta electoral resultan de cualquiera de las siguientes combinaciones $\{P_{i,1}\}$, $\{P_{i,2}\}$ y $\{P_{i,1}, P_{i,2}\}$. Esto, en otras palabras, significa que cuando hay una coalición con n partidos en el distrito i, $\mathcal{C}_i = \{P_{i,j}\}_{j=1}^{n}$, entonces los votos válidos vendrán dados por las combinaciones producidas por el conjunto potencia de $\mathcal{C}_i$, sin incluir al conjunto nulo. Es decir,

$$\begin{aligned} \text{Partidos en una coalición arbitraria en el distrito } i = \{P_{i,j}\}_{j=1}^{n} = \mathcal{C}_i, \quad \text{y así,} \\ \text{Votos válidos en la boleta electoral} = \mathcal{P}(\mathcal{C}_i) - \{\varnothing\} = \mathcal{P}_{\geq 1}(\mathcal{C}_i) = \mathcal{C}_i^* \end{aligned}$$

  Donde, en general, denotaremos como $\mathcal{P}_{\geq n}(A)$ al conjunto potencia de A cuyos elementos tienen una cardinalidad mayor o igual a $n \in \mathbb{N}$. Notando que si $m \leq n$ son dos números naturales, entonces $\mathcal{P}_{\geq n}(A) \subseteq \mathcal{P}_{\geq m}(A) \subseteq \mathcal{P}(A)$. Como un comentario adicional, en un mismo distrito pueden existir múltiples coaliciones compuestas por partidos P_j, sin embargo, no puede haber un mismo partido en dos o más coaliciones diferentes, es decir, su intersección es vacía. Esto es, si existen $\mathcal{C}_i^{(1)}$ y $\mathcal{C}_i^{(2)}$ con $\mathcal{C}_i^{(1)} \neq \mathcal{C}_i^{(2)}$, entonces $\mathcal{C}_i^{(1)} \cap \mathcal{C}_i^{(2)} = \varnothing$.

  En otro ejemplo, si en el i-ésimo distrito hay una coalición entre tres partidos, sin pérdida de generalidad digamos $\mathcal{C}_i = \{P_{i,1}, P_{i,2}, P_{i,3}\}$, entonces las formas válidas de votar serían seleccionando cualquiera de las combinaciones en

$$\mathcal{C}_i^* = \{\{P_{i,1}\}, \{P_{i,2}\}, \{P_{i,3}\}, \{P_{i,1}, P_{i,2}\}, \{P_{i,1}, P_{i,3}\}, \{P_{i,2}, P_{i,3}\}, \{P_{i,1}, P_{i,2}, P_{i,3}\}\}$$

[4] https://centralelectoral.ine.mx

Nota: Si no hay coaliciones en un determinado distrito, los votos válidos son sólo aquellos para cada partido de manera directa. Otra manera de pensarlo es que $\mathcal{C}_i$ puede estar compuesta de un solo partido, significando que éste no se encuentra coaligado con algún otro.

- Si hay una coalición en el distrito i-ésimo, la suma de los votos totales obtenidos por todas las formas válidas de votar por la coalición se usará en favor del candidato de la coalición. Haciéndo énfasis en que esta parte es particular para MR.

- Si una coalición obtiene el triunfo en el distrito i-ésimo, se otorgará dicho triunfo al partido coaligado en términos del acuerdo INE/CG193/2021 tomando en cuenta la *afiliación efectiva*, en su caso la adscripción a grupo parlamentario o bien lo establecido en el convenio de coalición.

- Para determinar la votación nacional de cada partido coaligado se considerará la regla dispuesta en el artículo 311, numeral 1, inciso c) de la LEGIPE. En virtud de lo anterior los votos individuales para cada partido no se reparten. Por ejemplo y sin pérdida de generalidad, en el caso de una coalición de los partidos $P_{i,1}$ y $P_{i,2}$ en el distrito i-ésimo, se tendrían que repartir los votos obtenidos por la combinación $\{P_{i,1}, P_{i,2}\}$ entre los partidos $P_{i,1}$ y $P_{i,2}$. Esto se consigue siguiendo el siguiente procedimiento:

  1. Se calcula el número total de votos de la combinación $P_{i,1}$ y $P_{i,2}$, *i.e.*

  $$\nu_1 = V(\{P_{i,1}, P_{i,2}\})$$

  2. Se obtienen la cantidad $\nu_2 = \left\lfloor \frac{\nu_1}{2} \right\rfloor$,[5] y el residuo $\nu_3 = \nu_1 - 2\nu_2$.
  3. Se calculan los votos para cada partido considerando tanto los votos directos, como los votos conjuntos, *i.e.*

  $$VTP_{i,1} = V(\{P_{i,1}\}) + \nu_2 \quad \text{y} \quad VTP_{i,2} = V(\{P_{i,2}\}) + \nu_2$$

  4. Se asigna el residuo al partido con más votos (de forma individual), *i.e.* si $VTP_{i,1} \geq VTP_{i,2}$, entonces se actualiza el valor $VTP_{i,1} = VTP_{i,1} + \nu_3$ y en otro caso se actualiza $VTP_{i,2} = VTP_{i,2} + \nu_3$.

  Donde $V : \mathcal{P}(\text{Fuerzas Políticas}) \to \mathbb{N}$ es una función tal que $V(A)$ denota la cantidad de votos marcados directamente en la boleta electoral de forma válida para A. Obsérvese que, $\mathcal{P}(\text{Fuerzas Políticas})$ son todas las formas posibles en las que una persona puede marcar su voto en la boleta electoral, sin embargo, existen casos donde el voto es nulo y por lo tanto dicho total sería cero. Por ejemplo, $V(\varnothing) = 0$ pues al entregar la boleta vacía no se registra como voto en favor de alguno de los candidatos, $V(\{P_{i,j}, I_{i,s}\}) = 0$ pues un candidato

[5] $f(x) = \lfloor x \rfloor$ es la función mayor entero menor o igual de $x \in \mathbb{R}$.

independiente no puede pertenecer a una coalición y por lo tanto si algún votante marca la boleta electoral de esta manera se considerará anulada, etc. Es decir,

$$Ker(V) = \{A : A \text{ no es una opción válida para votar.}\}.$$

Suponiendo que en el distrito i-ésimo existiera una coalición entre tres partidos, siguiendo un procedimiento análogo se tendrían que repartir los votos obtenidos por las combinaciones, *s.p.g.*, $\{P_{i,1}, P_{i,2}\}$, $\{P_{i,1}, P_{i,3}\}$, $\{P_{i,2}, P_{i,3}\}$ y $\{P_{i,1}, P_{i,2}, P_{i,3}\}$ entre cada partido de la combinación. Podemos describir el algoritmo anterior para obtener las $VTP_{i,j}$ en el caso de que el partido P_j pertenezca a una coalición de n partidos en el distrito i generalizando el procedimiento anterior.

Nota: En el ejemplo mencionado de una coalición compuesta por 3 partidos políticos, en particular $\{P_{i,1}, P_{i,2}\}, \{P_{i,1}, P_{i,2}, P_{i,3}\} \in \mathcal{P}_{\geq 2}(\mathcal{C}_i)$. Esta observación es importante ya que, sin pérdida de generalidad, $P_{i,1}$ está recibiendo votos por coalición tanto de $\{P_{i,1}, P_{i,2}\}$ como de $\{P_{i,1}, P_{i,2}, P_{i,3}\}$ por lo que es necesario cuidar la manera en que se reparten los votos.

En propuesta de un algoritmo de generalización[6], se deberán inicializar los votos totales asignados a $P_{i,j}$ como $VTP_{i,j} = 0$, para todo partido y en todos los distritos, de tal manera que iremos acumulando los votos para el partido P_j dependiendo de su participación en $\mathcal{C}_i^*$. En el caso en el que en el distrito i el partido $P_{i,j}$ no pertenezca a una coalición, sus votos serán los otorgados directamente, es decir, $VTP_{i,j} = V(\{P_{i,j}\})$. En otro caso, supongamos la existencia de un acuerdo de coalición en el distrito i entre los partidos que conforman el conjunto $\mathcal{C}_i$. Luego, consideremos la combinación $C \in \mathcal{P}_{\geq 2}(\mathcal{C}_i)$, procedemos a repartir los votos de C entre los $m = |C|$[7] partidos que lo conforman de acuerdo al siguiente algoritmo:

1. Se calcula el número total de votos de la combinación C, *i.e.*

$$\nu_1 = V(C)$$

2. Se obtienen la cantidad $\nu_2 = \left\lfloor \frac{\nu_1}{m} \right\rfloor$, y el residuo $\nu_3 = \nu_1 - m\nu_2$.
3. Se actualizan los votos para cada uno de los m partidos de C considerando tanto los votos directos, como los votos conjuntos. Con una variable indicadora se contemplará si ya anteriormente fueron agregados los votos procedentes de otro elemento $A \in \mathcal{P}_{\geq 2}(\mathcal{C}_i)$ con $A \neq C$, *i.e.*,

$$VTP_{i,j} = V(\{P_{i,j}\})\mathbb{1}\left(VTP_{i,j} = 0\right) + VTP_{i,j}\mathbb{1}\left(VTP_{i,j} > 0\right) + \nu_2$$

[6]Esta generalización es relevante ya que, aunque no ha ocurrido en las últimas elecciones, han habido registros de coaliciones con, por ejemplo, 5 partidos políticos involucrados. Tal es el caso de la coalición "*Alianza por México*" en las elecciones del año 2000.

[7]Sea A un conjunto, denotamos como $|A|$ a la cardinalidad del conjunto A.

Nota: La indicadora nos ayuda a actualizar el total de votos de tal manera que no estemos sumando los votos otorgados de forma directa únicamente al partido $P_{i,j}$ si ya anteriormente lo hicimos.

4. Se asigna el residuo al partido con más votos (de forma individual), *i.e.* si

$$V(P_{i,x}) = \max_{P_{i,j} \in \mathcal{C}_i} \{V(P_{i,j})\}$$

Entonces actualizamos $VTP_{i,x} = VTP_{i,x} + \nu_3$

De esta manera tendremos una generalización de cómo distribuir los votos marcados en favor de una coalición en las boletas electorales entre los partidos participantes en la coalición. Este es un aspecto fundamental a contemplar al momento de otorgar curules bajo el principio de RP ya que, como se explicará en una sección posterior, dicho principio está cimentado en la cantidad de votos que resultaron en favor de un partido político.

3.2.2.2. Votación Total Emitida

De acuerdo con el artículo 15 de la LEGIPE, la *Votación Total Emitida* (VTE), es la suma de todos los votos depositados en las urnas. Para obtener los curules otorgados por RP es necesario obtener la VTE por partido a nivel nacional.

La VTE en el distrito i-ésimo se puede escribir de la siguiente manera:

$$VTE_{i,\cdot} = \{VTP_{i,1}, \ldots, VTP_{i,k}, VTI_{i,1}, \ldots, VTI_{i,r}, VTNR_i, VTN_i\}$$

en donde $VTP_{i,j}$, como fue construida anteriormente, es la votación total para P_j en el i-ésimo distrito, $VTI_{i,s}$ es la votación total[8] por el s-ésimo candidato independiente en el mismo distrito, y análogamente, $VTNR_i$ votos totales a candidatos no registrados y VTN_i que son los votos nulos.

Finalmente, la VTE desagregada por partido, candidatos independientes, votos por candidatos no registrados y nulos está dada por:

$$VTE_d = \{VTP_1, \ldots, VTP_k, VTI_1, \ldots, VTI_r, VTNR, VTN\},$$

donde la *Votación Total Emitida* (VTE) por partido a nivel nacional está dada por $VTP_j = \sum_{i=1}^{300} VTP_{i,j}$, para $j \in \{1, \ldots, k\}$ y de forma análoga se obtiene la votación total emitida para los candidatos independientes, candidatos no registrados y votos nulos sumando sobre todos los distritos. En la aplicación del artículo 15 de la LEGIPE, la VTE se obtiene sumando los componentes de la VTE_d, *i.e.*

$$VTE = \sum_{j=1}^{k} VTP_j + \sum_{s=1}^{r} VTI_s + VTNR + VTN.$$

[8]Dada directamente de las boletas electorales. Es decir $VTI_{i,s} = V(I_{i,s})$ sabiendo que los candidatos independientes no pertenecen a alguna coalición.

3.2.2.3. Votación Válida Emitida

De acuerdo con el artículo 15 de la LEGIPE, en la aplicación de la fracción II del artículo 54 de la CPEUM, se entiende como *Votación Válida Emitida* (VVE) a la VTE menos los votos nulos y los correspondientes a candidatos no registrados, esto es:

$$VVE = VTE - (VTNR + VTN) = \sum_{j=1}^{k} VTP_j + \sum_{s=1}^{r} VTI_s.$$

La VVE desagregada se define consecuentemente como:

$$VVE_d = \{VTP_1, \ldots, VTP_k, VTI_1, \ldots, VTI_r\}, \tag{3.2}$$

y análogamente podemos definir la VVE en el distrito i-ésimo se puede escribir de la siguiente manera:

$$VVE_{i,\cdot} = \{VTP_{i,1}, \ldots, VTP_{i,k}, VTI_{i,1}, \ldots, VTI_{i,r}\}.$$

De acuerdo con el párrafo adicionado DOF 10-02-2014 en la fracción I del artículo 41 de la CPEUM, el j-ésimo partido político preservará su registro a nivel nacional, si el porcentaje de su votación total con respecto a la VVE es mayor o igual a 3 %, i.e. si entonces el partido P_j conserva su registro a nivel nacional. En donde

$$q_j = 100 \times \left(\frac{VTP_j}{VVE}\right) \%, \qquad \text{para } j \in \{1, \ldots, k\}.$$

Por este motivo, uno de los objetivos del Conteo Rápido es estimar los porcentajes $q_1, \ldots, q_k$. Como puede apreciarse la suma de estos porcentajes no será 100 % pues no se están considerando los candidatos independientes que también son parte de la VVE.

3.2.2.4. Votación Nacional Emitida

Con fundamento en el artículo 15-2 de la LEGIPE, en la aplicación de la fracción III del artículo 54 de la CPEUM se define la *Votación Nacional Emitida* (VNE), que es la que resulte de deducir de la *Votación Total Emitida* (VTE) los votos a favor de los partidos políticos que no hayan obtenido el tres por ciento de dicha votación, los votos emitidos para candidatos independientes, los votos para candidatos no registrados y los votos nulos. La VNE estará conformada únicamente por partidos políticos (sin candidatos independientes) y dada por el total de votos de aquellos partidos que cumplen con tener al menos el 3 % de la VNE. Entonces, podemos definir a la VNE desagregada como:

$$VNE_d = \{VTP_1 \mathbb{1}(q_1 \geq 3\,\%), \ldots, VTP_k \mathbb{1}(q_k \geq 3\,\%)\} \tag{3.3}$$

y sumando cada componente se obtiene la VNE como

$$VNE = \sum_{j=1}^{k} VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right). \tag{3.4}$$

Nota: De acuerdo con LEGIPE, artículo 15-2, se entiende que la regla para que un partido cumpla con una cantidad de votos para ser contemplado en la repartición por RP se calcula sobre la VTE. Sin embargo, en ese mismo artículo se toma como fundamento la CPEUM, artículo 54-2, donde la regla se calcula sobre la VVE. Por lo que, debido a la antinomia, la regla de selección de partidos será tomada como se indica en la Constitución.

3.2.3. Mayoría Relativa

De acuerdo al Sistema de Información Legislativa[9], y aplicando a los artículos 54, 63, 77, 116 y 122 de la CPEUM, *Mayoría Relativa* es un "tipo de votación que tiene por principio elegir a quien tenga el mayor número de votos emitidos. Consiste en que el candidato o asunto sometido a votación obtiene el triunfo o aprobación con el mayor número de votos emitidos. Así, se reúne una mayoría relativa cuando un grupo o candidato tiene un número de votos mayor a los elementos que contiene cualquier otro grupo, considerados separadamente.".

Fundamentado en el artículo 53 de la CPEUM, el territorio nacional se divide en 300 Distritos Electorales Federales. Los 300 escaños otorgados por el principio de MR se obtienen al realizar elecciones independientes en cada uno de estos 300 Distritos Electorales Federales. En este caso, el contendiente que obtenga el mayor número de votos se queda con el escaño en la Cámara de Diputados correspondiente al distrito por el que compitió. En cada distrito puede ganar sólo un candidato de algún partido político o un candidato independiente. Entonces, el vector de escaños que obtiene cada fuerza política por la vía de MR está dado por:

$$MP = \{MP_1, \ldots, MP_k, NI\} \tag{3.5}$$

Donde MP_j representa el total de escaños otorgados por MR al partido P_j y NI representa el total de escaños otorgados a candidatos independientes por MR. Para obtener el vector MP es necesario considerar si hay coaliciones o no. En caso de que existan coaliciones, para determinar al partido al que pertenece el candidato ganador se deben tomar en cuenta los convenios de coalición, así como el acuerdo INE/CG193/2021 en lo referente a la afiliación efectiva y a la pertenencia a grupo parlamentario. Esto se retomará en un capítulo más adelante.

Entonces, tenemos tres casos para partidos políticos, consideremos a un partido arbitrario $P_{i,j}$:

[9] http://www.sil.gobernacion.gob.mx/Glosario/definicionpop.php?ID=153

1. En el caso de que $P_{i,j}$ pertenezca a una coalición en el distrito i ($\mathcal{C}_i$), y fuese el elegido por el convenio de coalición, entonces los votos totales para MR de dicho partido serán los acumulados por toda la coalición. Es decir,

$$VTP^*_{i,j} = \sum_{P_{i,t} \in \mathcal{C}_i} VTP_{i,j}.$$

2. En el caso de que $P_{i,j}$ pertenezca a una coalición en el distrito i ($\mathcal{C}_i$), y no fuese el elegido por el convenio de coalición, entonces los votos totales para MR de dicho partido serán cero, pues fueron otorgados al partido elegido. Esto es,

$$VTP^*_{i,j} = 0.$$

3. En el caso de que $P_{i,j}$ no pertenezca a una coalición en el distrito i, o bien no exista un acuerdo de coalición en ese distrito, entonces los votos totales para MR de dicho partido serán los mismos que los que fueron dados por $VTP_{i,j}$. Lo que significa que:

$$VTP^*_{i,j} = VTP_{i,j}.$$

Finalmente, definimos $VVE^*_{i,\cdot}$ de forma natural como su correspondiente $VVE_{i,\cdot}$ de la siguiente manera:

$$VVE^*_{i,\cdot} = \{VTP^*_{i,1}, \ldots, VTP^*_{i,k}, VTI_{i,1}, \ldots, VTI_{i,r}\}.$$

Por lo tanto, los elementos del vector de escaños que obtiene cada fuerza política por la vía de MR[10] están dados por:

$$MP_j = \sum_{i=1}^{300} \mathbb{1}\left(VTP^*_{i,j} = \text{máx}\{VVE^*_{i,\cdot}\}\right) \quad \text{y} \quad NI = \sum_{s=1}^{r} \sum_{i=1}^{300} \mathbb{1}\left(VTI_{i,s} = \text{máx}\{VVE^*_{i,\cdot}\}\right).$$

Es así como de los 500 diputados que conforman la cámara, 300 de ellos son seleccionados por MR de acuerdo a este algoritmo y a lo contemplado en la sección de coaliciones. Los otros 200 diputados son seleccionados por RP, tema que abordaremos en la siguiente sección.

3.2.4. Representación Proporcional

El objetivo de RP es que las mujeres y hombres que los partidos políticos consideren idóneos por su experiencia y conocimiento les sea garantizado un lugar y representatividad en la Cámara de Diputados. Asimismo, este principio permite que existan representantes en la cámara de diputados de partidos políticos no populares que posiblemente no alcancen escaños por el principio de MR. Los insumos para calcular los 200 escaños otorgados por el principio de Representación Proporcional, son:

[10] Dicho vector está denotado como MP en (3.5).

- El número de escaños obtenidos por MR de cada fuerza política (MP).
- La *Votación Nacional Emitida* (VNE).

A partir de esta información se reparten los 200 diputados por el principio de RP como se describe en la siguiente sección, comenzando primero con una repartición inicial. Debido a condiciones que establece la CPEUM puede ser necesario realizar más de una repartición con el objetivo de satisfacer los criterios correspondientes. Asimismo, únicamente los partidos que conservarán su registro, es decir tales que $q_j \geq 3\,\%$ podrán participar en la repartición de curules por este principio. De tal manera que se utilizará la variable indicadora $\mathbb{1}\left(q_j \geq 3\,\%\right)$ para tener control sobre los partidos que entran o no a la repartición por RP. En general, para esta explicación se buscará mantener contemplados a los k partidos políticos a través de técnicas ya anteriormente utilizadas como la actualización de variables.

3.2.4.1. Repartición Inicial

Inicialmente, se buscará repartir los 200 curules disponibles por el principio de RP entre los partidos que adquirieron este derecho. Con fundamento en la LEGIPE, artículo 16-2, se calcula el *Cociente Natural* (CN) como $CN = \frac{VNE}{200}$. Este valor se puede interpretar como la cantidad de votos necesarios para que se otorgue un escaño por el principio de RP. Para realizar la *primera repartición*, se determinarán los curules que se le asignarían a cada partido político como[11]

$$C_j = \left\lfloor \frac{VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)}{CN} \right\rfloor \quad \text{para } j \in \{1, \ldots, k\},$$

pues, así como se establece en la constitución, es la fragmentación entera de la VNE por partido político. Sin embargo, de esta forma existe el posible inconveniente de que los 200 curules a otorgar por RP no sean repartidos por completo ya que

$$200 = \frac{VNE}{CN} = \sum_{j=1}^{k} \frac{VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)}{CN} \geq \sum_{j=1}^{k} C_j.$$

Ya que que es necesario otorgar exactamente 200 escaños por RP y gracias a la última observación, es posible, dependiendo del comportamiento de los votos, que sea necesario repartir una cantidad de curules sobrantes dados por el residuo

$$R = 200 - \sum_{j=1}^{k} C_j \in \mathbb{N}.$$

En particular, existe una cota dada por $R < k$, pues

$$R = \sum_{j=1}^{k} \left(\frac{VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)}{CN} - C_j \right) < \sum_{j=1}^{k} 1 = k.$$

[11] Ver LEGIPE, artículos 16-3 y 17-1.

Estos escaños sobrantes, de acuerdo a lo establecido en la LEGIPE en los artículos mencionados, son repartidos siguiendo el procedimiento del *resto mayor*:

1. Se calculan los residuos particionando por partido político (P_j), *i.e.*

$$R_j = \frac{VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)}{CN} - C_j \quad \text{para } j \in \{1, \ldots, k\}.$$

2. Se ordenan los partidos políticos de mayor a menor de acuerdo a sus R_j.

3. Se le asigna un diputado a cada partido de acuerdo a este orden hasta que no quedan más diputados que repartir.

   Nota: Observe que $R < k$ implica que nunca será necesario otorgar más de un escaño a un mismo partido por este procedimiento, y de esta manera, se estarán repartiendo exactamente 200 diputados por RP.

Denotamos $M_j \in \{0, 1\}$ a los lugares otorgados al partido P_j, para $j \in \{1, \ldots, k\}$, por el procedimiento de *resto mayor*[12]. El número total de diputados por el principio de RP, que hasta este punto tiene cada partido, se obtiene vía

$$RP_j = C_j + M_j \quad \text{para } j \in \{1, \ldots, k\},$$

Donde, por construcción

$$\sum_{j=1}^{k} RP_j = 200.$$

Por lo tanto, el número total de curules que le corresponderían al partido P_j está dado por la suma de diputados obtenidos vía los principios de MR y RP, *i.e.*

$$NP_j = MP_j + RP_j.$$

Nota: No debe olvidarse que los candidatos Independientes forman parte de la conformación de la cámara de diputados. Recordemos que éstos no reciben curules por el principio de RP, sin embargo, al momento de presentar la conformación éstos sí deben ser incluidos. Contemplando además, que los candidatos independientes sí forman parte de la VVE.

Con esto tenemos una primera composición de la conformación de la cámara de diputados en (3.1). Esta repartición de escaños debe satisfacer ciertas condiciones establecidas tanto en la LEGIPE como en la CPEUM, de tal manera que si esto no ocurre, será necesario hacer modificaciones establecidas en estos mismos documentos para evitar la sobrerrepresentación.

[12] Por construcción, $\sum_{j=1}^{k} M_j = R$.

3.2.4.2. Verificación de no Sobrerrepresentación

En una democracia ideal, la conformación de la cámara de diputados representaría fielmente a cada uno de los sectores de la sociedad. En el caso de México, los 500 diputados deberían estar repartidos de forma proporcional de acuerdo a la VVE. Sin embargo, la ley permite tener una sobrerrepresentación limitada estableciendo dos cotas:

1. Ningún partido puede tener más de 300 diputados por ambos principios.
2. Ningún partido puede tener mayor número de escaños que lo que resulte de considerar el 8 % de su participación en la VNE.

La Cámara de Diputados está compuesta por 500 escaños. No obstante, algunos de estos los podrían ganar candidatos independientes y las reglas de no sobrerrepresentación aplican sólo a los partidos políticos. Recordemos que NI denota el número total de escaños que obtuvieron los candidatos independientes vía el principio de MR. Conforme al artículo 54 de la CPEUM, base V, ha de considerarse el número de diputaciones por ambos principios. Así, el criterio de mayor equilibrio[13] consiste en una cota dada por el porcentaje de la *Votación Nacional Emitida* (VNE) más ocho puntos porcentuales para cada partido que represente un porcentaje del total de la Cámara (500 curules), es decir:

$$NP_j \leq U_j \mathbb{1}\left(q_j \geq 3\,\%\right) \quad \text{para } j \in \{1, \ldots, k\}, \tag{3.6}$$

donde

$$U_j = \text{máx}\left\{0, \text{mín}\left\{300, 500\left(\frac{VTP_j}{VNE} + 8\,\%\right)\right\}\right\} \tag{3.7}$$

Obsérvese que este límite está cimentado en la VNE y depende de la cantidad de votos que se le dan directamente a cada partido o bien por distribución por coaliciones (VTP_j). Esto significa que cuando un partido recibe "pocos" votos, entonces su límite U_j será también bajo. Particularmente, este límite no está relacionado con los resultados provenientes de MR.

La LEGIPE, artículo 17-2, indica que el partido que no cumpla con esta cota, se le deberán quitar escaños otorgados por RP, hasta que se satisfaga la condición. Con fundamento en el artículo 15-3 de la misma, este principio no aplicará a los escaños obtenidos por el principio de MR. Por lo que, debido a los convenios de coalición, podría haber partidos que violen su correspondiente cota U_j, sin posibilidad a realizar ningún ajuste, pues dichos escaños son otorgados precisamente por el principio de MR.

[13] En concordancia con el acuerdo primero inciso b) del INE/CG193/2021

Si ningún partido supera el límite U_j entonces la conformación estará ya establecida y no será necesario realizar ningún procedimiento adicional. Por otro lado, en el caso de que algún partido exceda la cota de no sobrerrepresentación, (3.6), es necesario realizar un ajuste en donde se reasigne[14] el exceso de curules otorgados entre los demás partidos, en los términos del artículo 18-1 de la LEGIPE. Este punto involucra un proceso iterativo, ya que:

- Puede haber varios partidos que excedan la cota.
- Al reasignar escaños, es necesario verificar nuevamente que se cumpla (3.6).

En caso de que alguno de los partidos a los que se le reasignaron escaños exceda la cota, hay que repetir el proceso hasta que se hayan reasignado todos los escaños en exceso y se cumpla (Ecuación 3.6). Este proceso termina cuando se satisfaga el criterio por cada partido. El algoritmo es el siguiente, donde la parte iterativa comienza en el paso 2:

1. Se inicializan variables indicadoras que nos dirán si un partido está (o estuvo; ver paso 3) en exceso, adicionalmente se descartan partidos que no contribuyen a la VNE:

$$G_j = 1 - \mathbb{1}\left(q_j \geq 3\,\%\right) = \mathbb{1}\left(q_j < 3\,\%\right) \quad \text{para } j \in \{1, \ldots, k\},$$

   A priori, esta indicadora nos permitirá descartar a los partidos que no aportan a la VNE, lo cual nos permite mantener la distribución de plurinominales en k partidos.

2. Se calcula el exceso de escaños para cada partido como:

$$E_j = (NP_j - U_j)(1 - G_j)\mathbb{1}\left(NP_j > U_j\right), \quad \text{para } j \in \{1, \ldots, k\}.$$

   Donde $E_j = NP_j - U_j > 0$, si y sólo si:

   a) P_j es uno de los partidos que contribuyen a la VNE y no estuvo en exceso en iteraciones estrictamente anteriores (ver paso 3), y,

   b) Está en exceso en esta iteracion, *i.e.* $NP_j > U_j$.

   Nota: Recordemos que $NP_j = MP_j + RP_j$, *i.e.*, E_j depende también de los resultados derivados del principio de MR y no solamente del principio de RP. En otras palabras, la relación entre E_j como función de MP_j es al menos no decreciente, dejando fijos los VTP_j y variando las VTP_j^* con acuerdos de coalición.

[14] Al estar reasignando la cantidad de curules otorgados, estaremos *actualizando* los valores de NP_j y por ende, la conformación, *i.e.*, (3.1).

3. Se actualiza la variable G_j que lleva el control de los partidos que no contribuyen a la VNE o que han estado en exceso, *i.e.*

$$G_j = \begin{cases} 1, & \text{si} \quad E_j > 0, \\ G_j & \text{en otro caso.} \end{cases} \qquad \text{para } j \in \{1, \ldots, k\}.$$

Observando que $G_j \in \{0, 1\}$, la podemos interpretar como:

a) $G_j = 1$ *sii* el partido P_j no forma parte de la VNE, o bien, ha caído en exceso en esta o alguna iteración previa.

b) $G_j = 0$ en otro caso, *i.e.*, *sii* forma parte de la VNE y no ha caído en exceso en esta o alguna iteración previa.

4. Se actualiza el total de escaños adquiridos para los partidos que superaron el límite restando el exceso:

$$RP_j = \text{máx}\{G_j(RP_j - E_j), 0\}, \quad \text{para } j \in \{1, \ldots, k\}.$$

Esta actualización de RP_j ha sido construida de tal manera que el resultado sea el total de escaños corregidos por el límite únicamente de aquellos partidos que participan en RP y cayeron en exceso en esta o alguna iteración anterior, *i.e.*, puede tomar los siguientes valores contemplando estos casos.

a) Se actualiza $RP_j = RP_j - E_j = U_j > 0$ cuando el partido P_j forma parte de la VNE y ha caído en un exceso en esta o alguna iteración anterior, tal que los escaños otorgados por RP fueron mayores al excedente (E_j) dado por su correspondiente límite (U_j). En otras palabras, al partido que se encuentre en este caso se le asignan los escaños límite U_j.

b) Se actualiza $RP_j = 0$, en otro caso, *i.e.*, cuando el partido P_j cae en alguno de los tres siguientes casos: 1) Forma parte de la VNE y no ha caído en exceso en esta o alguna iteración previa, o bien, 2) Si el excedente en su total de escaños es mayor o igual que los curules otorgados por RP, o si simplemente 3) No forma parte de la VNE.

Nota: El caso b)-2), donde $RP_j \leq E_j$, puede darse derivado de los acuerdos de coalición. Recordando y haciendo énfasis en una observación anterior, estos acuerdos de coalición otorgan escaños por el principio de MR y no pueden ser penalizados o limitados por protección constitucional, mientras que los otorgados por RP están cimentados en la VNE y por lo tanto en las votaciones VTP_j y no en las VTP_j^* que son producto de las coaliciones. Este punto nos permite ver una debilidad constitucional que favorece una posible sobrerrepresentatividad de un partido vía la práctica de coaliciones

en el principio de MR.

A los partidos que cayeron en exceso y sus lugares por RP fueron corregidos en este paso, se les conceden de forma definitiva y el resto de los lugares serán repartidos nuevamente de forma análoga a la **repartición inicial**. Esto lo describimos más a detalle en el siguiente paso. En este punto, $\sum_{j=1}^{k} RP_j$ es el total de curules ya otorgados definitivamente por RP para partidos que cayeron en exceso en este o algún paso anterior.

5. Una vez ya asignados los curules para los partidos que cayeron en exceso, será necesario repartir los

$$M = 200 - \sum_{j=1}^{k} RP_j$$

escaños sobrantes, entre los partidos que no estén (o hayan estado) en exceso. Para lograr lo anterior, con fundamento en el artículo 18 de la LEGIPE, se calcula (o actualiza) la *Votación Nacional Efectiva* (VNE_f), restando de la VNE los votos de los partidos que están o estuvieron en exceso, *i.e.*, a los que ya se les concedieron sus curules. Esto lo hacemos calculando:

$$\begin{aligned} VNE_f &= VNE - \sum_{j=1}^{k} G_j(VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)) \\ &= \sum_{j=1}^{k} \left[VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right) - G_j(VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right))\right] \\ &= \sum_{j=1}^{k} (1 - G_j)(VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)) \end{aligned}$$

Donde el producto $G_j \mathbb{1}\left(q_j \geq 3\,\%\right) \in \{0, 1\}$ tomará el valor:

a) $G_j \mathbb{1}\left(q_j \geq 3\,\%\right) = 1$ *sii* el partido P_j ha caído en exceso en esta o alguna iteración previa y forma parte de la VNE. Lo que significaría que ya se le concedieron escaños por RP en el paso anterior.

b) $G_j \mathbb{1}\left(q_j \geq 3\,\%\right) = 0$ en otro caso, *i.e.*, *sii* no ha caído en exceso en esta o alguna iteración previa, o bien, no forma parte de la VNE. Lo que significaría que aún no se le conceden, o se le concederán escaños por RP.

Esto significa que si un partido ya estuvo en exceso en este o algún paso anterior, será excluido[15] de esta nueva repartición (pues ya se le otorgaron curules

[15]Gracias a las definiciones que estamos usando y al algoritmo que se está trabajando, podemos seguir trabajando con k partidos, sin necesidad de sacar algún partido en algún paso, como ya se dijo al principio de esta sección, este es uno de los objetivos que se buscan para facilitar la implementación de este principio.

por RP). Se define (o actualiza) ahora un *Nuevo Cociente Natural* (NCN), fundamentado en el último artículo mencionado, como:

$$NCN = \frac{VNE_f}{M}$$

Este *Nuevo Cociente Natural* tiene una interpretación análoga al *Cociente Natural* de la repartición inicial, solo que ahora no están participando los partidos que entraron a RP, ya cayeron en exceso en esta o alguna iteración anterior y por lo tanto ya se les otorgaron sus escaños por este principio. Posteriormente, el procedimiento continúa de forma similar a la **repartición inicial**. En lugar de los 200 lugares, se reparten ahora M, primero considerando para cada partido que aún no ha caído en exceso, P_j los escaños ganados por su participación en la VNE_f de la siguiente manera:

$$C_j = (1 - G_j)\left\lfloor \frac{VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)}{NCN} \right\rfloor \quad \text{para } j \in \{1, \ldots, k\},$$

actualizando así las C_j que se obtuvieron anteriormente. En el caso de que faltasen por asignar, se calculan los residuos dados por la actualización de R,

$$R = M - \sum_{j=1}^{k} C_j \in \mathbb{N}.$$

Este valor también se encontrará acotado por:

$$R = \frac{VNE_f}{NCN} - \sum_{j=1}^{k} C_j = \sum_{j=1}^{k} (1 - G_j)\left[\frac{(VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right))}{NCN} - C_j\right] < \sum_{j=1}^{k} (1 - G_j)$$

Donde $\sum_{j=1}^{k}(1 - G_j)$ es la cantidad de partidos que aún no caen en exceso en esta o alguna iteración anterior y que de hecho están participando en la VNE_f. Más aún, $\sum_{j=1}^{k}(1 - G_j) < k$ pues estamos en el caso donde ya al menos un partido que cayó en exceso en esta o una iteración anterior.

Ahora, como parte la metodología del *resto mayor*[16], se actualizan los residuos R_j fragmentando R para cada partido P_j como:

$$R_j = \frac{VTP_j \mathbb{1}\left(q_j \geq 3\,\%\right)(1 - G_j)}{NCN} - C_j$$

utilizando de esta metodología con: 1) NCN en lugar del CN, 2) $(1-G_j)VTP_j$ en lugar de VTP_j, y, 3) Las C_j y R_j actualizadas. Obteniendo así los $M_j \in \{0, 1\}$

[16]La metodología de *resto mayor* se vió en la Subsubsección 3.2.4.1 (**repartición inicial**).

repartidos del residuo R.

Nota: $R < \sum_{j=1}^{k}(1 - G_j)$ implica que en este caso nunca será necesario otorgar más de un escaño a un mismo partido por *resto mayor*. Por construcción, $\sum_{j=1}^{k} M_j = R$.

Por lo tanto, el número total de diputados obtenidos por el principio de RP hasta este punto para cada partido (RP_j) se obtiene actualizando con los nuevos valores de C_j y M_j como:

$$RP_j = C_j + M_j \quad \text{para } j \in \{1, \ldots, k\}.$$

Posteriormente, actualizamos el número total de curules otorgados que hasta este punto tiene cada partido P_j, lo cual está dado por la suma de los escaños obtenidos vía los principios de MR y RP, *i.e.*,

$$NP_j = MP_j + RP_j \quad \text{para } j \in \{1, \ldots, k\}.$$

Actualizando así la Conformación de la Cámara de Diputados (3.1).

6. Con fundamento en el artículo 17-2 de la LEGIPE se calcula el exceso de acuerdo al paso 2 de este algoritmo, donde hay finalmente 2 casos. 1) Si hay algún partido que cae en exceso, *i.e.*, $\exists P_j$ tal que $E_j > 0$, entonces se continúa con los pasos subsecuentes. 2) Si ningún partido cae en exceso, *i.e.*, $\forall P_j$ sucede que $E_j = 0$ entonces el algoritmo ha terminado y finalmente tenemos la conformación final en (Ecuación 3.1) con la cantidad de votos con los que se cuenta.

   Nota: El algoritmo debe terminar en algún momento pues en caso de que algún partido P_j esté en exceso, entonces tal y como se menciona en el paso 4-a), se actualiza $NP_j = U_j$, lo cual implica que $E_j = 0$. En otro caso, si P_j no está en exceso, entonces $\mathbb{1}(NP_j > U_j) = 0$ y nuevamente $E_j = 0$. Por lo tanto, en algún punto deberá existir una repartición tal que los escaños otorgados para cada partido no violen la cota de no sobrerrepresentación. De esta manera tendremos una eventual conformación de la Cámara de Diputados que satisfaga lo que está sujeto tanto en la CPEUM y la LEGIPE.

3.2.5. Resultados a presentar

El resultado final que se espera de estos cálculos es presentar:

1. El vector "Conformación" de escaños otorgados a cada entidad política mostrado en (3.1) (incluye a los candidatos independientes) que ya cumple con los requisitos de no sobrerrepresentación. Con el objetivo de conocer la composición de la cámara de diputados dados los votos presentados, y,

2. El vector $pVVE$ que lo definiremos a partir de lo visto en la Subsubsección 3.2.2.3 como

$$pVVE = \{q_1, \ldots, q_k, q_I\}, \tag{3.8}$$

Donde $q_I = 100 \times \left(\frac{\sum_{s=1}^{r} VTI_s}{VVE}\right)$ %. Esto tiene el objetivo de observar cuáles partidos mantendrán su registro para las siguientes elecciones. Además, se incluye el agregado de todos los candidatos independientes para lograr así un 100 % de la VVE.

Nota: Estos dos vectores deben tener la misma longitud, $k + 1$.

Además, se espera presentar el porcentaje de participación ciudadana, el cual se verá en la Ecuación 5.3 del Capítulo 5.

3.2.6. Afiliación efectiva para 2021

Con base en el artículo [7] publicado por el *Tribunal Electoral del Poder Judicial de la Federación* (TEPJF), el objetivo de la afiliación efectiva es evitar que, mediante un convenio de coalición, se distorsionen los límites de sobrerrepresentación en la Cámara de Diputados. Es un instrumento que busca garantizar el cumplimiento de los límites de sobrerrepresentación, los cuales ya se mencionaron en la subsección anterior. Esto, porque constituye un objetivo apegado a la CPEUM y a las leyes generales, y es tal que permite determinar el origen de una candidatura postulada por una coalición y que ganó en un distrito electoral por MR.

Históricamente, en 2015 se advirtió por primera vez que los partidos políticos postulaban militantes de otras entidades políticas con los cuales estaban coaligados. En ese año, determinadas fuerzas políticas impugnaron la asignación de curules vía el principio de RP porque había partidos sobrerrepresentados, debido precisamente a que sus militantes fueron postulados por otros partidos; es decir, algunos de los candidatos postulados eran partícipes en nombre de partidos políticos coaligados con su original pero con los que no se habían visto envueltos directamente hasta instancias cercanas a las elecciones. Sin embargo, la Sala Superior del TEPJF confirmó la asignación[17] y, en su momento, emitió una jurisprudencia[18] en donde se establecía:

> "*Candidatos a cargos de elección popular. Pueden ser postulados por un partido político diverso al que se encuentran afiliados, cuando exista convenio de coalición.*"

[17] SUP-REC-943/2018 y acumulados.
[18] Jurisprudencia 29/2015, derivada de la contradicción de criterios SUP-CDC-8/2015.

validando esa posibilidad.

Durante el 2018, se presentó nuevamente el problema. La controversia llegó a la Sala Superior del TEPJF con motivo de la asignación de curules vía RP hecha por el INE. En la sentencia[19], la Sala Superior del TEPJF confirmó la asignación, porque los actores pretendieron cuestionarla una vez emitidos los resultados de la elección, por lo que se pretendía modificar los candidatos ya sufragados. Sin embargo, en esa misma sentencia se consideró la necesidad de revisar la jurisprudencia que permite a los partidos políticos coaligados postular candidaturas de militantes que no están en sus filas, debido a que el criterio trasciende en el equilibrio de la Cámara de Diputados.

Es indispensable precisar que, en ambos precedentes, el problema fue planteado a la Sala Superior del TEPJF en un momento en el cual ya había transcurrido la jornada electoral, la ciudadanía ya había ejercido su derecho a votar y emitido su preferencia por cierta fuerza política, a partir de la manera en que los partidos y coaliciones contendieron en la elección. Por ello, aunque un partido político formalmente se ajustaba a los límites de sobrerrepresentación, materialmente en los hechos y en la integración real de la Cámara de Diputados, tenían una representación mayor por virtud de su convenio de coalición en comparación a su votación recibida; es decir, estaban sobrerrepresentados. La técnica que los partidos políticos empleaban para que esto sucediera, nacía de la matemática que involucraba a la VNE y el límite de sobrerrepresentación derivada de esta en el principio de RP, algo se menciona a lo largo de la Subsubsección 3.2.4.2 de este documento.

En respuesta a estos acontecimientos, en el 2021 y previo a las elecciones, el Consejo General del INE emitió un acuerdo[20] para establecer reglas al momento de asignar diputaciones de RP. Dentro de estas reglas, se estableció la figura de la "*afiliación efectiva*", es decir, aquella militancia real de la candidatura postulada por una coalición, al momento de ser registrada por el INE. En otras palabras, si en el convenio se señala que una persona es militante de un partido, se deberá verificar que realmente lo sea. Para ello, el INE estableció tres criterios para determinar cuál es la *afiliación efectiva* de las candidaturas postuladas por las coaliciones. Esos criterios son:

1. El INE verificará cuál era la *afiliación efectiva* al momento del registro.
2. Si el ganador no tiene *afiliación efectiva*, la diputación contará para el partido político que se haya establecido en el convenio de coalición.
3. Si el ganador fue por reelección y sin *afiliación efectiva*, el triunfo se contará para el grupo parlamentario al que haya pertenecido al momento del registro.

Mediante esos tres criterios, el INE buscó definir a qué partido político se le debe contar el triunfo de una candidatura postulada por una coalición, lo cual permite

[19] SUP-REC-943/2018 y acumulados.
[20] INE/CG193/2021

también determinar los límites de sobrerrepresentación. De igual forma, el acuerdo del INE resuelve, en el momento previo a los resultados, el problema presentado desde las elecciones federales de diputaciones de los años 2015 y 2018 porque los partidos políticos podrán postular candidaturas por conducto de otro partido con el cual estén coaligados, pero el triunfo se contará para el partido político en el cual realmente milita la candidatura.

Por otro lado, la Sala Superior del TEPJF tuvo que decidir sobre la constitucionalidad y legalidad del acuerdo, con base en los siguientes aspectos mencionados en [7]:

(a) La afiliación efectiva es un instrumento para garantizar el cumplimiento de los límites de sobrerrepresentación. Esto, porque constituye un parámetro objetivo y apegado a la Constitución y a las leyes generales, que permite determinar el origen de una candidatura postulada por una coalición y que ganó en un distrito electoral por MR.

(b) El acuerdo del INE se ajusta al artículo 105 constitucional[21], porque si bien fue emitido durante el proceso electoral, lo cierto es que fue anterior a la etapa de registros de candidaturas. Además, el acuerdo sólo fija un criterio para asignar diputaciones, sin modificar, alterar, afectar los derechos de las candidaturas, de los partidos políticos y de las coaliciones. Finalmente, se trata de una norma instrumental para que el INE asigne las diputaciones, con base en el procedimiento constitucional y legal establecido.

(c) No se vulnera la reserva de ley, porque el INE es el encargado de la función electoral, con facultades constitucionales y legales para verificar, entre otros aspectos, el cumplimiento a los límites de sobrerrepresentación.

(d) No se vulnera la autodeterminación de los partidos, porque no modifica las estrategias que planearon a fin de contender de manera coaligada ni la posibilidad de postular militantes de otros partidos políticos con los que estén coaligados. Únicamente es para fines de la asignación y a qué partido se le contará el triunfo.

(e) No se vulnera la jurisprudencia 29/2015 de la Sala Superior del TEPJF, porque los partidos políticos podrán postular militantes de otros partidos con los que estén coaligados. Es decir, la jurisprudencia no resultaba aplicable, porque ésta regula una manera en cómo pueden participar los partidos políticos en la postulación de candidaturas cuando están coaligados. En cambio, el acuerdo impugnado trata sobre una etapa distinta, en la cual actúa el INE al momento de asignar diputaciones de RP, es decir, cuando ya hay resultados electorales.

Esto finalmente logró convalidar el carácter constitucional y legal del acuerdo impugnado, por ser un mecanismo para garantizar la adecuada representatividad de los

[21] En el artículo 105 de la CPEUM se mencionan los asuntos que, en términos que señale la ley reglamentaria, deberán ser del conocimiento de la *Suprema Corte de Justicia de la Nación* (SCJN).

partidos políticos en la Cámara de Diputados, a partir de la votación realmente obtenida. Esto constituye resoluciones históricas, fundamentales y trascendentes para el cumplimiento de la Constitución y, de manera particular, para garantizar la debida integración de un órgano del Estado mexicano: la Cámara de Diputados, en la cual se respeten los límites de sobrerrepresentación.

En conclusión, esta sentencia pretende que las curules obtenidas por los partidos políticos representen de forma real y auténtica el número de votos obtenidos en la elección, en los términos constitucionalmente indicados. Es decir, trata de impedir que medidas artificiosas afecten o evadan la prohibición constitucional de la sobrerrepresentación. Por lo que, se resuelve el problema presentado en elecciones pasadas, en las cuales la controversia fue planteada de forma tardía a la Sala Superior del TEPJF. Teniendo de esta manera, una mejor relación entre votos y escaños, que es lo que toda democracia pretende.

3.3. Ejemplos poblacionales (2015 y 2018)

Para realizar los Cálculos de la Conformación de la Cámara de Diputados vistos en la Sección 3.2 es necesario contar con el agregado de votos por entidad política, incluyendo coaliciones, candidatos independientes, no registrados y votos nulos; esta es información que será presentada por el INE a nivel casillas de los distritos federales electorales. Recordemos que, en la Sección 2.1 se menciona el proceso que involucra la llegada de información, la cual es obtenida de una muestra representativa de la votación a nivel nacional.

En esta sección, se describirá a medida de ejemplo y utilizando información poblacional, es decir con la que se determinará la conformación final de la cámara salvo ajustes que considere el INE, para los años 2015 y 2018, los cálculos necesarios para obtener la conformación de la Cámara de Diputados, donde gracias al trabajo de recolección de los CAEs llegará la llamada *Base de datos de los Cómputos Distritales* que es el insumo requerido para realizar los cálculos.

Todos los cálculos, así como los insumos necesarios para obtener los resultados que se mostrarán en esta sección, están escritos en código de R. Todos estos se podrán localizar en el GitHub del autor. En particular, los correspondientes a esta sección se pueden encontrar dando clic aquí[22].

3.3.1. Base de datos de los Cómputos Distritales

La *Base de datos de los Cómputos Distritales* es una tabla de datos que contiene por renglones información de las casillas en los distritos federales electorales y por

[22] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/3.%20La%20C%C3%A1mara%20de%20Diputados

columnas los conteos agregados de votos para los partidos políticos, coaliciones, candidatos independientes, candidatos no registrados, votos nulos y otros identificadores de cada casilla. Suele variar de elección en elección debido a la naturaleza del cambio en las fuerzas y entidades políticas que sean partícipes, aunque en general tiene que tener las cantidades e identificadores anteriores. Asimismo, la cantidad de casillas así como sus identificadores pueden también ser sujetos a cambios en cada elección por requerimientos del COTECORA o del INE en cada año de elecciones.

3.3.1.1. Variables de la Base de Datos

En este apartado mostraremos las variables que fueron necesarias para hacer el cálculo de la conformación de la cámara de diputados para los años 2015 (Tabla 3.1) y 2018 (Tabla 3.2).

- Para el 2015, se contó con 149,075 casillas y las variables que se utilizan para el cálculo de la conformación se pueden observar en la Tabla 3.1. En estas elecciones, solo se contó con dos coaliciones de dos partidos cada una, de tal manera que la metodología del reparto de votos dados directamente a la coalición está descrita ya explícitamente en la Subsubsección 3.2.2.1. Cabe mencionar que dicha repartición se hace por coalición y sobre los partidos partícipes en la misma. Otra observación interesante es que en estas elecciones existen partidos que no forman parte de alguna coalición, de hecho este es el caso de la mayoría de los partidos.

- Para el 2018 se contó con 157,640 casillas y las variables que se utilizan para el cálculo de la conformación se pueden observar en la Tabla 3.2. En estas elecciones, se contó con tres coaliciones de tres partidos cada una, de igual manera, la metodología del reparto de votos dados directamente a la coalición está descrita también explícitamente en la Subsubsección 3.2.2.1. En contraste a las elecciones del 2015, para las del 2018 no existen partidos que no forman parte de alguna coalición.

3.3.2. Valores poblacionales: Votación Válida Emitida y Conformación

En esta sección se describirá de forma breve los cálculos involucrados para obtener *Votación Válida Emitida* (VVE) y la conformación de la Cámara de Diputados a partir de la *Base de datos de los Cómputos Distritales* ejemplificando de esta manera la Sección 3.2. Estos resultados están validados por el portal de Transparencia y Protección de Datos Personales del INE en el acuerdo INE/CG804/2015 para el 2015 y por el portal del Repositorio Documental del INE en el acuerdo INE/CG1181/2018 para el 2018.

Función	Variable	Descripción
Identificadores de Casilla	ESTADO	Identificador de estado
	DISTRITO	Identificador de distrito
Votos otorgados directamente a Partidos Políticos	PAN	Partido Acción Nacional
	PRI	Partido Revolucionario Institucional
	PRD	Partido de la Revolución Democrática
	PVEM	Partido Verde Ecologista de México
	PT	Partido del Trabajo
	MOVIMIENTO_CIUDADANO	Movimiento Ciudadano
	NUEVA_ALIANZA	Nueva Alianza
	MORENA	Movimiento Regeneración Nacional
	PH	Partido Humanista
	PS	Partido Socialista
Votos otorgados directamente a Coaliciones	C_PRI_PVEM	PRI - PVEM
	C_PRD_PT	PRD - PT
Votos otorgados directamente a Candidatos Independientes	CAND_IND_1	Candidato independiente i
	CAND_IND_2	
Votos realizados ajenos a la VVE	NO_REGISTRADOS	Candidatos No Registrados
	NULOS	Votos Nulos

Tabla 3.1: Variables de la Base de datos de los Cómputos Distritales para 2015

- Para el 2015, se comienza calculando la VTE de la *Base de datos de los Cómputos Distritales* agrupando las casillas con base en las variables ESTADO y DISTRITO en una variable llamada arbitrariamente ID_EDO_DIST de tal manera que se obtendrán las $VTE_{i,\cdot}$ que son los agregados de votos para cada $i \in \{1, \ldots, 300\}$ de los distritos federales electorales como en la Subsubsección 3.2.2.2 utilizando las variables correspondientes a los votos de la Tabla 3.1.

  Posteriormente se tiene el objetivo de encontrar al vector de escaños que obtiene cada fuerza política por la vía de MR (MP) en (3.5) tal y como se indica en

la Subsección 3.2.3. Para esto, es necesario conocer los *convenios de coalición* (véase la Subsubsección 3.2.2.1) los cuales se presentan al COTECORA en forma de una tabla con variables indicadoras[23] de tamaño $300 \times k$, considerando que se tuvieron $k = 10$ partidos políticos participando en las elecciones del 2015, véase la Tabla 3.3.

Una vez obtenido el vector MP, se suman sobre los 300 distritos federales electorales los elementos de $VTE_{i,\cdot}$ como en la Subsubsección 3.2.2.2 para obtener el vector VTE_d. De tal manera que los resultados en 2015 de los vectores MP y VTE_d los podemos ver en la Tabla 3.4. A partir de este punto ya tenemos parte de la conformación de la cámara de diputados, que son los 300 escaños otorgados por MR. Se procede entonces con la otra parte de la conformación que son aquellos 200 curules dados por el principio de RP.

Procediendo con el cálculo de la conformación por parte del principio de RP, se debe ahora obtener la *Votación Nacional Emitida* (VNE) simplemente extrayendo de la VTE los votos nulos y a candidatos independientes y no registrados (Subsubsección 3.2.2.4). Esto se hace de tal manera que obtengamos el valor de VNE_d en (3.3) y el vector VNE de (3.4) de la Tabla 3.4, basta con eliminar los tres últimos renglones, quedándonos así únicamente con los Partidos Políticos, que son precisamente los que podrán participar en el principio de RP. Una vez teniendo estos datos se realiza una *repartición inicial* como en la Subsubsección 3.2.4.1, los resultados de esta primera repartición se muestran en la Tabla 3.5.[24]

De aquí, se debe verificar que los curules asignados de todos los partidos caen debajo de su cota de no sobrerrepresentación (Subsubsección 3.2.4.2), en particular, podemos notar que los curules otorgados al partido `PRI` son 220, sin embargo, la cota de no sobrerrepresentación está dada por (Ecuación 3.7) es de 203, por lo tanto esta asignación de curules por RP no satisface la condición de no sobrerrepresentación. Entonces, continuando con el algoritmo de no sobrerrepresentación, la cota será asignada a este partido político, y con los escaños restantes se volverán a repartir entre los demás partidos y repitiendo el proceso hasta que la condición se cumpla para todos los partidos.

Este procedimiento se repite cuantas veces sea necesaria hasta que todos los partidos que participan en RP satisfagan la cota, en este caso, el resultado y

[23]Estas variables indicadoras mostraban 1 si el partido fue el seleccionado por la coalición para dicho distrito y 0 en otro caso. Si ambos partidos presentaban 0, significa que no había un acuerdo de coalición en el distrito en cuestión y por lo tanto cada partido se quedaba con su correspondiente parte de la VTE, véanse los casos 1-3 en la Subsección 3.2.3.

[24]Nótese que no todos los partidos recibieron escaños por el principio de RP. Esto se debe a que, así como se toca en la Subsección 3.2.4, los partidos con un $q_j < 3\,\%$ no recibirán curules bajo este principio.

objetivo final de los cálculos para la conformación (ver la Subsección 3.2.5) se presentan en la Tabla 3.6.

Con lo que quedan finalmente mostrados los escaños que les corresponden a cada una de las fuerzas políticas participantes y aquellos partidos políticos que perderán su registro en las elecciones del 2015.

- En 2018, el procedimiento es completamente análogo, se comienza calculando la VTE de la *Base de datos de los Cómputos Distritales* agrupando las casillas con base en las variables `ID_ESTADO` y `ID_DISTRITO` en una variable llamada arbitrariamente `ID_EDO_DIST` de tal manera que se obtendrán las $VTE_{i,\cdot}$ que son los agregados de votos para cada $i \in \{1, \ldots, 300\}$ de los distritos federales electorales como en la Subsubsección 3.2.2.2 utilizando las variables correspondientes a los votos de la Tabla 3.2.

  Posteriormente se tiene el objetivo de encontrar al vector de escaños que obtiene cada fuerza política por la vía de MR (MP) en (3.5) tal y como se indica en la Subsección 3.2.3. Para esto, es necesario conocer los *convenios de coalición* (véase Subsubsección 3.2.2.1) los cuales se presentan al COTECORA en forma de una tabla con variables indicadoras[25] de tamaño $300 \times k$, considerando que se tuvieron $k = 9$ partidos políticos participando en las elecciones del 2018, véase la Tabla 3.7.

  Una vez obtenido el vector MP, se suman sobre los 300 distritos federales electorales los elementos de $VTE_{i,\cdot}$ como en la Subsubsección 3.2.2.2 para obtener el vector VTE_d. De tal manera que los resultados en 2018 de los vectores MP y VTE_d los podemos ver en la Tabla 3.8. A partir de este punto ya se cuenta con parte de la conformación de la cámara de diputados, que son los 300 escaños otorgados por MR. Se procede entonces con la otra parte de la conformación que son aquellos 200 curules dados por el principio de RP.

  Procediendo con el cálculo de la conformación por parte del principio de RP, se debe ahora obtener la *Votación Nacional Emitida* (VNE) simplemente extrayendo de la VTE los votos nulos y a candidatos independientes y no registrados (Subsubsección 3.2.2.4). Esto se hace de tal manera que obtengamos el valor de VNE_d en (3.3) y el vector VNE de (3.4) de la Tabla 3.8, basta con eliminar los tres últimos renglones, quedándonos así únicamente con los Partidos Políticos, que son precisamente los que podrán participar en el principio de RP. Una vez teniendo estos datos se realiza una *repartición inicial* como en la Subsubsección 3.2.4.1, los resultados de esta primera repartición se muestran en la Tabla 3.9.

[25] Estas indicadoras tienen la misma interpretación que 2015, véanse los casos 1-3 en la Subsección 3.2.3.

Haciendo algunas observaciones a la Tabla 3.9; en particular, nótese la cantidad de escaños otorgados al partido ES en contraste de su $pVVE$, de hecho, este partido recibió muchos más curules que el partido PANAL aún cuando este último obtuvo un $pVVE$ mayor, contemplando que, ambos partidos perdieron su registro y no están participando por escaños otorgados bajo el principio de RP. Este efecto es debido precisamente a los acuerdos de coalición, mostrándonos así un ejemplo claro de sobrerrepresentación por este medio. Recordemos, de la Tabla 3.2 que el partido ES se encuentra coaligado con los partidos PT y MORENA en la así llamada coalición "Juntos haremos historia". En particular obsérvese el dominio del partido MORENA en la VVE. En contraste, el partido PANAL está coaligado con los partidos PRI y PVEM en la coalición "Todos por México", los cuales ni siquiera juntos se acercan al $pVVE$ de MORENA. Como última observación, véase la amplia diferencia que hay entre los curules asignados de MORENA y su cota de no sobrerrepresentación U_j. Esto fue uno de los detonantes más importantes que dio pie a la implementación de la *Afiliación Efectiva* (Subsección 3.2.6).

De aquí, se debe verificar que los curules asignados de todos los partidos caen debajo de su cota de no sobrerrepresentación (Subsubsección 3.2.4.2), en particular, podemos notar que los curules otorgados al partido PT son 67, sin embargo, la cota de no sobrerrepresentación está dada por (3.7) es de 61, por lo tanto esta asignación de curules por RP no satisface la condición de no sobrerrepresentación. Entonces, continuando con el algoritmo de no sobrerrepresentación, la cota será asignada a este partido político, y con los escaños restantes se volverán a repartir entre los demás partidos y repitiendo el proceso hasta que la condición se cumpla para todos los partidos.

Este procedimiento se repite cuantas veces sea necesario hasta que todos los partidos que participan en RP satisfagan la cota, en este caso, el resultado y objetivo final de los cálculos para la conformación (ver la Subsección 3.2.5) se presentan en la Tabla 3.10.

Nota: Generalmente este proceso se termina con dos iteraciones en total.

Con lo que quedan finalmente mostrados los escaños que les corresponden a cada una de las fuerzas políticas participantes y aquellos partidos políticos que perderán su registro en las elecciones del 2018.

Función	Variable	Descripción
Identificadores de Casilla	ID_ESTADO	Identificador de estado
	ID_DISTRITO	Identificador de distrito
Votos otorgados directamente a Partidos Políticos	PAN	Partido Acción Nacional
	PRI	Partido Revolucionario Institucional
	PRD	Partido de la Revolución Democrática
	PVEM	Partido Verde Ecologista de México
	PT	Partido del Trabajo
	MC	Movimiento Ciudadano
	PANAL	Nueva Alianza
	MORENA	Movimiento Regeneración Nacional
	ES	Encuentro Social
Votos otorgados directamente a Coaliciones	PAN_PRD_MC	Por México al frente
	PAN_PRD	
	PAN_MC	
	PRD_MC	
	PRI_PVEM_PANAL	Todos por México
	PRI_PVEM	
	PRI_PANAL	
	PVEM_PANAL	
	PT_MORENA_ES	Juntos haremos historia
	PT_MORENA	
	PT_ES	
	MORENA_ES	
Votos otorgados directamente a Candidatos Independientes	CAND_IND_01	Candidato independiente i
	CAND_IND_02	
Votos realizados ajenos a la VVE	CNR	Candidatos No Registrados
	VN	Votos Nulos

Tabla 3.2: Variables de la Base de datos de los Cómputos Distritales para 2018

Variables de Referencia		PRI-PVEM		PRD-PT	
i	ID_EDO_DIST	PRI	PVEM	PRD	PT
1	0101	0	0	0	0
2	0102	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
98	1208	0	1	1	0
99	1209	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
299	3203	1	0	0	0
300	3204	1	0	0	0

Tabla 3.3: Fragmento de tabla de los convenios de coalición para las elecciones del 2015.

Partido	VTE_d	MP
PAN	8,377,535	56
PRI	11,636,957	155
PRD	4,335,321	28
PVEM	2,757,170	29
PT	1,134,101	6
MOVIMIENTO_CIUDADANO	2,431,063	10
NUEVA_ALIANZA	1,486,626	1
MORENA	3,345,712	14
PH	856,716	0
PS	1,325,032	0
CAND_IND	225,029	1
NO_REGISTRADOS	52,371	0
NULOS	1,900,449	0

Tabla 3.4: *Votación Total Emitida* (VTE) y escaños otorgados por *Mayoría Relativa* (MR) para cada fuerza política en 2015.

Partido	Conformación	U_j	$pVVE$
PAN	103	157	22.10 %
PRI	**220**	**203**	30.70 %
PRD	52	100	11.44 %
PVEM	45	78	7.27 %
PT	6	**No RP**	**2.99 %**
MOVIMIENTO_CIUDADANO	24	74	6.41 %
NUEVA_ALIANZA	9	60	3.92 %
MORENA	33	86	8.83 %
PH	0	**No RP**	**2.26 %**
PS	7	78	3.50 %
CAND_IND	1	No Aplica	0.59 %

Tabla 3.5: Repartición inicial de la conformación de la cámara de diputados (3.1), cota U_j (3.7) y $pVVE$ (3.8) para cada fuerza política en 2015.

Partido	Conformación	$pVVE(\%)$
PAN	109	22.10 %
PRI	203	30.70 %
PRD	55	11.44 %
PVEM	47	7.27 %
PT	6	2.99 %
MOVIMIENTO_CIUDADANO	25	6.41 %
NUEVA_ALIANZA	11	3.92 %
MORENA	35	8.83 %
PH	0	2.26 %
PS	8	3.50 %
CAND_IND	1	0.59 %

Tabla 3.6: Conformación de la cámara de diputados (3.1) por ambos principios MR + RP y $pVVE$ (3.8) para cada fuerza política en 2015.

Variables de Referencia		Por México al frente			Juntos haremos historia			Todos por México		
i	ID_EDO_DIST	PAN	PRD	MC	MORENA	ES	PT	PRI	PANAL	PVEM
1	0101	0	1	0	0	0	1	0	0	0
2	0102	1	0	0	0	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
98	1208	0	1	0	1	0	0	0	0	1
99	1209	0	1	0	0	0	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
299	3203	0	1	0	0	0	1	1	0	0
300	3204	0	0	1	0	0	1	1	0	0

Tabla 3.7: Fragmento de tabla de los convenios de coalición para las elecciones del 2018.

Partido	VTE_d	MP
PAN	10,093,012	40
PRI	9,307,233	7
PRD	2,967,452	9
PVEM	2,694,654	5
PT	2,210,988	58
MC	2,484,185	17
PANAL	1,390,882	2
MORENA	20,968,859	106
ES	1,353,499	56
CAND_IND	538,964	0
CNR	32,938	0
VN	2,241,811	0

Tabla 3.8: *Votación Total Emitida* (VTE) y escaños otorgados por *Mayoría Relativa* (MR) para cada fuerza política en 2018.

Partido	Conformación	U_j	$pVVE$
PAN	80	139	18.69 %
PRI	44	131	17.23 %
PRD	21	69	5.49 %
PVEM	15	66	4.99 %
PT	**67**	**61**	4.09 %
MC	27	64	4.60 %
PANAL	2	**No RP**	**2.58 %**
MORENA	188	246	38.82 %
ES	56	**No RP**	**2.51 %**
CAND_IND	0	No Aplica	1.00 %

Tabla 3.9: Repartición inicial de la conformación de la cámara de diputados (3.1), cota U_j (3.7) y $pVVE$ (3.8) para cada fuerza política en 2018.

Partido	Conformación	$pVVE$
PAN	81	18.69 %
PRI	45	17.23 %
PRD	21	5.49 %
PVEM	16	4.99 %
PT	61	4.09 %
MC	27	4.60 %
PANAL	2	2.58 %
MORENA	191	38.82 %
ES	56	2.51 %
CAND_IND	0	1.00 %

Tabla 3.10: Conformación de la cámara de diputados (3.1) por ambos principios MR + RP y $pVVE$ (3.8) para cada fuerza política en 2018.

Capítulo 4

Muestreo probabilístico

4.1. Teoría básica del muestreo probabilístico

Para este capítulo, se tomará como base lo escrito en [1]. Como introducción, el muestreo probabilístico es un método de selección y estimación en donde cada elemento en la población tiene una probabilidad de selección conocida y distinta de cero. Considerando estas probabilidades, se emplea un mecanismo aleatorio para elegir los elementos que se incluirán en la muestra.

Si el diseño del muestreo probabilístico se implementa bien, un investigador puede usar una muestra relativamente pequeña para hacer inferencias sobre una población arbitrariamente grande. Generalmente, el interés recae en la estimación de totales, promedios y porcentajes poblacionales.

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

1. Se tiene una población de interés con un total de $N < \infty$ elementos, cuyos índices denotaremos como:

   $$U = \{1, 2, 3, \ldots, N\}.$$

   A este conjunto se le denomina *población* o *universo*.

2. Nos interesa conocer alguna característica de la población, a la medición de esta característica en cada elemento de la población, se le denotará como:

   $$\{y_1, y_2, \ldots, y_N\}.$$

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

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

Para la tesis, se mencionará particularmente la teoría de muestreo probabilístico que fue requerida para los resultados a presentar sobre la estimación de la conformación de la cámara de diputados (ver 3.2.5).

4.1.1. Bases del Muestreo Aleatorio Simple (MAS)

Antes de entrar de lleno con el estimador de razón, se deben aclarar ciertos puntos con respecto al *Muestro Aleatorio Simple* (MAS). Si se selecciona aleatoriamente una muestra de $n \leq N$ elementos de la población de manera equiprobable, se le llama MAS. Dicho conjunto con n elementos de U se le denotará como $\mathcal{S}$. Esto proporciona una base para metodologías más complejas. Existen dos maneras de seleccionar una *muestra aleatoria simple*:

- **Con reemplazo:** Cuando cada elemento de la población se selecciona de manera equiprobable $(1/N)$, y siempre tomando algún elemento del universo. Lo que da pie a la posibilidad de que los elementos seleccionados en la muestra $\mathcal{S}$ se encuentren repetidos.

  Nota: Observe que bajo este esquema se pueden tomar muestras tales que $n > N$ ya que existe la posibilidad de que los elementos sean repetidos. Más aún, éste tipo de muestras podrían incluso no incluir todos los elementos del universo.

  De esta observación, resulta intuitivo que $n \to \infty$ implicaría que el universo está contenido en $\mathcal{S}$. Esto se puede inferir mostrando la probabilidad de que un elemento fijo pero arbitrario, i, del universo sea seleccionado al menos una vez en la muestra. Para esto, se observa que

$$\mathbb{P}\left[i \in \mathcal{S}\right] = 1 - \mathbb{P}\left[i \notin \mathcal{S}\right],$$

  donde el evento $\{i \notin \mathcal{S}\}$ significa que i nunca será seleccionado y para eso, es porque se están tomando únicamente los $N-1$ elementos restantes. Además, la probabilidad de que, cada vez que se toma un elemento, éste sea diferente de i es $(N-1)/N = 1 - 1/N$. Como esto debe pasar n-veces que son la cantidad de elementos que se toman, entonces:

$$\mathbb{P}\left[i \notin \mathcal{S}\right] = \left(1 - \frac{1}{N}\right)^n.$$

  Por lo tanto, la probabilidad de que un elemento arbitrario i aparezca, al menos una vez en la muestra seleccionada es:

$$\mathbb{P}\left[i \in \mathcal{S}\right] = 1 - \left(1 - \frac{1}{N}\right)^n,$$

lo cual da fundamento matemático a la idea intuitiva derivada de la nota anterior.

- **Sin reemplazo:** Cuando cada elemento de la población se selecciona en cada extracción de manera equiprobable, con la diferencia de que cada vez que se extrae un elemento éste no podrá ser seleccionado nuevamente. Lo que impide que los elementos seleccionados en $\mathcal{S}$ se encuentren repetidos. Desde Otro punto de vista, se puede pensar que en cada selección, el universo se va reduciendo hasta conformar $\mathcal{S}$ con n elementos.

  Nota: En este otro esquema no se pueden tomar muestras tales que $n > N$ ya el universo tiene $N < \infty$ elementos y éstos no pueden repetirse. Más aún, si se considera $n = N$ entonces la muestra resultante será, como conjunto, exactamente igual al universo.

  De esta nota, resulta intuitivo que a medida que $n \to N$ implicaría que la muestra S se parecerá cada vez más al universo, pues poco a poco todos los elementos estarán en la muestra. Así como en el caso anterior, esto se puede inferir mostrando la probabilidad de que un elemento fijo pero arbitrario, i, del universo sea seleccionado (una vez) en la muestra.

  Para esto, sabemos por combinatoria, que existen $\binom{N}{n}$ posibles maneras de seleccionar n elementos de una población de tamaño N. Además, sabemos que si la i-ésima observación está en la muestra, entonces quedan $\binom{N-1}{n-1}$ posibles maneras de seleccionar a los $n-1$ elementos restantes del total $N-1$. Entonces,

$$\mathbb{P}\left[i \in \mathcal{S}\right] = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}. \tag{4.1}$$

  Dando así, fundamento teórico a la idea intuitiva derivada de la nota anterior.

4.1.1.1. Propiedades básicas

El autor en [1] menciona que a menudo se está interesado en conocer algunos parámetros de la población, tales como el total y la media de los valores de la medida. Otro parámetro de especial interés es la razón entre los valores observados y alguna otra variable auxiliar. En la Tabla 4.1 se describen algunos de los parámetros y sus estimadores más importantes del área estadística. En particular, se identificarán a los parámetros poblacionales con letras mayúsculas y a los estimadores con minúsculas.

	Total	**Media**	**Razón**	**Varianza**
Expresión Real	$Y = \sum_{i=1}^{N} y_i$	$\overline{Y} = \frac{Y}{N}$	$R = \frac{\overline{Y}}{\overline{X}}$	$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2$
Estimador	$y = N\overline{y}$	$\overline{y} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i$	$r = \frac{\overline{y}}{\overline{x}}$	$s^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(y_i - \overline{y}\right)^2$

Tabla 4.1: Estimadores fundamentales en la teoría del muestreo por MAS con sus respectivas expresiones reales.

Donde, en el caso del *Estimador de Razón*, las variables y se piensan como la variable de interés y las x se podría pensar como una variable auxiliar donde lo que se busca es estimar el cociente. A lo largo de este capítulo se estará describiendo más la implementación de este estimador particularmente.

Ahora mostraremos algunas propiedades importantes que tiene el estimador $\overline{y}$. Para esto, recordemos que bajo el MAS sin reemplazo, la probabilidad de seleccionar un elemento arbitrario es $\left(\frac{n}{N}\right)$ véase (4.1) y denotemos como:

$$Z_i = \mathbb{1}\left(i \in \mathcal{S}\right) \sim Bernoulli\left(\frac{n}{N}\right). \tag{4.2}$$

En estadística clásica, un estimador $\hat{\theta}$ del parámetro θ se le llama *insesgado*, si su valor esperado es el parámetro de interés, es decir, θ. Esto es,

$$\mathbb{E}\left[\hat{\theta}\right] = \theta.$$

A la diferencia $\mathbb{E}\left[\hat{\theta}\right] - \theta$ se le llama *sesgo* del estimador y usualmente es denotado como:

$$b(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta.$$

Ahora, al calcular el valor esperado del estimador $\overline{y}$ se observa que:

$$\mathbb{E}\left[\overline{y}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i \in \mathcal{S}} y_i\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right] = \frac{1}{n}\sum_{i=1}^{N} \mathbb{E}\left[Z_i\right] y_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{Y}. \tag{4.3}$$

Lo cual significa que $\overline{y}$ es un estimador insesgado de $\overline{Y}$ (4.3). En consecuencia, sucede que:

$$\mathbb{E}\left[y\right] = \mathbb{E}\left[N\overline{y}\right] = N\overline{Y} = Y, \tag{4.4}$$

lo que significa que y también es un estimador insesgado de Y. Se definen en la Tabla 4.2, las varianzas para los estimadores y y $\overline{y}$, así como sus respectivas estimaciones. Donde el factor $\left(1 - \frac{n}{N}\right)$ se denomina *corrección por población finita* (CPF). Concorde [8], este factor permite que a medida que $n \uparrow N$ la varianza se reduzca, pues se

Estimador	Varianza Real	Varianza Estimada
Media	$V(\overline{y}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$	$v(\overline{y}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$
Total	$V(y) = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$	$v(y) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$
Razón	$V(r) = \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n\overline{X}^2}$	$v(r) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n\overline{x}}$

Tabla 4.2: Estimadores con sus expresiones reales para la varianza de los estimadores de la suma y media poblacional. Para el caso del estimador de razón, ver (4.8) y (4.9).

tiene más información sobre la población, mientras que $n \downarrow 0$ provoca un aumento en la varianza del estimador.

De acuerdo con [9], en la práctica el CPF puede ser ignorado cuando la fracción de muestreo no excede del 5 %, y, para muchos propósitos, incluso si es tan alto como el 10 %. El efecto de ignorar la corrección es sobreestimar el error estándar de la estimación $\overline{y}$.

Para mostrar la fórmula para la varianza $V(\overline{y})$ vista en la Tabla 4.2, realizaremos algunos cálculos primero. Si $i \neq j$ entonces, al estar considerando un *Muestro Aleatorio Simple Sin Remplazo* (MASSR)

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

De tal manera que, recordando (4.2),

$$\begin{aligned} Cov(Z_i, Z_j) &= \mathbb{E}\left[Z_i Z_j\right] - \mathbb{E}\left[Z_i\right]\mathbb{E}\left[Z_j\right] \\ &= \left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right) - \left(\frac{n}{N}\right)^2 \\ &= -\frac{n}{N(N-1)}\left(1 - \frac{n}{N}\right). \end{aligned} \tag{4.5}$$

Con lo cual, podemos ahora mostrar la fórmula para la varianza de $\overline{y}$,

$$
\begin{aligned}
V(\overline{y}) &= \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_i\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j Cov(Z_i, Z_j)\right] \\
&\overset{(i)}{=} \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j\right] \\
&\overset{(ii)}{=} \frac{1}{n}\left(1-\frac{1}{n}\right)\frac{1}{N(N-1)}\left[(N-1)\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2 + \sum_{i=1}^{N} y_i^2\right].
\end{aligned}
$$

Donde (i) sucede por (4.2) y (4.5) factorizando el término común. La igualdad (ii) sucede al simplificar $\sum_{i=1}^{N}\sum_{i\neq j}^{N} y_i y_j = \left(\sum_{i=1}^{N} y_i\right)^2 - \sum_{i=1}^{N} y_i^2$. De esta manera,

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

Este resultado implica que la varianza de y se puede deducir utilizando la Tabla 4.1 como:

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

Un resultado igual de importante es que estos estimadores resultan ser *insesgados*. Para esto, vemos que s^2 de la Tabla 4.1 se puede reescribir como:

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

De tal manera que se facilita el cálculo del valor esperado para el estimador, pues,

$$\mathbb{E}\left[\sum_{i\in\mathcal{S}}\left(y_i-\overline{Y}\right)^2\right]=\mathbb{E}\left[\sum_{i=1}^{N}Z_i\left(y_i-\overline{Y}\right)^2\right]\overset{(4.2)}{=}\frac{n}{N}\sum_{i=1}^{N}\left(y_i-\overline{Y}\right)^2=\frac{n(N-1)}{N}S^2,$$

y también,

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

De tal manera que

$$\begin{aligned}\mathbb{E}\left[s^2\right]&=\frac{1}{n-1}\left[\mathbb{E}\left[\sum_{i\in\mathcal{S}}(y_i-\overline{Y})^2\right]-\mathbb{E}\left[n\left(\overline{y}-\overline{Y}\right)^2\right]\right]\\&=\frac{1}{n-1}S^2\left[\frac{n(N-1)}{N}-\frac{N-n}{N}\right]\\&=\frac{1}{n-1}S^2\left[\frac{N(n-1)}{N}\right]=S^2.\end{aligned}\tag{4.6}$$

Y así, de las definiciones de la Tabla 4.2, tendremos que por linealidad de la esperanza y aplicando (4.6), los estimadores $v(y)$ y $v(\overline{y})$ son insesgados.

4.1.2. Estimador de razón

Se define al parámetro de razón como un cociente entre el valor medio de la variable de interés (y) y la media de otra variable auxiliar (x). Se propone entonces un estimador de R como en la Tabla 4.1 a través de una estadística r la cual denotaremos a continuación y observaremos algunas de sus propiedades. Al ser la razón una función no lineal:

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

El primer objetivo a realizar con respecto la estadística r es el cálculo de su varianza, es decir, buscamos un estimador para $V(r)=\mathbb{E}\left[(r-R)^2\right]$. Para esto, una demostración realizada por [1], consiste en linealizar a r y realizar una aproximación vía series de Taylor de primer orden alrededor del punto $\left(\overline{Y},\overline{X}\right)$ tal y como se muestra a continuación.

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

que, al calcular las derivadas parciales y al evaluar el punto indicado en f, se llega a que:

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

Llegando así a la siguiente expresión de aproximación para r:

$$r \approx R + \frac{1}{\overline{X}}(\overline{y} - R\overline{x}) \quad \Leftrightarrow \quad r - R \approx \frac{1}{\overline{X}}(\overline{y} - R\overline{x}).$$

Donde R y $\overline{X}$ son constantes que asumimos conocidas cuando se conoce toda la población y a las variables auxiliares. Así, definimos $d_i := y_i - Rx_i$, donde $\overline{d}$ será estimador de $\overline{D}$ definida análogamente, luego,

$$\overline{d} := \frac{1}{n}\sum_{i\in\mathcal{S}} d_i = \frac{1}{n}\sum_{i\in\mathcal{S}}(y_i - Rx_i) = \overline{y} - R\overline{x}.$$

De esta manera, tendremos que:

$$\overline{D} = \mathbb{E}\left[\,\overline{d}\,\right] = \mathbb{E}\left[\overline{y} - R\overline{x}\right] = \overline{Y} - R\overline{X} = 0.$$

Lo cual muestra en consecuencia que r es un estimador *aproximadamente* insesgado para R, *i.e.*,

$$\mathbb{E}[r] = \mathbb{E}\left[\frac{\overline{y}}{\overline{x}}\right] \approx R = \frac{\overline{Y}}{\overline{X}} \tag{4.7}$$

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

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

Luego, de los resultados vistos en la Tabla 4.2 aplicados a $\overline{d}$, se deduce que la varianza aproximada es:

$$V(r) \approx \left(1 - \frac{n}{N}\right)\frac{S_d^2}{n\overline{X}^2}, \tag{4.8}$$

donde $S_d^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(d_i - \overline{D})^2$ como en la Tabla 4.1. Por otra parte, un estimador insesgado para $V(r)$ es entonces:

$$v(r) = \left(1 - \frac{n}{N}\right)\frac{s_d^2}{n\overline{x}} \qquad \text{con} \qquad s_d^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\hat{d}_i - \hat{\overline{d}}\right)^2, \tag{4.9}$$

donde $\overline{x}$ es la media muestral de la variable auxiliar; y, $\hat{\overline{d}}$ es la media de la variable $\hat{d}_i = y_i - rx_i$. Análogamente a lo visto en la Subsubsección 4.1.1.1, $v(r)$ es insesgada ya que $\mathbb{E}[s_d^2] = S_d^2$. Como último comentario, de acuerdo con [1], r es un estimador *sesgado* de la razón poblacional, R, pues tiene un sesgo de orden $1/n$ debido a la linealización por series de Taylor. En la práctica, [9] menciona que este valor no es importante en muestras de "tamaño moderado".

4.1.3. Cálculo del Tamaño de Muestra

4.1.3.1. Aproximación asintótica de la distribución de los estimadores

Para todo buen diseño de muestra, es de interés encontrar un tamaño de muestra n suficientemente grande como para que alguno de los estimadores y o bien $\overline{y}$ se acerque estadísticamente lo más posible al valor real. Esto tiene una relevancia importante ya que dependiendo del tamaño de la muestra se pueden deducir diversos costos operativos que finalmente se traducen en pérdidas monetarias que deben ser contempladas por los encargados de recaudar la información.

El cálculo del tamaño de muestra está cimentado en resultados asintóticos de probabilidad, en particular en el *Teorema del Límite Central* (TLC) para *Muestro Aleatorio Simple Sin Remplazo* (MASSR), probado por Hájek en 1960. En términos prácticos, este resultado dicta que bajo ciertas condiciones, y si n, N y $N-n$ son todos *suficientemente grandes*, entonces:

$$Z = \frac{\overline{y} - \overline{Y}}{\sqrt{V(\overline{y})}} \sim N(0,1). \tag{4.10}$$

Que significa que Z tiene un comportamiento asintóticamente (convergencia en distribución) Normal (Gaussiano) Estándar. Asumiendo (4.10) como cierto y tomando de forma general a z_p con $p \in [0,1]$ como el percentil del $p \times 100\,\%$ de una Normal Estándar, se tendrá que:

$$\mathbb{P}\left[z_{\alpha/2} < Z = \frac{\overline{y} - \overline{Y}}{\sqrt{V(\overline{y})}} < z_{1-\alpha/2}\right] = 1 - \alpha.$$

Esto, pensando a $\overline{y}$ como una variable aleatoria, pues su valor depende de la muestra seleccionada por MASSR de tal manera que pensar esto como una probabilidad es válido. Sin embargo, al tener valores observados de la muestra y obtener un valor fijo (estimado), se tendrá que usar un argumento del tipo frecuentista. De tal manera que se puede construir un intervalo de confianza del $(1-\alpha) \times 100\,\%$ para la media $(\overline{y})$, el total (y) y la razón (r) tal y como se muestra en la Tabla 4.3.

Estimador	Intervalo de confianza
Media	$\overline{y} + z_{\alpha/2}\sqrt{V(\overline{y})}, \overline{y} + z_{1-\alpha/2}\sqrt{V(\overline{y})}$
Total	$y + z_{\alpha/2}N\sqrt{V(\overline{y})}, y + z_{1-\alpha/2}N\sqrt{V(\overline{y})}$
Razón	$r + z_{\alpha/2}\sqrt{V(r)}, r + z_{1-\alpha/2}\sqrt{V(r)}$

Tabla 4.3: Intervalos de confianza por TLC al $(1-\alpha) \times 100\,\%$ de confianza para los estimadores fundamentales del MAS. Véanse la Tabla 4.1 y la Tabla 4.2.

En [1] se menciona que usualmente las varianzas reales se desconocen, por lo que para calcular los intervalos de confianza mostrados en la Tabla 4.3 se emplean sus estimadores, para esto, basta con sustituir en las expresiones anteriores el estimador correspondiente. En la práctica, los niveles de confianza pueden variar y los más empleados son: 90 %, 95 % y 99 %, lo cual implícitamente obliga a α a tomar los valores de 0.05, 0.025 y 0.005 respectivamene.

Usualmente elegir un tamaño de muestra "suficiente" para que la aproximación normal dada por el TLC sea adecuada, no es una tarea trivial, sin embargo, diversos autores proponen ciertas reglas. Por ejemplo, [10] recomienda un tamaño mínimo de:

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

donde $\displaystyle\sum_{i=1}^{N}(y_i - \overline{Y})^3$ es la asimetría de la población; de esta manera, cuando la asimetría es grande, se necesita un tamaño de muestra grande para que la aproximación normal sea válida. Otras propuestas más usuales como $n = 30$ citado como un número suficientemente grande, a menudo no es suficiente en problemas de muestreo de poblaciones finitas [8].

4.1.3.2. Tamaño de muestra dada una precisión

Para lograr una *precisión* o margen de error deseado, que se denota usualmente como ε, lo que se hace es expresar la distancia (en términos matemáticos del valor absoluto) que tiene el estimador $(\hat{\theta})$ de su valor objetivo (θ), razón por la cual a menudo se busca que el estimador sea insesgado. Todo esto ligado de una confianza del $(1-\alpha)\times 100\,\%$ a través de una medida de probabilidad. En términos generales, esto lo podemos expresar en términos matemáticos como:

$$\mathbb{P}\left[\left|\hat{\theta} - \theta\right| \leq \varepsilon\right] = 1 - \alpha. \tag{4.11}$$

De tal manera que aplicando (4.11) con el objetivo de buscar una precisión para el estimador de la media $(\overline{y})$ se tiene que al aplicar al TLC y tomando $Z \sim Normal(0,1)$ donde su función de distribución acumulada será denotada por $\Phi(\cdot)$,

$$\begin{aligned} 1 - \alpha = \mathbb{P}\left[\left|\overline{y} - \overline{Y}\right| \leq \varepsilon\right] &= \mathbb{P}\left[-\frac{\varepsilon}{\sqrt{V(\overline{y})}} \leq \frac{\overline{y} - \overline{Y}}{\sqrt{V(\overline{y})}} \leq \frac{\varepsilon}{\sqrt{V(\overline{y})}}\right] \\ &\approx \mathbb{P}\left[-\frac{\varepsilon}{\sqrt{V(\overline{y})}} \leq Z \leq \frac{\varepsilon}{\sqrt{V(\overline{y})}}\right] \overset{(i)}{=} 2\Phi\left(\frac{\varepsilon}{\sqrt{V(\overline{y})}}\right) - 1, \end{aligned}$$

donde (i) ocurre al aplicar la simetría respecto al cero de la distribución Normal Estándar. De este último resultado, se sigue que recordando lo visto en la Tabla 4.2,

$$
\begin{aligned}
1-\alpha &= 2\Phi\left(\frac{\varepsilon}{\sqrt{V(\overline{y})}}\right)-1 \\
&\iff \frac{\varepsilon}{\sqrt{V(\overline{y})}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right) := z_{1-\alpha/2} \\
&\iff \varepsilon = z_{1-\alpha/2}\sqrt{V(\overline{y})} = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}.
\end{aligned} \tag{4.12}
$$

Derivado de (4.12), se tiene que si se conoce la varianza (S^2), el error deseado se alcanza con un tamaño de muestra igual a:

$$
n = \frac{z_{1-\alpha/2}^2 S^2}{\varepsilon^2 + \frac{z_{1-\alpha/2}^2 S^2}{N}}. \tag{4.13}
$$

De tal manera que este error se observará con una confianza del $(1-\alpha)\times 100\,\%$.

Nota: Si se cambia (4.11) de tal manera que $\mathbb{P}\left[\left|\hat{\theta}-\theta\right| \leq \varepsilon\right] \geq 1-\alpha$, lo cual significaría que se desea observar el fenómeno con una confianza de al menos el $(1-\alpha)\times 100\,\%$; se puede mostrar bajo un procedimiento análogo y recordando que $\Phi(\cdot)$ es una función de distribución acumulada de una variable aleatoria continua sobre los reales, lo cual se sabe que es creciente y por lo tanto con inversa igualmente creciente y que en consecuencia preserva las desigualdades, que entonces para observar el error deseado con una confianza de al menos el $(1-\alpha)\times 100\,\%$ es necesaria una muestra de tamaño al menos $n \geq \frac{z_{1-\alpha/2}^2 S^2}{\varepsilon^2+\frac{z_{1-\alpha/2}^2 S^2}{N}}$. De tal manera que (4.13) nos da un valor **mínimo** del tamaño de muestra para observar el error deseado con la confianza solicitada.

Análogamente, para el estimador de razón se tiene que:

$$
\varepsilon = z_{1-\alpha/2}\sqrt{V(r)} = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{S_d^2}{n\overline{X}^2}}. \tag{4.14}
$$

Despejando n de (4.14), cuando S_d^2 es conocida, el tamaño de muestra necesario para alcanzar un error ε resulta ser:

$$
n = \frac{N z_{1-\alpha/2}^2 S_d^2}{N\varepsilon^2\overline{X}^2 + z_{1-\alpha/2}^2 S_d^2}. \tag{4.15}
$$

Nota: Bajo un procedimiento análogo al mencionado en la nota anterior, se puede mostrar que lo que se establece en (4.15) es un valor **mínimo** del tamaño de muestra n para el cual se alcanza el error deseado con la confianza solicitada.

4.2. Muestreo Aleatorio Estratificado

En un muestreo estratificado, la población total de N elementos es dividida en H subpoblaciones llamadas **estratos** de tal manera que se forma una partición de la población original, *i.e.*, cada uno de los estratos son ajenos dos a dos y la unión constituye la población total. De modo que cada unidad obtenida del muestreo pertenece a exactamente un estrato. Para poder trabajar con un muestreo estratificado, se debe conocer previamente los elementos de la población y la cantidad de estos que pertenecerá a cada uno de los H estratos; el h-ésimo estrato tendrá un tamaño dado N_h, con $h \in \{1, 2, \ldots, H\}$ y tal que:

$$\sum_{h=1}^{H} N_h = N.$$

En cada estrato se toman muestras independientes de n_h elementos seleccionados aleatoriamente. Cuando esto se realiza por MAS, entonces a todo el proceso se le conoce como *Muestro Aleatorio Estratificado* (MAE). Se define $\mathcal{S}_h$ como el conjunto de los n_h elementos tomados por MASSR en el estrato h. El tamaño total de la muestra será:

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

De tal manera que, para el h-ésimo estrato se tiene la notación mostrada en la Tabla 4.4 y en la Tabla 4.5 vemos las principales estadísticas para los estratos con sus valores reales y muestrales.

Notación	Descripción
N_h	Número total de elementos en el estrato h.
n_h	Número de elementos en la muestra para el estrato h.
y_{hj}	Valor del j-ésimo elemento en el estrato h.

Tabla 4.4: Notación de entidades para el h-ésimo estrato.

En la Tabla 4.6 podemos observar las estadísticas globales de la muestra vistas desde el punto de vista estratificado y a partir de lo que se observa en las Tablas 4.4 y 4.5.

4.2.1. Propiedades de los estimadores estratificados

Las propiedades de los estimadores en la Tabla 4.6 se derivan directamente de las propiedades del MASSR.

Expresión Real	Expresión Muestral	Descripción
$Y_h = \sum_{j=1}^{N_h} y_{hj}$	$y_h = \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} = N_h \overline{y}_h$	Total acumulado de los elementos de $\mathcal{S}_h$.
$\overline{Y}_h = \frac{Y_h}{N_h}$	$\overline{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}$	Media dada por todos los elementos en $\mathcal{S}_h$.
$S_h^2 = \sum_{j=1}^{N_h} \frac{\left(y_{hj} - \overline{Y}_h\right)^2}{N_h - 1}$	$s_h^2 = \sum_{j \in \mathcal{S}_h} \frac{\left(y_{hj} - \overline{y}_h\right)^2}{n_h - 1}$	Varianza dada por todos los elementos en $\mathcal{S}_h$.

Tabla 4.5: Expresiones reales y muestrales de las estadísticas para el h-ésimo estrato.

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

Tabla 4.6: Parámetros poblacionales por estratificación con su correspondiente estimador (el subíndice e denota estratificación).

Insesgamiento: Dado que en cada estrato se toma un MASSR, entonces $\overline{y}_h$ es un estimador insesgado para la **media** real del estrato, *i.e.*, $\mathbb{E}\left[\overline{y}_h\right] = \overline{Y}_h$. En consecuencia, $\overline{y}_e$ es también insesgado para la media poblacional $\overline{Y}_e$. Esto se puede ver de la siguiente manera teniendo como referencia la Tabla 4.6:

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

Consecuentemente y basados en las mismas razones, se sigue el insesgamiento para el **total** estratificado y_e como estimador del total poblacional Y_e:

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

Varianza: Apoyados de las propiedades de un MASSR sobre la varianza de y_h, $V(y_h)$, y del hecho que las muestras son independientes entre los estratos, se tendrán los resultados expresados en la Tabla 4.7.

Estadística	Varianza real	Varianza Estimada
Total	$V(y_e) = \sum_{h=1}^{H} V(N_h \overline{y}_h) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$	$v(y_e) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$
Media	$V(\overline{y}_e) = \frac{1}{N^2} V(y_e)$	$v(\overline{y}_e) = \frac{1}{N^2} v(y_e)$

Tabla 4.7: Expresión de la varianza real y estimada de los estimadores estratificados (el subíndice e denota estratificación).

Donde:

$$S_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} \left(y_{hj} - \overline{Y}_h\right)^2 \qquad \text{y} \qquad s_h^2 = \frac{1}{n_h - 1} \sum_{j \in \mathcal{S}_h} \left(y_{hj} - \overline{y}_h\right)^2 .$$

Es importante notar que, $v(y_e)$ es un estimador insesgado para $V(y_e)$ ya que s_h^2 lo es para S_h^2. Esto se sigue del MAS tomado en cada estrato. Análogamente, $v(\overline{y}_e)$ es un estimador insesgado para $V(\overline{y}_e)$ [8].

4.2.2. Estimador de razón estratificado

De acuerdo con [1], y apoyados de lo visto en la Subsección 4.1.2, para un diseño muestral por muestreo aleatorio estratificado existen dos métodos que normalmente son empleados para estimar la razón (R como en la Tabla 4.1) entre dos variables, la principal (y) y otra auxiliar (x): **estimador separado** (r_s) y **estimador combinado** (r_c).

4.2.2.1. Estimador Separado

Para el caso del estimador separado, r_s, y definiendo a $r_h = \frac{\overline{y}_h}{\overline{x}_h}$, la estimación del total de la variable y es:

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} X_h \frac{\overline{y}_h}{\overline{x}_h} = \sum_{h=1}^{H} X_h r_h. \tag{4.16}$$

Donde $\overline{y}_h$, y, $\overline{x}_h$ son las medias estimadas de cada variable en el estrato $h \in \{1, 2, \ldots, H\}$, véase la Tabla 4.5. Asumiendo que X_h es el total en el estrato h (debe ser conocido previamente), el estimador separado de razón poblacional es

$$r_s = \frac{\hat{t}_y}{X}, \tag{4.17}$$

donde $X = \sum_{h=1}^{H} X_h$ es el total de la variable auxiliar y debe ser conocida previamente.

Por otro lado, la esperanza de este estimador depende fuertemente de la razón en cada estrato, *i.e.*, $r_h = \frac{\overline{y}_h}{\overline{x}_h}$. Esto es,

$$\mathbb{E}\left[r_s\right] \overset{(4.17)}{=} \mathbb{E}\left[\frac{\hat{t}_y}{X}\right] \overset{(4.16)}{=} \frac{1}{X} \sum_{h=1}^{H} X_h \mathbb{E}\left[\frac{\overline{y}_h}{\overline{x}_h}\right] = \frac{1}{X} \sum_{h=1}^{H} X_h \mathbb{E}\left[r_h\right] \overset{(4.7)}{\approx} \frac{1}{X} \sum_{h=1}^{H} X_h R_h. \tag{4.18}$$

Obsérvese que en este caso X y en general las X_h deben ser constantes conocidas. Además, definimos $R_h = \frac{\overline{Y}_h}{\overline{X}_h}$ como en MAS y más en particular en la Subsección 4.1.2 y con base en la Tabla 4.5 para el h-ésimo estrato. Resulta ser que este valor es en

general sesgado, para notar esta característica, calculemos la covarianza con MAS de tamaño n_h en el h-ésimo estrato, de las cantidades r_h y $\overline{x}_h$. Recordando nuevamente la Tabla 4.5 y sus propiedades aplicadas a las variables y, y x, se tiene que:

$$Cov(r_h, \overline{x}_h) = \mathbb{E}\left[\frac{\overline{y}_h}{\overline{x}_h}\overline{x}\right] - \mathbb{E}\left[r_h\right]\mathbb{E}\left[\overline{x}_h\right] = \overline{Y}_h - \overline{X}_h\mathbb{E}\left[r_h\right].$$

Despejando de la ecuación anterior,

$$\mathbb{E}\left[r_h\right] = \frac{\overline{Y}_h}{\overline{X}_h} - \frac{1}{\overline{X}_h}Cov(r_h, \overline{x}_h) = R_h - \frac{1}{\overline{X}_h}Cov(r_h, \overline{x}_h). \tag{4.19}$$

Recordando (4.7) aplicado al h-ésimo estrato, obtenemos en (4.19) que el sesgo del estimador r_h es:

$$sesgo(r_h) = -\frac{1}{\overline{X}_h}Cov(r_h, \overline{x}_h).$$

De donde se concluye que r_h, como el estimador del valor real $R_h = \frac{\overline{Y}_h}{\overline{X}_h}$ es sesgado para cada estrato. En consecuencia, y derivado de (4.18), de forma general para el **estimador separado** (r_s) se tiene que:

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

Lo anterior demuestra que el **estimador separado** (r_s) es sesgado de la razón poblacional (R), con un sesgo dado por:

$$sesgo(r_s) = -\frac{1}{X}\sum_{h=1}^{H} N_h Cov(r_h, \overline{x}_h). \tag{4.20}$$

La varianza de r_s dado por (4.17) depende totalmente de la varianza de $\hat{t}_y$, la cual a su vez, depende de las varianzas de las razones estimadas en cada estrato, *i.e.*, r_h dadas por un MAS. Obsérvese entonces que, al ser X_h fijo y cada uno de los estratos se muestrea de forma independiente, entonces:

$$
\begin{aligned}
V(\hat{t}_y) &\overset{(4.16)}{=} \sum_{h=1}^{H} X_h^2 V\left(\frac{\overline{y}_h}{\overline{x}_h}\right) = \sum_{h=1}^{H} X_h^2 V\left(r_h\right) \\
&\overset{(4.8)}{\approx} \sum_{h=1}^{H} X_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{dh}^2}{n_h \overline{X}_h^2} \\
&= \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{dh}^2}{n_h}.
\end{aligned}
\tag{4.21}
$$

Donde

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

Además, tendremos que:

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

Lo anterior, pues se definen análogamente a (4.8) y (4.9) aplicado para el h-ésimo estrato. Por lo tanto, se construye en la Tabla 4.8 con las Expresiones de las varianzas real y estimada para el **estimador separado** r_s.

Varianza Real	**Varianza Estimada**
$V(r_s) \overset{(4.17)}{=} \frac{1}{X^2} V(\hat{t}_y) \overset{(4.21)}{\approx} \sum_{h=1}^{H} \left(\frac{N_h}{X}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{dh}^2}{n_h}$	$v(r_s) = \sum_{h=1}^{H} \left(\frac{N_h}{X}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{dh}^2}{n_h}$

Tabla 4.8: Expresión de la varianza real y estimada del **estimador separado**, r_s.

Por último, el insesgamiento de s_{dh}^2 se sigue de las propiedades de MAS para cada estrato, *i.e.*, $\mathbb{E}\left[s_{dh}^2\right] = S_{dh}^2$ y en consecuencia, $v(r_s)$ es también un estimador insesgado para $V(r_s)$. Sin embargo, tal y como vemos en la Tabla 4.8, derivado de (4.21), la varianza no es exacta y por lo tanto existe una aproximación que puede desembocar en sesgo al momento de estimar la misma.

4.2.2.2. Estimador Combinado

A diferencia del **estimador separado**, el **estimador combinado**, r_c, se obtiene utilizando los totales estratificados Y_e y X_e definidos como en la Tabla 4.6, con sus

correspondientes estimadores y_e y x_e. De esta manera, el estimador combinado para la razón poblacional[1], $R_c = \frac{Y_e}{X_e} = \frac{\overline{Y}}{\overline{X}} = R$ apoyados de la Tabla 4.6, es:

$$r_c = \frac{y_e}{x_e}. \tag{4.22}$$

Y a continuación mostraremos algunas de sus propiedades, recordando lo visto en la Subsección 4.2.1. Para calcular la esperanza de este estimador, comenzaremos calculando la covarianza entre las cantidades r_c y x_e, de forma análoga al **estimador separado**, *i.e.*,

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

Despejando de esta última ecuación, se sigue que:

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

Esto significa que el **estimador combinado** está sesgado en su estimación a R_c con un sesgo dado por:

$$sesgo(r_c) = -\frac{1}{X_e}Cov(r_c, x_e). \tag{4.23}$$

Por lo tanto, para realizar el cálculo de la varianza, se realizará un procedimiento similar al visto en la Subsección 4.1.2. Esto es, considerando:

$$r_c = f(y_e, x_e) = \frac{y_e}{x_e},$$

y después linealizando esta expresión vía series de Taylor de primer orden alrededor del punto (Y_e, X_e) de tal manera que:

$$r_c = f(y_e, x_e) \approx f(Y_e, X_e) + \left(\partial_{y_e}|_{Y_e,X_e}\right)(y_e - Y_e) + \left(\partial_{x_e}|_{Y_e,X_e}\right)(x_e - X_e),$$

que, al calcular las derivadas parciales y sustituir la evaluación del punto en f, se llega a que:

$$\left(\partial_{y_e}|_{Y_e,X_e}\right) = \frac{1}{X_e}, \qquad \left(\partial_{x_e}|_{Y_e,X_e}\right) = -\frac{Y_e}{X_e^2} = -\frac{R_c}{X_e} \quad \text{y} \quad f(Y_e, X_e) = \frac{Y_e}{X_e} = R_c.$$

Llegando así a la siguiente expresión de aproximación para r_c:

[1]La estadística R_c es equivalente a la R de la Tabla 4.1, solo que visto desde el punto de vista estratificado como en la Tabla 4.6. Lo mismo sucede para las demás estadísticas referentes a la Tabla 4.6, son equivalentes con su versión original salvo que se piensan que vienen dadas por los estratos.

$$r_c \approx R_c + \frac{1}{X_e}\left(y_e - R_c x_e\right). \tag{4.24}$$

Partiendo de esta última ecuación, recordando la definición de y_e en Tabla 4.6 aplicada también para x_e, por propiedades de la varianza, y, por el hecho de que los estratos que se muestrean de forma independiente, podemos calcular ahora:

$$\begin{aligned}
V(r_c) &\overset{(4.24)}{\approx} \frac{1}{X_e^2} V(y_e - R_c x_e) \\
&= \frac{1}{X_e^2} V\left(\sum_{h=1}^{H} N_h(\overline{y}_h - R_c \overline{x}_h)\right) \\
&= \frac{1}{X_e^2} V\left(\sum_{h=1}^{H} N_h \sum_{j\in\mathcal{S}_h} \frac{1}{n_h}(y_{hj} - R_c x_{hj})\right) \\
&= \frac{1}{X_e^2} V\left(\sum_{h=1}^{H} N_h \overline{d}_h\right) \\
&= \frac{1}{X_e^2}\left[\sum_{h=1}^{H} V\left(N_h \overline{d}_h\right)\right] \\
&\overset{(i)}{=} \frac{1}{X_e^2}\left[\sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_{dh}^2}{n_h}\right].
\end{aligned} \tag{4.25}$$

Donde (i) se sigue de aplicar lo visto en la Tabla 4.7 sobre el producto de N_h con la variable auxiliar creada $\overline{d}_h$. Derivado entonces de los resultados vistos en la Subsección 4.2.1, procedemos a construir la Tabla 4.9 la cual contiene la Varianza Real (aproximada) y la propuesta de Varianza Estimada.

Varianza Real	**Varianza Estimada**
$V(r_c) \approx \frac{1}{X_e^2}\left[\sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_{dh}^2}{n_h}\right]$	$v(r_c) = \frac{1}{x_e^2}\left[\sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{s_{dh}^2}{n_h}\right]$

Tabla 4.9: Expresión de la varianza real y estimada del **estimador combinado**, r_c.

Donde

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

y,

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

Además, de forma análoga al **estimador separado** al final de la Subsubsección 4.2.2.1, $\mathbb{E}\left[s_{dh}^2\right] = S_{dh}^2$, por lo tanto, la Varianza Estimada es insesgada para la Varianza Real en la Tabla 4.9. De acuerdo con [8], a pesar que el estimador combinado tiene menos sesgo cuando el tamaño de la muestra en algunos estratos es pequeño, si las proporciones varían mucho de un estrato a otro, entonces se aprovecha la eficiencia adicional que brinda la estratificación, como lo hace el estimador de razón separado. Véase además, que la así llamada Varianza "Real" está siendo de cualquier manera aproximada vía series de Taylor con fundamento en (4.25).

4.2.3. Cálculo del tamaño de muestra

Análogamente a la Subsubsección 4.1.3.2, si se desea determinar el tamaño de muestra tal que sea suficiente para satisfacer un nivel fijo de precisión dado (ε), resulta que, de los estimadores estratificados descritos anteriormente, esta labor es complicada como para ser realizada analíticamente. Sin embargo, se puede calcular la precisión considerando diferentes tamaños de muestra n. Por ejemplo, para el **estimador combinado**, se tiene que:

$$\varepsilon = z_{\alpha/2}\sqrt{V(r_c)} = \frac{z_{\alpha/2}}{X_e}\left[\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{dh}^2}{n_h}\right]^{1/2}.$$

De acuerdo con [1], para calcular estas precisiones se requiere primero, asignar tamaños de muestra n_h a cada estrato, donde éstas dependen directamente del tamaño de muestra n. En la Tabla 4.10 se presentan las asignaciones más comunes.

Tipo	Asignación
Igual	$n_h = \frac{n}{H}$
Proporcional	$n_h = n\frac{N_h}{N}$
Óptima	$n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$

Tabla 4.10: Algunos tipos de asignaciones de tamaños de muestra para el muestreo estratificado.

De la Tabla 4.10, concorde con [8], si se desea elegir entre las asignaciones Proporcional y Óptima, se debe tener en cuenta que si las varianzas S_h^2 son aproximadamente iguales en todos los estratos, la asignación proporcional es probablemente la mejor para aumentar la precisión, mientras que si éstas varían mucho, la asignación óptima puede resultar mejor.

Para efectos del Conteo Rápido, lo que se hará para seleccionar un tamaño de muestra será basarse en una estadística que medirá el número de escaños mal asignados dado cierto tamaño de muestra, que se tomará inicialmente como en el Tipo Igual de la

Tabla 4.10. Partiendo de esto, lo que se hace es ajustar acordando mutuamente el INE y el COTECORA un tamaño de muestra apropiado mayor o igual dependiendo de las necesidades geográficas y demográficas de cada estrato. Esto es algo que se comentará más a detalle en futuras secciones.

4.3. Re-muestreo Bootstrap

4.3.1. Función de distribución empírica

Si se considera una muestra de v.a.i.i.d. $X_1, \ldots, X_n \sim F$ donde F es una *Función de distribución acumulada* (CDF). Entonces se puede realizar una estimación de F a través de la *Función de distribución acumulada empírica* (ECDF), definida como:

$$F_n(x) = \frac{\sum_{i=1}^{n} \mathbb{1}\left(X_i \leq x\right)}{n}, \tag{4.26}$$

y puede interpretarse como la CDF que acumula una masa de probabilidad de $1/n$ en cada punto de la recta real dado por X_i. En la Subsección 9.1.1 se agrega un apartado donde se muestra cómo construir intervalos de confianza para la ECDF basándose en las propiedades dadas por ésta definición.

Adicionalmente, nosotros entenderemos como **cuantil** ($q_F(\alpha)$) del $\alpha \times 100\,\%$ de CDF F a través del concepto de inversa generaliza, *i.e.*, como:

$$q_F(\alpha) = \text{ínf}\left\{x \in sop\{F\} : \int_{(-\infty, x]} dF = F(x) \geq \alpha\right\}. \tag{4.27}$$

Donde $sop\{F\}$ es el conjunto de números donde la distribución acumula probabilidad. Si se aplica (4.27) a la ECDF, F_n, entonces a $q_{F_n}(\alpha)$ se le conoce como *cuantil muestral o empírico de* F.

De acuerdo con [11], derivado de la construcción de la ECDF, F_n, se puede definir una manera de obtener estimaciones de medidas derivadas de la distribución de interés F vía un "estimador conector" mejor conocido por su nombre en inglés **plug-in estimator**. Primeramente definimos a $T(F)$ como una **estadística funcional** de la distribución de interés tal como, por mencionar algunos ejemplos, la media $\mu = \int x dF(x)$, la varianza $\sigma^2 = \int (x-\mu)^2 dF(x)$ o la mediana $F^{-1}(1/2)$. Luego, el estimador **plug-in** de $\theta = T(F)$ se define como:

$$\hat{\theta}_n = T(F_n).$$

Lo que significa, en palabras simples que, se utilizará la medida de probabilidad inducida por la ECDF, F_n, dada por la muestra, en lugar de la versión desconocida de la CDF, F.

De igual manera, si se define a $T(F) = \int r(x)dF(x)$ para alguna función $r(x)$, entonces T se le conoce como un "funcional lineal". Este nombre, ya que se satisface, por construcción, que para cualesquiera $a, b \in \mathbb{R}$ y F, G CDF se tiene que:

$$T(aF + bG) = aT(F) + bT(G).$$

Lo cual significa que T es lineal en sus argumentos. Nótese que se están definiendo estos operadores para el caso generalizado de una medida de probabilidad, sin importar el caso continuo donde $\int r(x)dF(x) = \int r(x)f(x)dx$ donde f es la función de densidad asociada a la medida de probabilidad F, el caso discreto donde $\int r(x)dF(x) = \sum_j r(x_j)f(x_j)$ donde f es la función de masa asociada a la medida de probabilidad F, o bien el caso mixto. Sin embargo, sin importar cualquiera de estos últimos casos mencionados, por definición la ECDF, F_n, es discreta, la cual acumula una masa de probabilidad de tamaño $1/n$ a cada observación X_i.

Debido a lo anterior, se tiene que el estimador **plug-in** para un funcional lineal $T(F) = \int r(x)dF(x)$ es:

$$T(F_n) = \int r(x)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i). \tag{4.28}$$

Gracias al resultado (4.28), se tienen múltiples formas de cómo estimar medidas de tendencia central usando la muestra:

1. **Media**: Tomando $\mu = T(F) = \int xdF(x)$, entonces su estimador **plug-in** será:

$$\hat{\mu} = \int xdF_n(x) = \overline{X}_n.$$

2. **Varianza**: Tomando $\sigma^2 = T(F) = \int (x-\mu)^2 dF(x) = \int x^2 dF(x) - \left(\int xdF(x)\right)^2$, entonces su estimador **plug-in** será:

$$\begin{aligned}\hat{\sigma}^2 &= \int x^2 dF_n(x) - \left(\int xdF_n(x)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.\end{aligned}$$

3. **Asimetría/Skewness**: Denotando a la media y la varianza como μ y σ^2 respectivamente de la variable aleatoria X, entonces su *skewness* se define como:

$$\alpha = \frac{\mathbb{E}\left[X-\mu\right]^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left[\int (x-\mu)^2 dF(x)\right]^{3/2}}.$$

Entonces, recordando los dos estimadores anteriores $\hat{\mu}$ y $\hat{\sigma}^2$, su estimador **plug-in** será:

$$\hat{\alpha} = \frac{\int (x-\mu)^3 dF_n(x)}{\left[\int (x-\mu)^2 dF_n(x)\right]^{3/2}} = \frac{\frac{1}{n}\sum_i \left(X_i - \hat{\mu}\right)^3}{\hat{\sigma}^3}.$$

4. **Cuantiles**: Recuperando (4.27), podemos definir a $T(F) = F^{-1}(p) = q_F(p)$ con $p \in [0,1]$. Por lo que su estimador **plug-in** será $T(F_n) = F_n^{-1}(p) = q_{F_n}(p)$ que es lo que definíamos como el cuantil muestral o empírico de F.

Entre otros momentos, medidas de tendencia central y estadísticas derivadas de F.

4.3.2. El Bootstrap

De acuerdo con [11], el **bootstrap** es una metodología para estimar errores estándar y calcular intervalos de confianza. Consideremos $T_n = g(X_1, \ldots, X_n)$ una estadística, es decir, cualquier función de la muestra. Supongamos que queremos conocer la varianza de T_n bajo la distribución F, esto lo denotaremos como $\mathbb{V}_F(T_n)$ para hacer énfasis en que la distribución de la estadística vendrá derivada del comportamiento distribucional de la muestra y por lo tanto se piensa en general como desconocida. Por ejemplo, si $T_n = \overline{X}_n$ entonces $\mathbb{V}_F(T_n) = \sigma^2/n$ donde $\sigma^2 = \int (x-\mu)^2 dF(x)$ y $\mu = \int x dF(x)$. Por lo tanto, la varianza de T_n es una función de F. La idea del bootstrap consiste en dos pasos:

Paso 1: Estimar $\mathbb{V}_F(T_n)$ con $\mathbb{V}_{F_n}(T_n)$. Es decir, utilizando la ECDF de F.

Paso 2: Aproximar $\mathbb{V}_{F_n}(T_n)$ usando simulaciones.

En particular, en el ejemplo de $T_n = \overline{X}_n$, se tiene por el **Paso 1** que $\mathbb{V}_{F_n}(T_n) = \hat{\sigma}^2/n$ donde $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. En este caso, el **Paso 1** es suficiente. Sin embargo, hay casos más complicados donde no podemos escribir en una fórmula "simple" el valor de $\mathbb{V}_{F_n}(T_n)$, razón por la que se necesita el **Paso 2**. Antes de proceder se explicará la idea de *simulación* en bootstrap.

4.3.2.1. Justificación del bootstrap

Supóngase que se tiene una muestra de v.a.i.i.d. $\{Y_i\}_{i=1}^B$ de una distribución G. Entonces, invocando uno de los resultados más clásicos e importantes en probabilidad que es la *Ley de los grandes números* se tiene que:

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow[B\uparrow\infty]{P} \int y dG(y) = \mathbb{E}[Y].$$

Entonces, al tener una muestra grande de G, la media muestral, $\overline{Y}_n$, se puede utilizar para aproximar a $\mathbb{E}[Y]$. En una simulación, se puede tomar a B tan grande como se desee, de tal manera que la diferencia entre $\overline{Y}_n$ y $\mathbb{E}[Y]$ sea despreciable. Más aún, si h es cualquier función con media finita, entonces:

$$\overline{h(Y)}_n = \frac{1}{B}\sum_{j=1}^{B} h(Y_j) \xrightarrow[B\uparrow\infty]{P} \int h(y) dG(y) = \mathbb{E}[h(Y)].$$

En particular,

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j - \overline{Y})^2 = \frac{1}{B}\sum_{j=1}^{B} Y_j^2 - \left(\frac{1}{B}\sum_{j=1}^{B} Y_j\right)^2$$
$$\xrightarrow[B\uparrow\infty]{P} \int y^2 dG(y) - \left(\int y dG(y)\right)^2 = \mathbb{V}_G(Y).$$

Por lo cual, se puede utilizar la varianza muestral de simulaciones para aproximar el valor de $\mathbb{V}_G(Y)$.

4.3.2.2. Estimación de la Varianza Boostrap

Con base en [11] y lo visto en la Subsubsección 4.3.2.1, se ha mostrado que se puede aproximar $\mathbb{V}_{F_n}(T_n)$ por simulación. Entonces $\mathbb{V}_{F_n}(T_n)$ se puede interpretar como "la varianza de T_n si la distribución de los datos es F_n." La pregunta natural que sigue es ¿cómo simular de la distribución de T_n cuando se asume que los datos tienen una distribución F_n? La respuesta consiste en simular una muestra aleatoria $\{X_i^*\}_{i=1}^n$ proveniente de F_n y calcular posteriormente $T_n^* = g(\underline{X}^*) = g(X_1^*, \ldots, X_n^*)$. Esto constituye una observación proveniente de la distribución de T_n. La idea es ilustrada por [11] en el siguiente diagrama:

$$\begin{array}{rccccc} \text{Mundo real} & F & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & T_n = g(X_1, \ldots, X_n) \\ \text{Mundo bootstrap} & F_n & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & T_n^* = g(X_1^*, \ldots, X_n^*) \end{array}$$

Ahora, para poder simular $X_1^*, \ldots, X_n^*$ de F_n, simplemente hay que recordar lo visto en la Subsección 4.3.1, donde vimos que F_n es tal que acumula una masa de probabilidad de $1/n$ en cada uno de los puntos de los datos originales $X_1, \ldots, X_n$. Por lo tanto, [11] hace un énfasis importante sobre que **obtener una observación de F_n es equivalente a obtener un punto aleatorio del conjunto original de datos.**

En otras palabras, para simular $X_1^*, \ldots, X_n^* \sim F_n$ es suficiente con extraer n observaciones con reemplazo de $X_1, \ldots, X_n$. Por lo que, en resumen, **el algoritmo de**

estimación de varianza por bootstrap consiste en:

Paso 1: Obtener $X_1^*, \ldots, X_n^* \sim F_n$, o lo que es lo mismo, una muestra de tamaño n con remplazo de $X_1, \ldots, X_n \sim F$.

Paso 2: Calcular $T_n^* = g(X_1^*, \ldots, X_n^*)$.

Paso 3: Repetir los **Pasos 1 y 2** B-veces para obtener $T_{n,1}^*, \ldots, T_{n,B}^*$.

Paso 4: Tomar

$$v_{boot} = se_{boot}^2 = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2.$$

Todo esto se hace recordando que se están realizando dos aproximaciones de manera implícita en este algoritmo:

$$\mathbb{V}_F(T_n) \approx \mathbb{V}_{F_n}(T_n) \approx v_{boot}.$$

En [11] se propone a medida de ejemplo un pseudo código que permita ejemplificar el uso de bootstrap para estimar el error estándar de la mediana, el cual se muestra a continuación (con su debida traducción al español).

Pseudo código (inspirado en código R) para estimar el error estándar de la mediana vía bootstrap.

```
Dados los datos X = (X(1), ..., X(n)):
T <- median(X) # Esta es la mediana real (sólo referencia).
Tboot <- Vector de longitud B
for(i in 1:B){
    Xstar <- Muestra de tamaño n de X (Con remplazo)
    Tboot[i] <- median(Xstar) # Vector de medianas bootstrap.
}
se_boot <- sqrt(variance(Tboot)) # Error estándar bootstrap.
```

4.3.2.3. Intervalos de Confianza Bootstrap

Dada cierta estadística $T_n = T_n(F)$ existen diferentes métodos para construir intervalos de confianza usando bootstrap. En [11] se muestran algunos de ellos, los cuales presentaremos a continuación.

Método 1: Intervalo Normal. Este es el método más simple, sin embargo, para que sea certero, es necesario que la distribución de T_n sea cercana a la de una Normal:

$$T_n \pm z_{\alpha/2} se_{boot}.$$

Donde: se_{boot} se obtiene como en la Subsubsección 4.3.2.2, $\alpha \in (0,1)$ y en general, z_p con $p \in (0,1)$ es el cuantil del $p \times 100\,\%$ de una Normal Estándar.

Método 2: Intervalo Pivotal. Recordemos que $T_n = T(F_n)$, que es la estadística de interés obtenida a través de la distribución empírica (dada por los datos observados) y denotemos como $T = T(F)$ a la estadística dada por la distribución subyacente que tienen los datos. Luego se define al **pivote** como $R_n = T_n - T$. Posteriormente se toman $T^*_{n,1}, \ldots, T^*_{n,B}$ que denotan las B replicaciones bootstrap de T_n (como en el **Paso 3** de la Subsubsección 4.3.2.2). Ahora, se denota como $H(r)$ a la CDF del **pivote** (R_n):

$$H(r) = \mathbb{P}_F\left(R_n \leq r\right).$$

Luego, se define a $C_n = (a,b) = \left(T_n - H^{-1}\left(1 - \frac{\alpha}{2}\right), T_n - H^{-1}\left(\frac{\alpha}{2}\right)\right)$.

De donde se sigue que:

$$\begin{aligned}
\mathbb{P}\left[a \leq T \leq b\right] &= \mathbb{P}\left[a - T_n \leq T - T_n \leq b - T_n\right] \\
&= \mathbb{P}\left[T_n - b \leq T_n - T \leq T_n - a\right] \\
&= \mathbb{P}\left[T_n - b \leq R_n \leq T_n - a\right] \\
&= H(T_n - a) - H(T_n - b) \\
&= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.
\end{aligned}$$

Por lo cual, C_n resulta ser un intervalo de exactamente $(1-\alpha) \times 100\,\%$ de confianza para la estadística T. El detalle con este método es que los límites de este intervalo, a y b dependen de la distribución desconocida H, sin embargo, se puede hacer una estimación bootstrap de tamaño B de esta CDF utilizando su correspondiente ECDF como en (4.26) de la siguiente manera:

$$H_n(r) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(R^*_{n,b} \leq r\right).$$

Donde $R^*_{n,b} = T^*_{n,b} - T_n$. Ahora, tomando $p \in (0,1)$ denotaremos como $R^*_{n,(p)}$ y $T^*_{n,(p)}$ como los cuantiles del $p \times 100\,\%$ de confianza para las muestras $\{R^*_{n,1}, \ldots, R^*_{n,B}\}$ y $\{T^*_{n,1}, \ldots, T^*_{n,B}\}$ respectivamente. Note que $R^*_{n,(p)} = T^*_{n,(p)} - T_n$. Se sigue entonces que un intervalo de aproximadamente $(1-\alpha) \times 100\,\%$ de confianza es:

$$\begin{aligned} C_n^* &= \left(\hat{a}, \hat{b}\right) \\ &= \left(T_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right), T_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &= \left(T_n - R^*_{n,(1-\alpha/2)}, T_n - R^*_{n,(\alpha/2)}\right) \\ &= \left(2T_n - T^*_{n,(1-\alpha/2)}, 2T_n - T^*_{n,(\alpha/2)}\right). \end{aligned}$$

El cual es el **intervalo bootstrap pivotal** a un nivel de confianza del $(1-\alpha)\times 100\,\%$. Asimismo, se puede probar de acuerdo a [11] que bajo condiciones débiles para $T(F)$,

$$\mathbb{P}_F\left[T(F) \in C_n^*\right] \underset{n\uparrow\infty}{\longrightarrow} 1 - \alpha.$$

Método 3: Intervalo Percentil.

Con base en la definición de $T^*_{n,(p)}$ del **Método 2**, el **intervalo bootstrap percentil** al nivel de confianza $(1-\alpha)\times 100\,\%$ se define como:

$$C_n = \left(T^*_{n,(\alpha/2)}, T^*_{n,(1-\alpha/2)}\right).$$

En [11] se muestra una justificación de porqué la implementación de este intervalo es correcta, en el sentido de que proporciona una aproximación de intervalo de confianza al nivel mencionado.

Para esto, el autor supone la existencia, aunque no sea conocida, de una transformación monótona $U = m(T)$ tal que $U \sim N(\phi, c^2)$ donde $\phi = m(T_n)$. Luego, se toma una muestra boostrap de tamaño B como $U_b^* = m(T_{n,b})$ para cada $b \in \{1, \ldots, B\}$ y se denotan sus cuantiles a nivel $p \times 100\,\%$ como $U^*_{n,(p)}$. Como ϕ es una función monótona, por el teorema de equivarianza de cuantiles, se tiene que $U^*_{n,(p)} = m\left(T^*_{n,(p)}\right)$ para cualquier p dada. Además, como $U \sim N(\phi, c^2)$, el cuantil a nivel $\alpha/2$ de U viene dado por $q_U(\alpha) = \phi - z_{\alpha/2}c$.[2] Debido a esto, $U^*_{n,(\alpha/2)} = \phi + z_{\alpha/2}c$ y análogamente $U^*_{n,(1-\alpha/2)} = \phi + z_{1-\alpha/2}c$. Por lo tanto,

$$\begin{aligned} \mathbb{P}\left[T^*_{n,(\alpha/2)} \leq T \leq T^*_{n,(1-\alpha/2)}\right] &= \mathbb{P}\left[m\left(T^*_{n,(\alpha/2)}\right) \leq m(T) \leq m\left(T^*_{n,(1-\alpha/2)}\right)\right] \\ &= \mathbb{P}\left[U^*_{n,(\alpha/2)} \leq U \leq U^*_{n,(1-\alpha/2)}\right] \\ &= \mathbb{P}\left[\phi + z_{\alpha/2}c \leq U \leq \phi + z_{1-\alpha/2}c\right] \\ &= \mathbb{P}\left[z_{\alpha/2} \leq \frac{U-\phi}{c} \leq z_{1-\alpha/2}\right] \\ &= 1 - \alpha. \end{aligned}$$

[2] En este caso $q_U(\alpha/2)$ bajo la definición (4.27) hace referencia a la CDF, F, subyacente que tiene U. En este caso una Normal.

Nota: El autor en [11] comenta que una transformación de normalización exacta rara vez existirá, sin embargo, puede existir una aproximación de transformación de normalización.

En resumen, los intervalos dados por los tres métodos anteriores, los podemos visualizar en la Tabla 4.11.

Método	**Intervalo**
Normal	$T_n \pm z_{\alpha/2} se_{boot}$
Pivotal	$\left(2T_n - T^*_{n,(1-\alpha/2)}, 2T_n - T^*_{n,(\alpha/2)}\right)$
Percentil	$\left(T^*_{n,(\alpha/2)}, T^*_{n,(1-\alpha/2)}\right)$

Tabla 4.11: Intervalos de confianza para bootstrap a nivel α de significancia por diferentes metodologías.

En la Subsección 9.1.2 se muestra un ejemplo de cómo se ve la programación de un intervalo de confianza por esta metodología.

4.3.2.4. Algoritmo bootstrap y su versión estratificada

En la Sección 4.2 se abordó el tema del *Muestro Aleatorio Estratificado* (MAE) y cómo es su implementación. Recordando que esta metodología consiste en separar a la población total de tamaño N en H subpoblaciones dadas por los estratos y cada una con un tamaño de N_h con $h \in \{1, \ldots, H\}$, en lo que consiste realizar un "*Bootstrap Estratificado*" es simplemente respetar la estructura dada por los H para realizar un re-muestreo bootstrap convencional. Es decir, se aplicará un MAE con reemplazo sobre los estratos con el objetivo de calcular la estadística de interés en el bootstrap.

4.3.2.4.1 Algoritmo Bootstrap (Convencional)

En la Subsubsección 4.3.2.2 se muestra un caso particular de la idea Bootstrap (convencional) con el objetivo de encontrar una varianza para cierta estadística T. **De forma general**, dada una muestra aleatoria $X_1, \ldots, X_n$, que puede ser a su vez una sub-población seleccionada de una más grande de tamaño N, una implementación del Bootstrap (convencional) con el objetivo de obtener una muestra de tamaño B de una estadística T, para posteriormente agregar esta nueva muestra con otra estadística τ_{boot}, vista como un algoritmo, lo que se hace es seguir los siguientes pasos:

Paso 1: Obtener $X^*_1, \ldots, X^*_n \sim F_n$, o lo que es lo mismo, una muestra de tamaño n con remplazo de $X_1, \ldots, X_n \sim F$.

Paso 2: Calcular $T_n^* = g(X_1^*, \ldots, X_n^*)$, *i.e.*, calcular la estadística T de la muestra resultante en el **Paso 1**.

Paso 3: Repetir los **Pasos 1 y 2** B-veces para obtener $T_{n,1}^*, \ldots, T_{n,B}^*$.

Paso 4: Aplicar la estadística de agregación τ_{boot} a la muestra resultante del **Paso 3**.

Nota: En la Subsubsección 4.3.2.2 se tomó $\tau_{boot} = v_{boot}$.

La desventaja de este algoritmo ante una muestra que puede ser estratificada radica en el **Paso 1**, y es que, cuando se muestrea de forma aleatoria de la población $X_1, \ldots, X_n$, cabe la (gran) posibilidad de que exista un desbalance entre observaciones provenientes de cada estrato. En otras palabras, puede haber más observaciones de un estrato que de otro, lo cual no es siempre deseado ya que en muchas ocasiones lo que se busca a través de los estratos, es tener un balance de información entre ellos para no dar más peso a un estrato que a otro.

4.3.2.4.2 Algoritmo Bootstrap Estratificado

Debido a lo anterior, lo que se hace es tomar un MAE en el **Paso 1** del Párrafo 4.3.2.4.1 sobre la muestra de $X_1, \ldots, X_n$ y continuar el algoritmo con esta nueva muestra extraída de forma estratificada.

Recordando la notación de la Sección 4.2 y asumiendo la existencia de un total de H estratos, entonces para cada uno de los $h \in \{1, \ldots, H\}$ estratos de tamaño n_h, denotaremos como F_h a la CDF de la muestra aleatoria $X_{h,1}, X_{h,2}, \ldots, X_{h,n_h}$. De tal manera que, si la muestra se puede dividir en H estratos entonces el algoritmo bootstrap para el caso estratificado se realizaría de la siguiente manera:

Paso 1: Para cada $h \in \{1, \ldots, H\}$ Obtener $X_{h,1}^*, \ldots, X_{h,n_h}^* \sim F_{n_h}$, o lo que es lo mismo, una muestra de tamaño n_h con remplazo de $X_{h,1}, \ldots, X_{h,n_h} \sim F_h$.

Paso 2: Calcular

$$T_n^* = g(X_{1,1}^*, \ldots, X_{1,n_1}^*, X_{2,1}^*, \ldots, X_{2,n_2}^*, \ldots, X_{h,1}^*, \ldots, X_{h,n_h}^*, \ldots, X_{H,1}^*, \ldots, X_{H,n_H}^*),$$

i.e., calcular la estadística T de **toda** la muestra resultante de tamaño $n = \sum_{h=1}^{H} n_h$ del **Paso 1**.

Paso 3: Repetir los **Pasos 1 y 2** B-veces para obtener $T_{n,1}^*, \ldots, T_{n,B}^*$.

Paso 4: Aplicar la estadística de agregación τ_{boot} a la muestra resultante del **Paso 3**.

De esta manera estaremos evitando que exista un desbalance entre los estratos que se están trabajando y se estará controlando de forma estratificada la manera de obtener las estadísticas de interés de la muestra.

Esto cobra vital importancia para la aplicación de este capítulo al Conteo Rápido, ya que los estratos, en este caso, vendrán dados por los distritos federales electorales y ya que cada uno de éstos estará otorgando un escaño por el principio de *Mayoría Relativa* (MR) es importante no dar pie a que se pierda muestra de algún estrato o exista un sesgo dado por un estrato de mayor tamaño, N_h. En el siguiente capítulo se abordará la manera en cómo se trabajó el diseño de muestreo para el Conteo Rápido por los miembros del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA).

Capítulo 5

Muestreo aplicado al Conteo Rápido

5.1. Estimación de los porcentajes de votos para cada partido con respecto a la votación válida emitida

Con base en lo establecido en [2], uno de los objetivos particulares del Conteo Rápido Federal es estimar el porcentaje de votos con respecto a la *Votación Válida Emitida* (VVE) (Subsubsección 3.2.2.3), es decir estimar el vector $pVVE$ en (3.8). Notando que este vector está conformado por partidos políticos y el agrupado de candidatos independientes, para fines prácticos consideraremos, sin pérdida de generalidad, $p \in pVVE$ como alguno de estos porcentajes; la aplicación del muestreo probabilístico será el mismo $\forall p \in pVVE$. Esta cantidad es el valor real y se obtiene una vez realizados los Cómputos Distritales de la elección. Sin embargo, siguiendo la teoría del muestreo probabilístico, se selecciona una muestra aleatoria de n casillas (muestra total), del total de las N casillas instaladas, y con la información recuperada n_e (muestra efectiva) en la tarde-noche del día de la elección se calcula el estimador $\hat{p}$. Utilizando las herramientas de muestreo, es posible definir **estrategias de selección, tamaños de muestra** y **estimadores** para asegurar que, con un nivel de confianza del 95 %:

$$|p - \hat{p}| \leq \varepsilon. \tag{5.1}$$

A la cantidad ε se le conoce comúnmente como **precisión, margen de error** o **error máximo aceptable** en la estimación (Subsubsección 4.1.3.2), y se fija de acuerdo con las exigencias de la elección y con la capacidad operativa de campo. La expresión (5.1) se puede escribir de forma equivalente en términos de intervalos de confianza como:

$$\hat{p} - \varepsilon \leq p \leq \hat{p} + \varepsilon. \tag{5.2}$$

Esto se puede interpretar tomando como referencia el siguiente ejemplo general: se extraen B muestras distintas e independientes una de otra (cada una siguiendo la

misma estrategia de selección, usando el mismo tamaño de muestra y el mismo estimador) y con cada muestra se genera una estimación, de tal manera que se obtengan B estimaciones (independientes) $\{\hat{p}_1, \ldots, \hat{p}_B\}$ entonces aproximadamente $B \times 0.95$ de estas estimaciones cumplirán con que su distancia a p, es menor a ε.

5.2. Porcentaje de Participación Ciudadana

5.2.1. Fórmulas asintóticas Vs. Bootstrap

Para esta sección, recordaremos lo visto en la Sección 4.1 y Sección 4.2 para obtener intervalos de confianza asintótica utilizando la teoría del muestreo probabilístico, en particular con los temas de intervalos de confianza en la aproximación asintótica de la distribución de los estimadores (Subsubsección 4.1.3.1) y el estimador combinado (Subsubsección 4.2.2.2) como una versión estratificada del estimador de razón (Subsección 4.1.2). El objetivo será poner a prueba lo dicho anteriormente contra lo visto en la Sección 4.3 correspondiente a bootstrap.

Para hacer este contraste en las fórmulas asintóticas y el bootstrap, lo que se hará es tomar como ejemplo otro de los objetivos que tiene el Conteo Rápido Federal, que es estimar el porcentaje de participación ciudadana ($PART$). Este porcentaje resulta ser uno de los más difíciles de estimar por el Conteo Rápido debido a su amplia volatilidad que involucra razones no solamente matemáticas sino también sociales. El valor real del porcentaje de participación ciudadana viene dado por los resultados dados por los cómputos distritales y se calcula como el cociente del total de la *Votación Total Emitida* (VTE) entre el total de ciudadanos mostrados en la *Lista Nominal* (LN), es decir:

$$PART = \frac{VTE}{LN} \tag{5.3}$$

De tal manera que, recordando la definición de VTE en la Subsubsección 3.2.2.2, la participación (5.3) se puede interpretar como el porcentaje de personas en la *Lista Nominal* (que tienen derecho a votar) que decidieron ejercer el sufragio (ir a votar), aunque sea en favor de candidatos no registrados, o bien, anulando su voto.

Esta comparación se hará a través de intervalos al 95 % de confianza para la estimación de la participación. Este procedimiento será realizado a través de un diseño MAE (Sección 4.2), tomando como estratos los $H = 300$ distritos federales electorales, los cuales cuentan cada uno con cierta cantidad de casillas (N_h), las cuales se tomarán de forma estratificada con diferentes tamaños de muestra por estrato de n_h iguales a 5, 10 15 y 20 casillas para cada $h \in \{1, \ldots, H\}$, *i.e.*, se está tomando una asignación de tamaño de muestra del **Tipo Igual** para el MAS (Tabla 4.10). Lo anterior significa implícitamente que se está tomando una muestra total $n = n_h \times H$ de $5 \times (300) = 1,500$, $10 \times (300) = 3,000$, $15 \times (300) = 4,500$ y $20 \times (300) = 6,000$

casillas. En otras palabras, calcularemos un intervalo de confianza al 95 % asintótico y otro boostrap de la participación (5.3) primero tomando $n_h = 5$ casillas en cada estrato, posteriormente lo haremos tomando $n_h = 10$, luego $n_h = 15$ y finalmente $n_h = 20$.

Para hacer esto estaremos utilizando la base de datos de Cómputos Distritales para 2018, ya que, además de las variables mencionadas en la Tabla 3.2, también existen otras dos llamadas `TOTAL_VOTOS_CALCULADOS`, que es el total agregado de los votos hacia los partidos políticos, candidatos independientes, no registrados y votos nulos; es decir la VTE, y `LISTA_NOMINAL_CASILLA` que indica el total de personas que pueden votar para la casilla en cuestión. Nótese que para cada casilla, `TOTAL_VOTOS_CALCULADOS` $\leq$ `LISTA_NOMINAL_CASILLA`. Tomando como base estos dos datos, así como el identificador de estrato dado por la concatenación de las variables `ID_ESTADO`, y, `ID_DISTRITO` , los pasos a realizar son las siguientes:

Paso 1 (Selección de la muestra): Se tomará una muestra aleatoria estratificada del **Tipo Igual** (Tabla 4.10) de tamaño n_h por estrato. Tomando como estratos los $H = 300$ distritos federales electorales.

Paso 2 (Intervalo Asintótico): Con la muestra seleccionada en **Paso 1**, se calculará el intervalo de confianza para el estimador combinado de la participación utilizando como base (4.22) y la Tabla 4.9 para poder obtener dicho intervalo como en la Tabla 4.3 a un 95 % de confianza.

Paso 3 (Intervalo Bootstrap): Con la muestra seleccionada en **Paso 1**, se realizará un bootstrap estratificado[1] de tamaño $B = 10,000$ en donde para cada una de estas B muestras se calculará la participación, de tal manera que se tendrán un total B de estimaciones de la participación, con las cuales se construirá un intervalo de confianza por el **Método Percentil** como en la Tabla 4.11.

Paso 4 (Experimentación): Se repetirán los **Pasos 1 al 3**, un total de 500 (cantidad seleccionada de manera arbitraria para tomar una muestra "considerablemente" grande) veces para cada una de las n_h mencionadas anteriormente. De esta manera, para cada valor asignado de n_h, tendremos un total de 500 intervalos por cada metodología (Asintótica y Bootstrap). En cada valor asignado de n_h y para cada metodología, se tomará el promedio del extremo inferior de los 500 intervalos y el promedio del extremo superior de los 500 intervalos, dándonos así un único "intervalo promedio" para las dos metodologías en cuestión. En otras palabras, si llamamos a los intervalos resultantes de repetir los pasos anteriores 500 veces como $\left(L_j^{(n_h)}, U_j^{(n_h)}\right)$ con $j \in \{1, \ldots, 500\}$ y $n_h \in \{5, 10, 15, 20\}$, **en cada metodología**, se está tomando un "intervalo promedio" para cada n_h dado por:

[1] Tomando a consideración los estratos, se muestrea con reemplazo en cada uno de estos para tener una misma proporción de observaciones a lo largo de los distritos federales electorales, tal como se vio en el Párrafo 4.3.2.4.2.

$$\left(L^{(n_h)}, U^{(n_h)}\right) = \left(\frac{1}{500}\sum_{j=1}^{500} L_j^{(n_h)}, \frac{1}{500}\sum_{j=1}^{500} U_j^{(n_h)}\right).$$

Los resultados de este experimento se pueden ver en la Figura 5.1 y la Tabla 5.1, donde se puede apreciar una ligera mejora en el intervalo asintótico, **sin embargo, es numéricamente poco significativa** por lo que se deduce que utilizar la metodología asintótica proporciona una ganancia bastante pobre sobre la metodología asintótica. Además, para ciertas estadísticas que veremos más adelante, no existe una fórmula asintótica por lo que la opción a tomar será realizar bootstrap. Cabe mencionar que elecciones de tamaños de muestra por estrato de $15 \leq n_h \leq 25$ se tomaron casi siempre para la realización de conteos rápidos en años pasados. El desarrollo computacional de este experimento fue realizado en código de R y puede ser consultado en el GitHub del autor dando clic aquí[2].

Tamaño de Muestra por Estrato	**Método: Asintótico**		**Método: Bootstrap**	
n_h	$L^{(n_h)}$	$U^{(n_h)}$	$L^{(n_h)}$	$U^{(n_h)}$
5	0.6274	0.6383	0.6287	0.6383
10	0.6291	0.6368	0.6300	0.6372
15	0.6298	0.6361	0.6306	0.6366
20	0.6302	0.6356	0.6309	0.6361

Tabla 5.1: Comparación numérica de los intervalos de confianza para la estimación de la Participación (5.3) obtenidos por bootstrap y fórmulas asintóticas. El valor real (dado por toda la población) es $PART = 0.6329418$.

5.3. Tamaño de muestra

5.3.1. Medida de error para estimar la conformación de la cámara

Otro de los objetivos establecidos para el Conteo Rápido Federal, es la estimación de la conformación de la cámara de diputados (Subsección 3.2.5). Este cálculo lleva consigo implícitamente las estimaciones del porcentaje de votos dada por $pVVE$ y el porcentaje de participación nacional en la elección. La estimación de la conformación

[2]https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/5.%20Muestreo%20Aplicado%20al%20CR/5.2.1

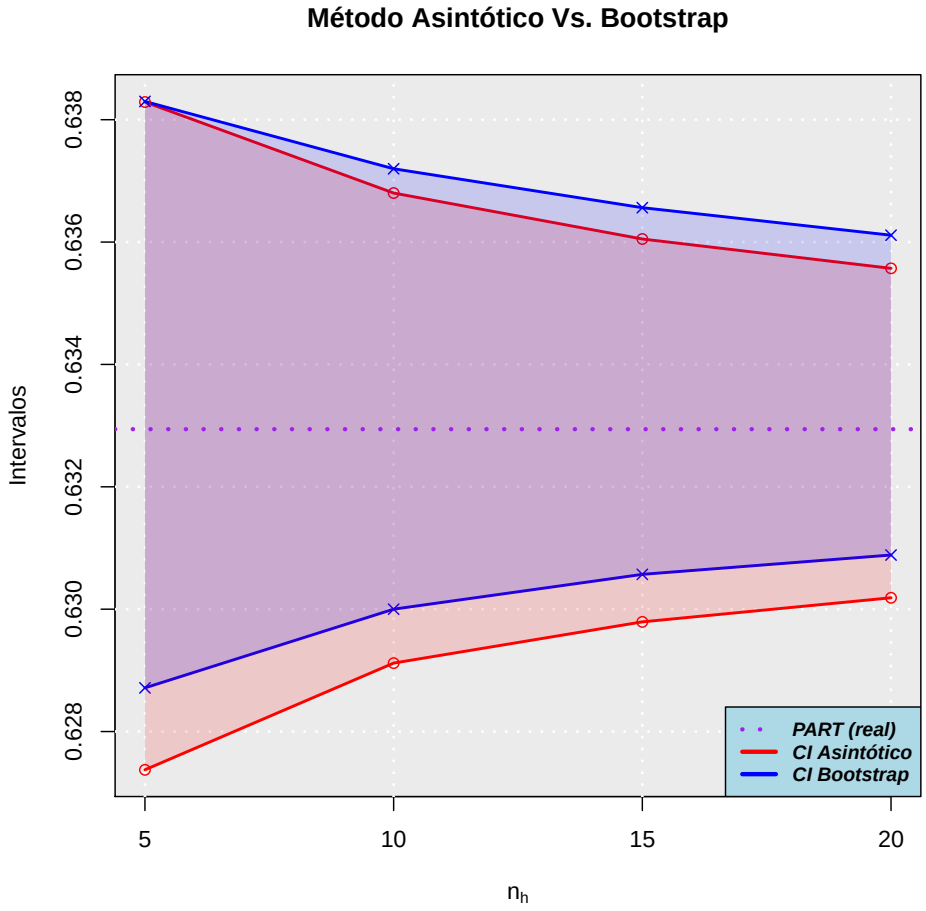


Figura 5.1: Comparación gráfica de los intervalos de confianza para la estimación de la Participación (5.3) obtenidos por bootstrap y fórmulas asintóticas.

de la cámara de diputados no es un problema estándar y la teoría del muestreo no se puede aplicar directamente.

Lo que se busca en esta sección es introducir un criterio análogo al descrito en la expresión (5.1), pero para la conformación de la cámara. Primero es necesario **definir una medida de error** a la cual se le ha llamado como el **número máximo de escaños mal asignados** ($MEMA$). Recordando que la conformación de la cámara viene dada por (3.1), formalmente este valor se define como el número de escaños mal asignados (EMA), para el partido con el mayor número de escaños mal asignados incluyendo el agrupado de los candidatos independientes, es decir,

$$\begin{aligned} MEMA &= \text{máx}\left\{\left|NP_1 - \widehat{NP}_1\right|, \ldots, \left|NP_k - \widehat{NP}_k\right|, \left|NI - \widehat{NI}\right|\right\} \\ &= \text{máx}\left\{EMA_1, \ldots, EMA_k, EMA_I\right\} \\ &= \max_{s\in\{1,\ldots,k,I\}}\left\{EMA_s\right\}, \end{aligned} \tag{5.4}$$

donde $\widehat{NP}_j$ es el número de escaños en la cámara de diputados asignados al partido j, y $\widehat{NI}$ el análogo para el agrupado de los candidatos independientes **obtenido mediante la estimación del conteo rápido**.

Es posible definir otros criterios de forma similar. Sin embargo, el criterio (5.4) es fácil de comprender y en los ejercicios realizados para elecciones pasadas se observaron buenos resultados (Tabla 5.2 y Tabla 5.3).

Para definir el tamaño de muestra en la Elección Federal, primero se establecerá un máximo error permisible en la asignación de escaños, denotado como d, y en las siguientes secciones se buscará determinar el tamaño de muestra necesario para garantizar que, con un 95 % de confianza se tenga que:

$$MEMA \leq d. \tag{5.5}$$

Al no ser posible encontrar expresiones analíticas que permitan despejar el tamaño de muestra necesario para alcanzar un margen de error d, en este caso, mediante ejercicios de simulación se aproximará la distribución para (5.4) usando diferentes tamaños de muestra. Esto ya tiene una argumentación vista en la Subsección 5.2.1 donde vimos que la implementación del *boostrap* sobre las fórmulas asintóticas no tiene una diferencia significativa para encontrar intervalos de confianza. Históricamente, los intervalos *bootstrap* han sido utilizados en múltiples ocasiones para Conteos Rápidos Federales pasados y han mostrado brindar buenos resultados.

5.3.2. Ejemplo del cálculo del número máximo de escaños mal asignados: Conteo Rápido del 2003

Para las elecciones federales del 2003 hubo 3 equipos integrados por miembros del COTECORA que se encargaron de estimar la conformación de la cámara de diputados, a continuación presentamos sus resultados y un ejemplo de cómo se obtiene el $MEMA$ de (5.4).

En la Tabla 5.2 se muestran los resultados de la estimación de la cámara de diputados por los 3 equipos encargados el día de la jornada electoral el 6 de Julio del 2003. También, se muestra el resultado final dado por los cómputos distritales el día 12 de Julio del 2003.

Partidos	Resultado Final (12 de Julio, 2003)	Estimación puntual de la conformación de la Cámara de Diputados 2003 (6 de Julio, 2003)		
		Equipo 1	Equipo 2	Equipo 3
PAN	153	155	152	155
PRI	226	223	226	224
PRD	95	96	96	95
PVEM	15	15	15	15
PT	6	6	6	6
Convergencia	5	5	5	5

Tabla 5.2: Conformación de la cámara de diputados en la Elección Federal de 2003: Estimaciones puntuales de los tres equipos.

Por otro lado, en la Tabla 5.3 se muestra la cantidad de escaños en los que se equivocó cada equipo por partido. Posteriormente, en el último renglón de esta tabla se muestra el número MEMA para cada uno de los equipos. De esta manera es como se obtiene la estadística dada en (5.4).

Partidos	Escaños mal asignados (EMA)		
	Equipo 1	Equipo 2	Equipo 3
PAN	2	1	2
PRI	1	2	0
PRD	1	1	0
PVEM	2	2	2
PT	0	0	0
Convergencia	0	0	0
$MEMA$	2	2	2

Tabla 5.3: Error usando el número Máximo de Escaños Mal Asignados ($MEMA$) en la estimación de la conformación de la cámara de diputados en la Elección Federal de 2003.

5.4. Diseño de muestreo para la estimación de la Cámara de Diputados

Para definir el diseño de muestreo para el Conteo Rápido 2021 se usaron las bases de datos con los resultados de los cómputos distritales en las elecciones de diputados 2012, 2015 y 2018 como referencia.

En todos los experimentos se siguió un diseño de muestreo estratificado por distrito federal electoral en donde, en cada iteración, se siguieron los siguientes pasos:

1. Selección por *Muestro Aleatorio Simple Sin Remplazo* (MASSR) (Subsección 4.1.1) de n_h casillas al interior de cada distrito, considerando una de las bases de referencia.

2. Se usó el estimador común del total (*Tabla* 4.6) para aproximar el total de votos en favor de cada fuerza política.

3. Considerando las coaliciones y los votos totales en cada distrito se obtuvieron los diputados por el principio de *Mayoría Relativa* (MR) (Subsubsección 4.2.2.2).

4. Se usó la teoría del *Muestro Aleatorio Estratificado* (MAE) (Sección 4.2) para **combinar** (Subsubsección 4.2.2.2) los resultados de los 300 distritos y estimar las distintas clases de votaciones (Subsección 3.2.2) (*Votación Total Emitida* (VTE), *Votación Válida Emitida* (VVE) y *Votación Nacional Emitida* (VNE)).

5. Utilizando el estimador de razón combinado (4.22) se realizó la estimación de la participación (5.3), así como del porcentaje de la VVE que alcanzó cada partido político.

6. Se estimó la conformación de la Cámara de Diputados.

    a) En este punto ya se cuenta con estimaciones para los diputados por MR, así como para la VTE y VVE.

    b) Se siguieron las reglas marcadas en la *Ley General de Instituciones y Procedimientos Electorales* (LEGIPE, [6]) y en la *Constitución Política de los Estados Unidos Mexicanos* (CPEUM, [4]) para distribuir escaños por *Representación Proporcional* (RP) (Subsección 3.2.4).

Lo enlistado anteriormente constituye únicamente una realización de la simulación, se realizaron 10,000 iteraciones y en cada iteración se calculó el margen de error mostrado en la Ecuación 5.4. Se probaron los tamaños de muestra al interior de cada estrato de $n_h = 5$, 10, 15, 20, 25 y 30, dado que son 300 estratos y en cada estrato se usará el mismo tamaño de muestra, lo anterior implica que los tamaños de muestra a considerar son de $n = 5 \times 300 = 1500$, $10 \times 300 = 3000$, $15 \times 300 = 4500$, $20 \times 300 = 6000$, $25 \times 300 = 7500$ y $30 \times 300 = 9000$ casillas respectivamente.

5.5. Ejemplos con Cómputos Distritales

5.5.1. Elecciones de Diputados 2012, 2015 y 2018

En la Figura 5.2 se muestra la distribución de muestreo para el número MEMA, utilizando como referencia la elección de 2015 y $n_h = 20$ ($n = 6,000$). En esta figura, a través de la función de masa de probabilidad empírica, se puede apreciar la forma distribucional que tiene la estadística de interés. Con base en esto, se toma el cuantil del $1 - \alpha = 95\,\%$ de confianza con el objetivo de satisfacer la condición (5.5). Para eso, de forma general, se tomará un **margen de error** dado por:

$$d = q_{F_n}(95\,\%), \tag{5.6}$$

recordando la definición en (4.27). Para el caso particular de este experimento se obtuvo un margen de error de $d = 9$, el desarrollo computacional del mismo fue realizado en código de R y puede ser consultado en el GitHub del autor dando clic aquí[3].

En la Tabla 5.4 se muestran las estimaciones de diferentes márgenes de error (d) por esta metodología a diferentes niveles de n_h y con los Cómputos Distritales de los años 2012, 2015 y 2018. En esta tabla se observó que un margen de error que se considera apropiado bajo estándares estadísticos del Conteo Rápido en el número MEMA, se da tomando una cantidad n_h entre 20 y 25 de casillas por estrato (distrito federal electoral), esto considerando los costos operativos que tiene incrementar el tamaño de muestra seleccionada y la poca o nula ganancia que se tiene al aumentar la muestra más allá de 25 casillas.

5.6. Diseño de muestreo en el Conteo Rápido 2021

De acuerdo con los datos oficiales establecidos en [2], para el Conteo Rápido en la Elección Federal del año 2021, se siguió un diseño de muestreo estratificado con el objetivo de estimar la conformación de la cámara de diputados, en donde:

1. Los estratos serán los distritos federales electorales por lo que se tendrán $H = 300$ estratos.

2. El tamaño de muestra total fue de $n = 6,240$ casillas a nivel nacional, en donde se seleccionaron:

[3] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/5.%20Muestreo%20Aplicado%20al%20CR/5.5.1

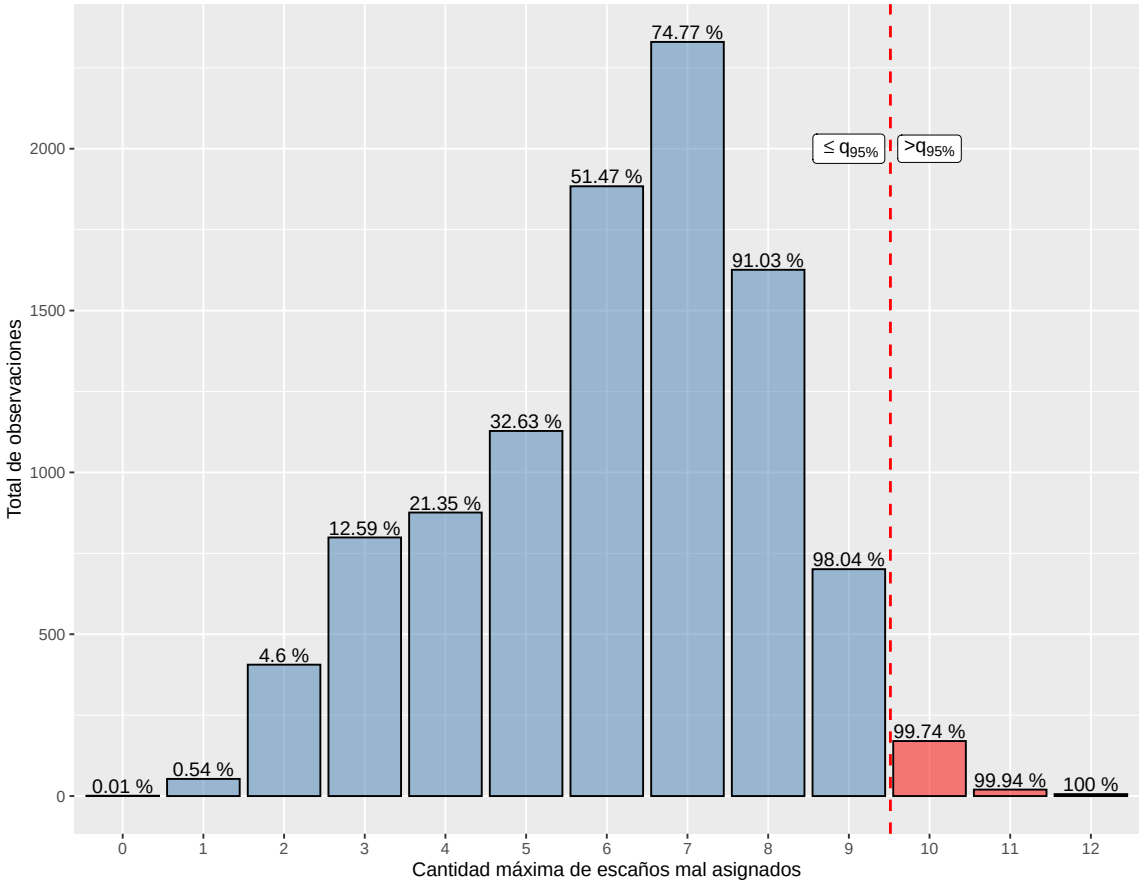


Figura 5.2: Densidad y distribución acumulada estimada por bootstrap del número *Máximo de Escaños Mal Asignados* (MEMA) (5.4). Experimento realizado con 10,000 muestras independientes.

n_h	**2012**	**2015**	**2018**
5	9	10	9
10	7	9	7
15	6	9	6
20	6	9	6
25	5	8	5
30	5	8	5

Tabla 5.4: Margen de error (d) dado por la Ecuación 5.6 para garantizar la condición establecida en la Ecuación 5.5 para la estimación de la conformación de la cámara de diputados vía un muestreo estratificado de diferentes tamaños n_h por el método del tipo de asignación "Igual" de la Tabla 4.10, considerando los cómputos distritales 2012, 2015 y 2018.

- 20 casillas en cada distrito, para un total de 6,000 casillas.
- 10 casillas de sobre-muestra debido a diferencias de huso horario y a que en el Conteo Rápido de 2015 se observó una muy baja recepción de casillas en estos estados. La sobre-muestra total será de 240 casillas. Los estados con sobre-muestra son; Sonora (7 distritos - 70 casillas), Baja California

(8 distritos - 80 casillas) y Guerrero (9 distritos - 90 casillas).

3. En el Conteo Rápido para la Elección Federal de Diputados de 2015 se observó un nivel de respuesta del 74 % y la estimación final se realizó a las 10:15 pm. Por lo tanto, se pronosticó que para el Conteo Rápido Federal de 2021, de las 6,240 casillas totales se recibiera una muestra efectiva de entre 4,000 y 5,500 casillas alrededor de las 10:30 pm del 6 de junio de 2021.

4. Con la muestra efectiva, se estimaba que el margen de error en las estimaciones fuese del:

   - 1 % al estimar la participación (5.3) en la Elección Federal.
   - 0.05 % al estimar la *Votación Válida Emitida* (VVE) por partido.
   - Un valor del número MEMA entre 6 y 7 al estimar la conformación de la cámara de diputados.

Esta estrategia de selección y márgenes de error esperados se obtuvieron tomando como referencia los Cómputos Distritales para las elecciones de diputados en los años 2012, 2015 y 2018.

Para armonizar los diseños para las 15 elecciones locales, se consideró en el diseño que:

- Las estratificaciones para los diseños locales son refinaciones de los distritos federales.
- En los estados con elección local, la muestra para el Conteo Rápido federal fue una sub-muestra de la muestra usada en el Conteo Rápido para la elección local.

Esto significa que, en el Conteo Rápido 2021, se estimaron 15 gubernaturas, así como la conformación de la cámara de diputados. Por tal motivo, en los estados donde hubo elección a Gobernador no se siguió un diseño estratificado por Distrito Federal. De manera general, en los estados donde hubo elecciones para gobernador, se usaron estratificaciones más finas. En palabras simples, uniendo las estratificaciones para las elecciones locales es posible recuperar la estratificación por Distrito Federal. Para las elecciones locales, los miembros del COTECORA usaron distintas ideas para construir las estratificaciones:

- Estratificar por municipio, en este caso, se puede ver que uniendo municipios se obtienen los distritos federales.
- Intersección entre distritos federales y alguna otra variable. En este caso, uniendo todas las intersecciones de un mismo distrito se obtiene el distrito completo.
- Estratificar usando distritos locales siempre y cuando uniendo varios distritos locales se obtengas distritos federales.

El manejo de las estratificaciones más finas en los estados tuvo dos objetivos:

1. Alcanzar mayor precisión para la elección local con menos muestra.
2. Lograr una mayor dispersión de la muestra.

Otra diferencia fundamental es que los tamaños de muestra para las elecciones locales implicaban tamaños de muestra significativamente más grandes por distrito federal a los buscados para la elección federal. Esto se daría únicamente en los estados en los que hubo elección de gobernador. Por ejemplo, en algún estado en donde hubo elección a gobernador en un estrato federal, la muestra, una vez acumulando la muestra para los estratos más finos, pudo ser de 100 casillas. Esto debido a que el diseño en ese estado fue planeado de esta manera. En cambio, en la elección Federal se buscan solamente 20 casillas por distrito federal. En este caso, la muestra para la elección federal será una sub-muestra de la elección local.

Bajo las consideraciones descritas anteriormente, es posible demostrar analíticamente que las probabilidades de selección de casillas para la elección federal son las mismas que se obtendrían asumiendo un diseño estratificado por distrito federal directamente. Por lo tanto, estos detalles no representan ningún problema técnico.

Con el objetivo de que exista un **compulsado** de metodologías, se realizaron dos estimaciones de la conformación de la cámara de diputados, una realizada por el equipo del Dr. Luis Enrique Nieto y el Maestro Carlos Pérez siguiendo una perspectiva Bayesiana de la inferencia estadística y la otra realizada por el equipo del Dr. Carlos E. Rodríguez y el Act. Edgar Alarcón, vía inferencia clásica usando las ideas básicas de re-muestreo Bootstrap (Sección 4.3). Ambas estimaciones se consolidaron en una sola estimación, mediante la unión de los correspondientes intervalos de estimación. Por lo que se obtuvo una sola estimación de la conformación de la cámara de diputados que cubrirá la conformación real con al menos un 95 % de confianza. Se siguió el mismo procedimiento para consolidar las estimaciones para la participación (5.3) y los porcentajes con respecto a la VVE.

5.6.1. Selección de la muestra días antes de la Jornada Electoral

La selección de la muestra se realizó en vivo y fue transmitida por diversas redes sociales el día 4 de junio del 2021. El vídeo de este proceso se puede consultar dando clic aquí[4] y el desarrollo computacional del mismo fue realizado en código de R y puede ser consultado en el GitHub del autor dando clic aquí[5].

[4] https://www.youtube.com/watch?v=SbySl89rwLA

[5] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/5.%20Muestreo%20Aplicado%20al%20CR/5.6.1

La aplicación que fue desarrollada para este proceso fue explicada durante este evento en donde se muestra y explica el uso de la misma con detalle. En la Figura 5.3 se ve un esquema de cómo la aplicación es utilizada para generar la muestra y en la Tabla 5.5 se explican las marcas de esta figura.

En resumen, este proceso consistía en la participación de 3 grupos conformados por algunos de los presentes con el objetivo de ingresar una semilla elegida y con ésta de manera pseudo-aleatoria generar la muestra. Este proceso fue realizado ante notario público para su validez formal. Uno a uno, cada uno de estos 3 grupos ingresaban parte de la semilla con la cual finalmente se seleccionaba la muestra de las casillas que serían utilizadas para el Conteo Rápido.

Figura 5.3: Diagrama del flujo operativo de la aplicación en *Shiny* que fue creada para seleccionar la muestra el día 4 de Junio de 2021 por el *Instituto Nacional Electoral* (INE).

Marca	Funcionamiento
1	**Campo de llenado para la semilla.**
2	**Rectificación de la semilla.**
3	**Ingresar la semilla.**
4	**Borrar para reescribir la semilla.**
5	**Continuar con el siguiente grupo.**
6	**Cancelar y repetir el paso.**
7	**Comenzar con la generación de la muestra.**
8	**Cancelar y repetir el paso.**

Tabla 5.5: Descripción de las marcas en la aplicación *Shiny* para generar la muestra en la Figura 5.3. La muestra generada se guarda en una carpeta nueva.

Capítulo 6

Imputación

Debido al modo de operar del Conteo Rápido, en donde las casillas se van presentando en tiempo real y a una velocidad humana que depende de la presión que se le está ejerciendo al *Capacitador Asistente Electoral* (CAE), al comienzo de este proceso se cuenta con muy poca información como para poder realizar las estimaciones de la Cámara de Diputados. Debido a esto, es necesario enfrentarse a un problema de *datos faltantes* y es debido a esto que la implementación de metodologías tales como la *Imputación Múltiple* pueden ser útiles para compensar la falta de información que se tiene principalmente al inicio de este ejercicio estadístico.

6.1. Ideas básicas

Esta sección tiene como principal objetivo ilustrar las ideas más relevantes del porqué y cómo se utilizan los algoritmos de imputación, ya que es una de las herramientas que fueron utilizadas para lograr efectuar la estimación de la conformación de la Cámara de Diputados con base en el tamaño de muestra que fuese arribando dada por los CAEs.

6.1.1. Datos faltantes y la no-respuesta

En general, toda encuesta busca recolectar la mayor cantidad de información de interés posible del universo de estudio en un momento determinado. Por ejemplo, el censo de Población y Vivienda por el *Instituto Nacional de Estadística y Geografía* (INEGI) se realiza con el objetivo de ser la fuente de información básica más completa para conocer la realidad demográfica y social del país. Sin embargo, debido a diversas circunstancias a las que se enfrenta el proceso de recolección de datos, muchas veces la información no llega completa. Un ejemplo dentro de este mismo contexto; si una vivienda no es accesible ya sea por razones sociales o geográficas, entonces no podría ser censada. De tal manera que existen casos donde la información se presenta de esta forma dando pie a datos faltantes.

6.1.1.1. ¿Qué es la no-respuesta?

El fenómeno de *no-respuesta*, usualmente denotado como *Not Available* (`NA`) computacionalmente, o bien, simplemente como una celda vacía en la matriz de datos; ocurre

en muchos censos y encuestas, y se da cuando las unidades de las cuales se obtiene la información no proporcionan algún dato solicitado, por ejemplo omitiendo una pregunta. El problema creado por la no-respuesta subyace en que valores con los que se esperaba contar finalmente resultan en observaciones o datos faltantes. Estos valores faltantes no solamente hacen que las estimaciones sean menos eficientes al reducir el tamaño de la base de datos, sino que también los métodos estadísticos estándares asumen que la información está completa y por lo tanto no pueden ser utilizados al momento para analizar la información. Más aún, pueden existir sesgos debido a que las unidades que suelen responder las encuestas seguido difieren sistemáticamente con los que no responden; por ejemplo, al preguntar el salario promedio de un trabajador. El sesgo puede resultar difícil de eliminar pues las razones de no-respuesta son usualmente desconocidas.

De acuerdo con [12], se puede extender la definición de la no-respuesta incluyendo a aquellas situaciones en las que los datos faltantes se dan del proceso de información dado por las unidades, mas que por el no contestar o proveer la información. Esto es por ejemplo, editando procesos que pudiesen eliminar respuestas que parecen imposibles tales como que una persona tenga una edad de 187 años. Otro caso surge a partir de recursos restringidos que pudiesen limitar la codificación de respuestas a una submuestra de unidades, en pocas palabras, toda la información que sí está disponible no fue posible de capturar por el personal. Una versión aún más extendida de no-respuesta incluye cualquier situación en la cual haya valores faltantes en una matriz de datos donde se registró información de manera ya sea incompleta o bien mal planteada, por ejemplo, solo preguntando el salario a personas que satisfacen ciertas características o bien preguntando por un año de nacimiento cuando se necesitaba información del mes de nacimiento.

En [13] establece que se pueden clasificar las fuentes de ausencia de respuesta de acuerdo con:

1. **El contenido de la encuesta**. Por ejemplo, una encuesta sobre drogas o de asuntos financieros o personales puede tener una gran cantidad de rechazos o información incompleta.

2. **Métodos de recolección de datos**. Por ejemplo, las encuestas por correo, fax o internet tienen bajas tasas de respuesta y las encuestas personales son las que tienen mayor tasa de respuesta.

3. **Características de quienes responden**. Por ejemplo, disponibilidad de las personas que responden. Así, una encuesta breve puede reducir el agobio de las personas que responden.

Asimismo, se menciona que la importancia de la no-respuesta depende de dos principales aspectos:

- **La magnitud o tamaño de la no-respuesta**, que al reducir el número de observaciones útiles para hacer mediciones incrementa el error muestral. Además, en el caso de un MAE, como la falta de respuesta no se produce por igual en todos los estratos, desequilibra la muestra y hace necesario reponderar estimaciones.

- **Diferencia de características** entre los que responden y los que no responden, lo que introduce un sesgo importante. El sesgo mantiene una relación creciente con el porcentaje de los que no responden y cuanto mayor sean las diferencias entre los que contestan y los que no.

Esto, sin mencionar la posible información falsa que los contribuyentes estén proporcionando. En numerosas ocasiones la información incorrecta puede llevar a errores más graves que la información no disponible. Tristemente esto último es algo que precisamente se tiene que estudiar a través de otras técnicas tales como el análisis de datos atípicos. Sin embargo, para el uso de esta tesis y por la naturaleza misma del ejercicio estadístico que es el Conteo Rápido, éste no debería ser un asunto por el cual tengamos que preocuparnos del todo.

6.1.1.2. Patrón de los datos faltantes

Uno de los puntos considerados en [13] en la no-respuesta parcial es el patrón de pérdida de los datos faltantes, ya que esto puede influir en la selección del método de imputación. Si la "base de datos" se interpreta como una matriz (de datos)/*dataset*, en donde las filas/tuplas son las unidades/individuos bajo observación y las columnas representan las variables/mediciones de interés, la elección del método de imputación debiera tener en cuenta el comportamiento de los datos faltantes, ya que el análisis visual puede permitir identificar patrones como los que se muestran en la Figura 6.1.

En el caso del Conteo Rápido, la manera en como se presentan los datos faltantes es similar al caso "e) Emparejamiento de archivos" esto debido a que la información se va presentando como remesas en archivos que tienen un nombre del estilo `REMESAS0400062105.txt` (todas las remesas se pueden consultar dando clic aquí[1]) y con un formato similar al mencionado en la Tabla 3.1 y la Tabla 3.2 pero adaptado para el Conteo Rápido del año 2021. Estos archivos, aunque no presentaban datos faltantes como tal, al tener relación con una muestra estratificada dada por lo establecido en la Subsección 5.6.1, existía inicialmente (y durante prácticamente todo el

[1] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/6.%20Imputaci%C3%B3n/REMESAS

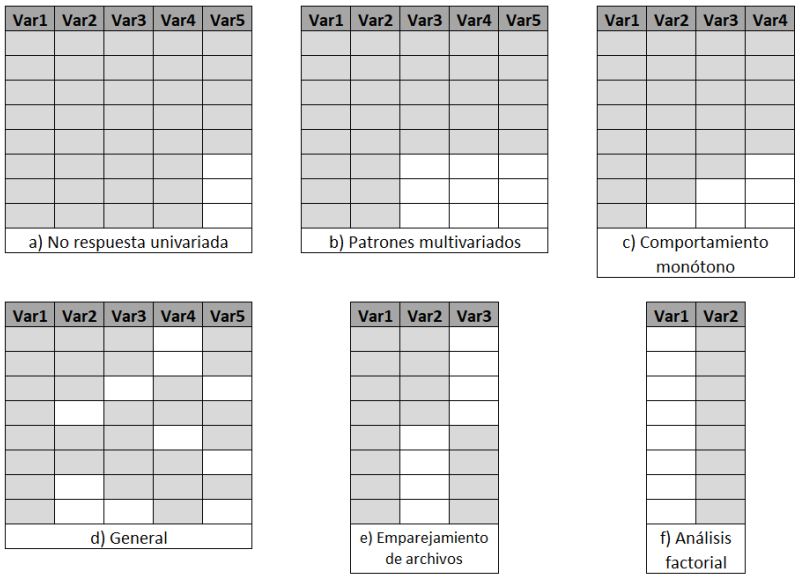


Figura 6.1: Patrones de datos faltantes. Las celdas en blanco representan la ausencia de información mientras que las sombreadas representan la presencia de información.

ejercicio estadístico) la presencia de datos faltantes en la espera de ser recolectados por los CAEs.

6.1.1.3. Modelos de datos faltantes

Con base en [14] existe una teoría en la que se pueden clasificar los problemas de datos faltantes en tres categorías. De acuerdo a esto, cada dato tiene cierta certidumbre de ser faltante. El proceso que gobierna estas probabilidades se le llama *mecanismo de datos faltantes*. El modelo de este proceso se le llama *modelos de datos faltantes*. Las categorías a las que se refiere anteriormente son las siguientes:

1. *Missing Completely At Random* (MCAR)

   La probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos. Es decir, la ausencia de la información no está originada por ninguna variable presente en la matriz de datos. Por ejemplo, en el caso de tener un estudio de las variables peso y edad, si existe el mismo porcentaje de datos faltantes a cualquier edad, sin considerar su peso o edad, entonces los datos son MCAR.

2. *Missing At Random* (MAR)

   La probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable pero es dependiente de los valores de otras variables del conjunto de datos. Es decir, la ausencia de datos está asociada a

variables presentes en la matriz de datos. Por ejemplo, en el caso de tener un estudio de las variables peso y sexo, si uno de los dos sexos tiene un porcentaje de datos faltantes mayor para la variable peso, entonces los datos son MAR.

3. *Not Missing At Random* (NMAR)

   La probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable. Por ejemplo, en el caso de tener un estudio de las variable peso y edad, si los sujetos con mayores valores de peso tienen un porcentaje de datos faltantes más elevados en esta variable para aquellos con la misma edad, entonces en este caso los datos son NMAR.

Los dos primeros mecanismos (MCAR y MAR) de datos faltantes mencionados se denominan también "ignorables", por cuanto producen efectos que se pueden ignorar si se controla adecuadamente por las variables que determinan la no respuesta. Por otro lado, para el tercero (NMAR) se le denomina en este sentido como "no ignorable".

6.1.1.4. Tratamiento de la no-respuesta

La *imputación* es una técnica relevante para trabajar aquellos problemas con datos deficientes o bien que presenten no-respuesta como se mencionó anteriormente. Existen diversas opciones para evitar la aparición de datos faltantes, desde la manera en cómo se plantea la encuesta hasta estos algoritmos matemáticos. El objetivo principal de la imputación es asignar un valor lo más coherente posible con base en el contexto de los datos a aquellos datos que se presentan como faltantes. Este tema se abordará con más detalle en la Subsección 6.1.2. Antes de abordar el tema de imputación, mencionaremos alternativas para tratar los datos faltantes mencionados en [13] y [14]:

- **Análisis con datos completos** *(Listwise)*

  Esta manera de proceder consiste en la eliminación de los registros / tuplas / renglones de la matriz de datos que presentan algún dato faltante y en **realizar el análisis estadístico únicamente con las observaciones que disponen de información completa para todas las variables**. Las ventajas de este enfoque son la facilidad de su implementación y la posibilidad de comparar los estadísticos univariantes. La mayor desventaja que tiene esta opción es que suele conllevar una importante pérdida de información principalmente cuando el número de variables es elevado, y puede generar sesgos en los resultados y las estimaciones, así como dificultar resultados asintóticos, los cuales son deseados para aplicar la teoría de inferencia estadística clásica, todo esto aunado a que desperdicia una importante cantidad de información que se conoce.

  Al eliminar la información, se asume que la submuestra excluida tiene las mismas características que los datos completos y que la falta de respuesta se generó de manera aleatoria, como en los casos MCAR y MAR (Subsubsección 6.1.1.3), lo cual en la mayoría de las situaciones prácticas no se cumple.

- **Análisis con datos disponibles** *(Pairwise deletion)*

  Una alternativa al análisis de datos completos consiste en **utilizar en el análisis de cada variable todos los datos que se disponga**. Una desventaja de este procedimiento es que utiliza distintos tamaños de muestra dependiendo de cada variable y que, en caso de utilizarse, no puede asegurar que la matriz de correlaciones entre las variables disponibles sea definida positiva. Con este método se obtienen buenos resultados únicamente en el caso de estar bajo un proceso de no respuesta tipo MCAR (Subsubsección 6.1.1.3). Cuando se le compara con el método *listwise*, esta opción tiene la ventaja de que hace uso de toda la información disponible pero al mezclar tamaños de muestra debilita su aplicación para algunas herramientas estadísticas, por lo que la elección de un método u otro es objeto de controversia.

- **Ponderación** (*Weighting*)

  Este método consiste en **utilizar un diseño de pesos sobre los casos completos para contrarrestar cualquier efecto producido por la falta de respuesta**. La esencia de todos los procedimientos ponderados es dar ciertos pesos (en forma de hiperparámetros[2]) a los individuos que respondieron de modo que representen a los que no respondieron. El objetivo de esta metodología es mejorar la precisión de las estimaciones y reducir el sesgo que introducen los que no respondieron, ya que el resultado final presupone que todos los sujetos contestaron. En general, este proceso requiere información auxiliar de los participantes y de los que no proporcionan información. Es posible aplicar distintos métodos para reponderar las observaciones que se mantienen en la muestra. Un problema es que la ponderación puede dar lugar a estimaciones con varianza muy grande.

6.1.2. Técnicas de Imputación

Se le denomina *imputación* al procedimiento que utiliza la información disponible en la muestra para asignar un valor a aquellas variables que tienen registros con el valor faltante, ya sea porque se carece de información o porque se detecta que algunos de los valores recolectados no corresponden con el comportamiento esperado. La razón principal por la cual se utiliza la imputación es obtener un conjunto de datos completo y consistente al cual se le puedan aplicar las técnicas estadísticas deseadas.

De acuerdo con [13], cuando un conjunto de datos se encuentra en la fase de imputación, se deben escoger cuidadosamente las variables objetivo y las auxiliares, los criterios de imputación y escoger el método preciso de imputación. Algunos de los criterios generales de calidad que se pueden considerar son:

[2]Los hiperparámetros son aquellos parámetros que son establecidos por el investigador con el fin de calibrar el modelo.

- **Mantener la distribución de la variable**. El objetivo es que la imputación llegue a producir una distribución de la variable aleatoria próxima a la distribución real.

- **Mantener las correlaciones entre variables**. Es deseable que las relaciones entre las variables no se vean alteradas por la imputación.

- **Consistencia**. Los valores imputados deben ser consistentes con las otras variables.

También, se pueden clasificar a grandes rasgos a los métodos de imputación como simples y múltiples, o bien, como determinísticos o aleatorios. Esto se comenta en términos generales más adelante.

6.1.2.1. Ventajas y desventajas de la imputación

La primer y más importante ventaja que tiene hacer esto es que de esta manera ya se pueden utilizar metodologías estadísticas que requieren el uso de datos completos, en contraste, se necesitarían de programas computacionales más especializados para manejar el problema de la no-respuesta. En particular, una ventaja que tiene la imputación simple es que, en muchos casos las imputaciones pueden ser creadas por el recolector de datos quien quizás tenga un mejor entendimiento de los procesos que ocasionan la no respuesta que cualquier otro usuario típico, y no simplemente utilizar procedimientos arbitrarios y rápidos tales como "llenar utilizando la media".

Por otro lado, el problema más evidente que tiene el uso de la imputación es que se están llenando valores que no son conocidos y, más aún, que se utilizan procedimientos estadísticos los cuales asumen que los datos son conocidos cuando realmente no lo son. En consecuencia, la inferencia estadística derivada de un conjunto de datos que fueron imputados puede causar sesgo en las conclusiones o bien no brindar resultados del todo confiables. Más aún, cuando la no-respuesta se da por causas que no se comprenden del todo, no hay manera de tomar en consideración la incertidumbre del por qué no se conocen los datos, lo cual puede llevar a no saber qué tipo de método de imputación es el apropiado.

6.1.2.2. Imputación Simple

La imputación simple **consiste en asignar un valor a cada valor faltante basándose en el valor de la propia variable o de otras variables, generando una base de datos completa**. De acuerdo con [12], es probablemente el método más común para tratar con una no-respuesta en la práctica de encuestas. A continuación se mencionan las ideas principales de algunas técnicas de imputación simple.

6.1.2.2.1 Imputación por Media

Con base en [13], este método es posiblemente uno de los procedimientos de imputación más antiguo y de acuerdo con [14] es una manera rápida de tratar a los datos faltantes simplemente tomando la media de la variable. La imputación por media tiene dos variantes:

- **Imputación por media no condicional**

  Consiste en estimar la media de los valores observados; *i.e.*, si y_{ij} es el valor de la variable-columna Y_j para la unidad/el renglón i, el método de imputación por medias incondicional trata de estimar los valores faltantes y_{ij} por $\overline{Y}_j$ que es la meda **de los valores observados** de Y_j. Esto significa que se tiene la filosofía de *Pairwise deletion* vista en la Subsubsección 6.1.1.4.

  En su aplicación se asume que los datos faltantes siguen un patrón MCAR (Subsubsección 6.1.1.3). Este procedimiento preserva el valor medio de la variable pero las estadísticas que definen la forma de la distribución (varianza, percentiles, sesgo, etc.) pueden verse afectadas, de la misma forma que también se distorcionan las relaciones entre las variables.

- **Imputación por media condicional**

  Imputa medias condicionadas a valores observados. Un método común consiste en agrupar los valores observados y no observados en clases e imputar los valores faltantes por la media de los valores en la misma clase.

6.1.2.2.2 Imputación Deductiva

Este es un método de imputación determinístico que se aplica en situaciones en que las respuestas que faltan (`NA`'s) se pueden deducir del resto de la información proveniente del conjunto de datos, *i.e.*, los valores se asignan mediante relaciones lógicas entre las variables. Una imputción determinística tiene, en términos de código de  , el siguiente formato

```
if(condición){
 acción
}
```

Un ejemplo bastante sencillo sería que, dada una matriz de datos tal que se cuenten con registros de individuos como su nombre y su sexo, si existe alguno que no haya proporcionado su sexo pero tiene un nombre femenino, podríamos imputar la variable faltante del sexo como femenino. Claro, esto como cualquier otro método tiene su margen de error, ya que, en este mismo ejemplo, podría existir un individuo que se llame, por ejemplo, "Andrea" y este nombre en México es generalmente asociado al sexo femenino, pero en Italia es un nombre masculino, por lo que existe la posibilidad de cometer este error de esta manera y de hecho, podría nutrirse más un modelo de

imputación por esta metodología que contemple este caso.

La programación en R de este último caso podría hacerse asumiendo que existe un objeto tipo `character`, digamos `nombres_femeninos`, que contenga todos los nombres femeninos más comunes, entonces una idea de código en este lenguaje de programación sería lo siguiente:

```
if(nombre %in% nombres_femeninos){
    sexo <- "Femenino"
}
```

6.1.2.2.3 Imputación *Cold Deck*

Con este procedimiento los valores faltantes (`NA`'s) se asignan a partir de una encuesta anterior o de otras informaciones, como datos históricos. La desventaja principal de este método es que la calidad de los resultados dependerá de la calidad de la información externa disponible. [13] afirma que a partir de este método se originó el procedimiento *Hot Deck*.

6.1.2.2.4 Imputación *Hot Deck*

De acuerdo con [14] la expresión "*Hot Deck*" se refiere literalmente a un paquete de control de tarjetas de computadora que contienen los datos de los casos que están en cierto sentido cerca. Históricamente, se remonta a cuando los datos eran almacenados en tarjetas perforadas y se indicaba si los donantes de información (en este caso haciendo referencia a los datos completos) provienen del mismo conjunto de datos que los destinatarios (en este caso los datos que se desean imputar). Se decía que estas tarjetas estaban "calientes" porque se estaban procesando al momento.

Con base en lo dicho por [13], la metodología *Hot Deck* consiste en un proceso de duplicación. Cuando falta un valor, se duplica un valor ya existente en la muestra para reemplazarlo. Su principal propósito es reducir el sesgo debido a la no-respuesta. Existen algunas variantes de este método:

- Imputación aleatoria *Hot Deck* (Imputación *Hot Deck* por MAS)

  Esta variante de la metodología consiste en asignar aleatoriamente un valor recogido en la muestra de la variable a imputar. Conserva la distribución de los individuos que sí responden pero no considera si es factible la imputación ni la correlación con otras variables. Es un método estocástico.

- Imputación *Hot Deck* por grupos

  En la mayoría de los casos, la metodología *Hot Deck* lleva consigo un proceso de clasificación asociado. Lo que se busca es que las unidades de la muestra

estén clasificadas en una partición, es decir, todas las unidades se encuentran en grupos disjuntos, esto de tal forma que las unidades sean lo más homogéneas posibles dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo, similar a un proceso de estratificación. De esta manera, el supuesto que se está utilizando es que dentro de cada grupo de clasificación la no-respuesta sigue la misma distribución que los que responden. Las variables de clasificación deben estar correlacionadas con los valores faltantes y con los valores de los que contestan. Si esto no se mantiene, el procedimiento *Hot Deck* puede llevar a resultados erróneos.

Este procedimiento entonces, consiste simplemente en imputar con un valor recogido de la muestra perteneciente al grupo. Al igual que el anterior, este es un método estocástico.

- Imputación *Hot Deck* secuencial

  Está técnica es **utilizada cuando la muestra tiene algún tipo de orden dentro de cada grupo de clasificación**. Cada valor faltante (`NA`) se reemplaza por un registro que sí cuente con dicho valor perteneciente al mismo grupo e inmediatamente anterior a él. Este es un método de imputación determinístico. Si el primer registro tiene un dato faltante, este es reemplazado por un valor inicial que puede obtenerse de la información externa. Las desventajas de este método son:

  - Cuando es necesario imputar muchos registros se tiende a emplear el mismo valor, llevando a una pérdida de precisión o un sesgo de las estimaciones.
  - Resulta complejo valorar la precisión de las estimaciones.

- Imputación *Hot Deck* por vecino más cercano

  Es un procedimiento no-paramétrico basado en la suposición de que los individuos cercanos en un mismo espacio tienen características similares. Este es un método de imputación determinístico y opera bajo ideas similares al análisis de *clusters* dentro de la teoría del análisis multivariado. Para aplicar este método se requiere definir una medida de distancia y en tal caso, imputar con el vecino más cercano, *i.e.*, con la observación cuya distancia sea menor a la que tiene una observación faltante.

  Supongamos que $x_i = (x_{i,1}, \ldots, x_{i,k})^T$ es el vector renglón i-ésimo de una matriz de datos con k-variables, el cual representa al i-ésimo individuo/tupla/registro y que presenta uno o más datos faltantes (`NA`'s). Entonces con la información disponible, cuyo conjunto de índices denotaremos como $\mathcal{I}$, se pueden definir diversas distancias $d(i, j)$, para imputar con otra x_j cuyas entradas en $\mathcal{I}$ son también conocidas y no presentan el dato faltante en la variable de interés. Existe una gran diversidad de distancias que se pueden utilizar para realizar esto, tales como:

- En general la norma-p:

$$d(i,j) = \left(\sum_{t \in \mathcal{I}} |x_{i,t} - x_{j,t}|^p \right)^p,$$

que cuando:

  - $p = 1$, se le conoce como distancia *Manhattan*.
  - $p = 2$, se le conoce como distancia *Euclidiana*.
  - $p \to \infty$, se le conoce como distancia *Chebyshev*.[3]

- Distancia de *Mahalanobis*:

$$d(i,j) = (x_i - x_j)^T S^{-1} (x_i - X_j),$$

donde S es la matriz de varianzas y covarianzas de los datos.

- Si se cuenta con algún registro en $\mathcal{I}$ que separe por grupos también es válido definir:

$$d(i,j) = \begin{cases} 0 & \text{si } x_i \text{ pertenece al mismo grupo que } x_j, \\ 1 & \text{si } x_i \text{ NO pertenece al mismo grupo que } x_j. \end{cases}$$

Y para cualquiera de estos casos, si se da un empate de distancia, imputar por alguna otra técnica utilizando las observaciones en cuestión.

Un posible peligro al usar el método *Hot Deck* es la duplicación del mismo valor en múltiples ocasiones. Esto ocurre cuando dentro de los grupos de clasificación existen muchos datos faltantes (`NA`'s) y pocos valores disponibles. Este método se recomienda utilizar cuando se trabaja con tamaños de muestra grandes para así disponer de valores que reemplacen a las unidades/tuplas faltantes.

6.1.2.2.5 Imputación por Regresión

Para este método utiliza una de las metodologías estadísticas más socorridas e implementadas para algoritmos de *machine learning* como parte del aprendizaje supervisado, para esto, se emplean modelos de regresión para imputar información en la variable respuesta (usualmente denotada como Y), a partir de las covariables (usualmente denotadas como $X_1, \ldots, X_k$) y que buscan estar correlacionadas con la respuesta. Este procedimiento consiste en eliminar las observaciones con datos incompletos y ajustar una regresión para predecir los valores faltantes con base en las covariables.

[3] En el caso de la distancia *Chebyshev* se puede demostrar que $d(i,j) = \text{máx}_t |x_{i,t} - x_{j,t}|$.

Supongamos que la forma de una matriz de datos viene dada de tal manera que Y sea el vector columna asociado a la variable que se desea imputar, X será la matriz de covariables con las que se busca predecir el valor de la variable a imputar. Adicionalmente, asumamos, sin pérdida de generalidad que la dimensión de la matriz de datos es de $(n+r)\times(k+1)$, de tal manera que hay $n+r$ individuos y $k+1$ variables, donde:

- Los primeros n individuos NO tienen datos faltantes, mientras que los restantes r los tienen únicamente en la variable Y.
- La matriz de datos cuenta con k covariables y una variable a imputar.

De tal manera que los últimos r individuos serán imputados con la predicción dada por el modelo de regresión utilizando a los primeros n individuos restado de ajustar, en general el modelo lineal generalizado:

$$g\left(\mathbb{E}[Y]\right) = \underline{X}\beta, \qquad Y \sim F. \tag{6.1}$$

Donde g es la función liga (*link*), Y de forma general puede seguir una distribución F y $\underline{X} = (X_1, \ldots, X_k)$ denota el vector de covariables asociadas al individuo en cuestión. Dependiendo de cómo sea la variable Y, su modelo distribucional y la función g, se obtiene un modelo de regresión en particular. Algunos de los más famosos son la regresión normal, logística y *probit*. Aunque en general puede ser cualquiera como regresión Poisson, Gamma, Binomial, etc.

Entonces, el procedimiento de imputación bajo esta metodología consiste precisamente en algo similar a cuando en *machine learning* se cuenta con un conjunto de prueba (*test*) y otro de entrenamiento (*training*). Con el conjunto de entrenamiento, que son los datos completos, se ajusta un modelo y con el de prueba, que son los datos que tienen faltante la respuesta, se realizan las predicciones que serán finalmente la imputación. A continuación describimos a grandes rasgos **algunos** casos clásicos de qué se puede elegir dependiendo del comportamiento de la variable a imputar (Y) recordando la Ecuación 6.1.

- **Si la variable a imputar es continua**

  Si te toma la función liga g como la identidad y asumimos una familia distribucional Normal para la variable a imputar Y, entonces se tiene el modelo de regresión lineal múltiple tradicional:

$$\mathbb{E}[Y] = \underline{X}\beta, \qquad Y \sim Normal. \tag{6.2}$$

  De donde podemos considerar dos variantes para realizar la imputación.

- **Imputación mediante regresión determinística**
  En este caso, se procede como un modelo de regresión múltiple tradicional utilizando la Ecuación 6.2 con los datos completos. De tal manera que los valores faltantes sean imputados usando la modelo ajustado con los datos con los que no son faltantes tomando así la imputación como:

$$\hat{y}_i = \underline{X}_i \beta. \tag{6.3}$$

  Donde $\hat{y}_i$ representa el valor imputado de la variable Y en la i-ésima entrada faltante y X_i son sus covariables correspondientes (las cuales se asumen como completas u observadas).
- **Imputación mediante regresión estocástica**
  Este caso es similar a lo que se hace en la parte determinística, solo que en la Ecuación 6.3 se incorpora una cantidad aleatoria a la predicción. De tal manera que la imputación se realizará mediante:

$$\hat{y}_i = \underline{X}_i \beta + z_i.$$

  Donde $z_i \sim N(0, \tilde{\sigma}_Y^2)$ con $\tilde{\sigma}_Y^2$ la varianza dada por las observaciones completas de Y.

- **Si la variable a imputar es binaria**

  Para este caso se considera una función liga $g(p) = logit(p) = ln\left(\frac{p}{1-p}\right)$ y asumimos una familia distribucional Bernoulli para la variable a imputar Y, entonces se tiene el modelo de regresión logística:

$$logit\left(\mathbb{E}[Y]\right)) = \underline{X}\beta, \qquad Y \sim Bernoulli. \tag{6.4}$$

  En este caso el modelo de imputación obtiene una probabilidad asociada al i-ésimo registro (p_i) de que el valor faltante sea $\hat{y}_i = 1$ y vendrá dada por:

$$p_i = logit^{-1}(\underline{X}_i \beta) = \frac{1}{1 + exp(\underline{X}_i \beta)}.$$

  Donde las entidades matemáticas involucradas son las estimadas por el modelo con datos completos de la Ecuación 6.4. Con base en las p_i se pueden imputar los valores para $\hat{y}_i \in \{0, 1\}$.

- **Si la variable a imputar es mixta**

  Una opción para realizar una imputación de una variable mixta es, por ejemplo, asumiendo que la variable a imputar (Y) tome el valor de cero con probabilidad positiva y algún otro real con una distribución continua, entonces la imputación se puede hacer en dos pasos:

1. Se imputa si vale cero según el **modelo logístico** anterior.
2. Si resulta que hay que imputar un valor diferente de cero, se hace un modelo lineal generalizado para variables continuas según sea el supuesto de la forma distribucional Y.

Todo esto ajustado para utilizar los datos que cuenten con la información completa.

6.1.2.2.6 Imputación *Predictive Mean Matching*

En [14] se menciona una técnica importante de imputación conocida como *Predictive Mean Matching* (`pmm`), misma que es un ejemplo del método *Hot Deck* (Párrafo 6.1.2.2.4), donde los valores son imputados utilizando valores de los casos completos con los que se hace una coincidencia por grupos (*match*) con respecto a una métrica, que es esta variante en particular.

Este método utiliza un algoritmo conocido como *Gibbs Sampling* que funciona para simular de la distribución final de un parámetro, por lo que este proceso utiliza técnicas de estadística Bayesiana. En este caso, este algoritmo se utiliza para que, dadas las variables de las observaciones completas, se ajuste una imputación sobre los datos faltantes por regresión (Párrafo 6.1.2.2.5) actualizando los parámetros del modelo y así lograr las imputaciones de la variable objetivo.

Por lo tanto, procedimiento de este método consiste, a grandes rasgos, en dos etapas:

1. **Matching**. Encontrar observaciones completas que, dado un nivel de tolerancia, se encuentren "cerca", en términos *Hot Deck* por vecino más cercano, de la observación cuya variable se desea imputar. Estas observaciones a su vez formarán un grupo de tamaño d (*donors*) que se toma por MAS con lo que se hace el siguiente paso.
2. **Predictive Mean**. Utilizando el grupo de observaciones formado (*donors*), se realiza una imputación y ésta puede variar dependiendo del tipo de variable objetivo, si es continua, discreta, etc. Por ejemplo, la imputación se puede hacer por regresión en el caso de que sea continua.

La idea es entonces imputar los datos faltantes utilizando los datos completos que potencialmente son similares a ellos. Dejando a un lado las observaciones completas que no parecen modelar correctamente la observación a imputar (Figura 6.2).

Este método en general prueba ser más robusto que utilizar, por ejemplo, datos que NO cumplen con tener varianza constante, es decir que son *heteroscedásticos* o bien que NO son *homoscedásticos*. En [15] se expone un ejemplo de la implementación de esta metodología lo cual brinda los resultados observados en la Figura 6.3 donde se aprecia la consistencia que tiene el uso de esta metodología.

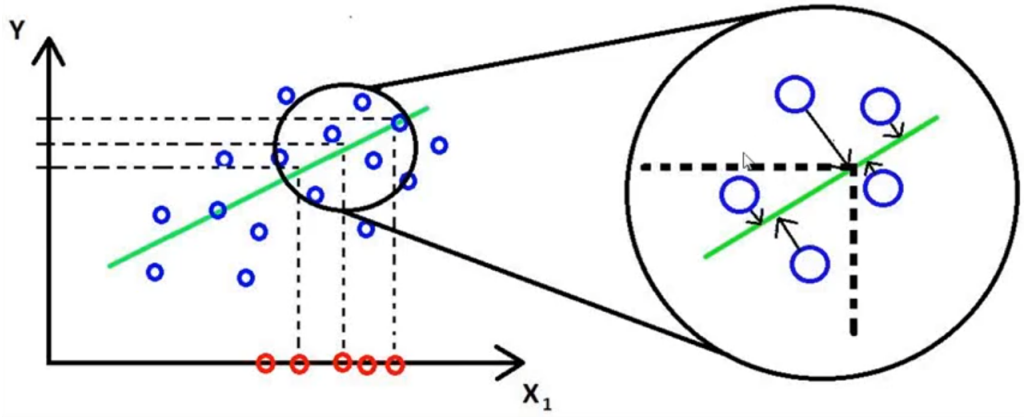


Figura 6.2: Ilustración de la idea básica de lo que hace el método *Predictive Mean Matching* (`pmm`). Los datos azules cuentan con la variable objetivo (Y) como conocida y las rojas la tienen como dato faltante.

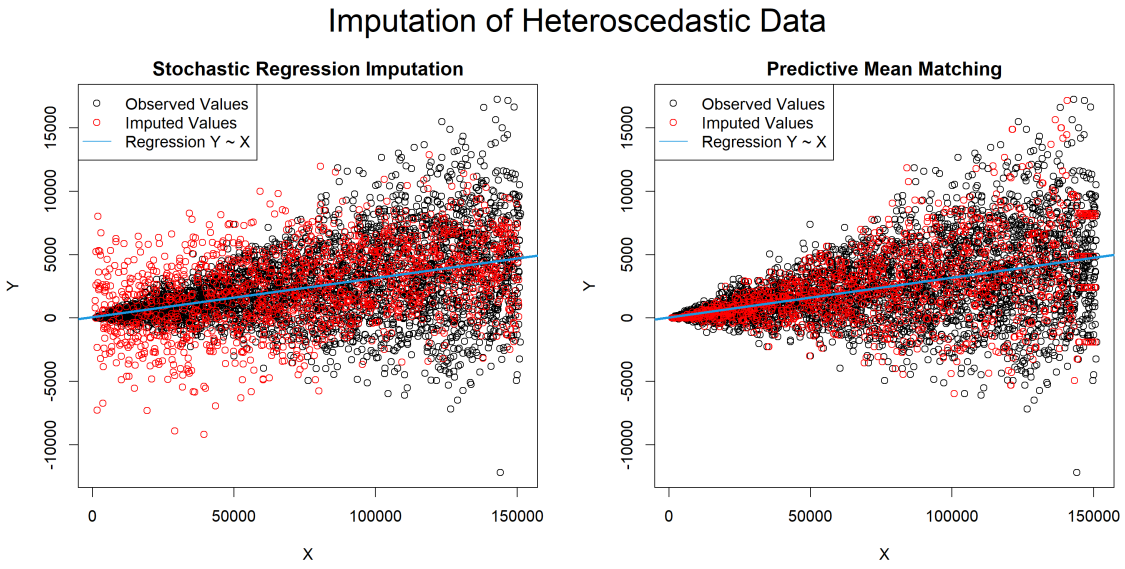


Figura 6.3: Gráfico realizado por [15] mostrando la efectividad de la metodología *Predictive Mean Matching* (`pmm`) (derecha) frente a la imputación por regresión estocásica (izquierda).

De tal manera que utilizando un modelo apropiado y asumiendo que los datos faltantes son *Missing At Random* (MAR), las estimaciones derivadas de este método de imputación mantienen una mayor consistencia.

6.1.2.2.7 Otros métodos de imputación

Tal y como se verá en la Sección 6.2, la librería *Multivariate Imputation by Chained Equations* (`mice`) del paquete estadístico R tiene implementados diversos métodos de imputación; si se desea conocer otros de los aquí mencionados, acudir a la sección antes citada. Un par de estos métodos de imputación los mencionaremos a continuación basado en lo dicho por [14].

Estos métodos están cimentados en técnicas de *Machine Learning* que involucran la implementación de árboles. Estos son métodos de aprendizaje supervisado que, como tales, constan en entrenar un algoritmo a partir de observaciones completas que aprendan del comportamiento que tiene la variable a imputar usando como referencia las covariables (*features*). Esto se hace con el objetivo de imputar con las observaciones que no tienen datos faltantes, similar a lo que se hace con la imputación por regresión (Párrafo 6.1.2.2.5). Además, los algoritmos relacionados con árboles tienen una efectividad bastante alta cuando se trata con variables discretas o categóricas (*classification*), aunque también pueden ser utilizados con variables continuas (*regression*). La manera en que éstos algoritmos imputan es análoga a los casos anteriores, se tomará como variable respuesta a la que se desea imputar y como explicativas a las demás que tengan datos completos y aplicando la metodología descrita.

- *Classification and regression trees* (`cart`). Este modelo busca covariables (*predictors*) basándose en puntos de corte que ayudan a dividir la muestra en submuestras homogéneas. El proceso de separación se repite a su vez en ambas sub-muestras, de tal manera que la serie de separaciones define un árbol binario (Figura 6.4). Asumiendo que la muestra es de tamaño m, esto se hace tantas veces como se desee de tal manera que al final se cuente con N grupos compuestos por n_i observaciones con $i \in \{1, \ldots, N\}$ tales que $m = \sum_{i=1}^{N} n_i$. En caso de que la variable objetivo sea discreta se la conoce a esta ramificación como árbol de clasificación (*classification tree*) y en caso de que sea continua como árbol de regresión (*regression tree*).

- *Random forest (imputations)* (`rf`). En general este método necesita de la implementación de árboles de decisión (*decision trees*) los cuales no son más que árboles de clasificación o de regresión. Es decir, esta es una generalización del método anterior, solo que tomado con distintas sub-muestras vía *bootstrap* como lo mencionaremos más adelante. Los árboles de decisión, aunque son bastante fáciles de construir, tienen la gran desventaja de que son poco flexibles. Esto significa que, a pesar de que son buenos con los datos con los que fueron creados, en general no funcionan del todo bien cuando se les ingresa nuevos datos. La ventaja que tienen los `rf` es que combinan la simplicidad y la flexibilidad, lo que resulta en una mejora para la exactitud del método.

  En pocas palabras, para realizar un `rf` se siguen los siguientes pasos:

  1. Crear un conjunto de datos por *bootstrap* (Párrafo 4.3.2.4.1) a partir de los datos originales (completos y con la variable objetivo).
  2. Con el conjunto de datos creado, construir un árbol de decisión pero únicamente utilizando un subconjunto de covariables en cada paso, el cual se elige al azar.
  3. Luego se repiten los pasos 1 y 2 creando en cada iteración un nuevo árbol de decisión, pero con la particularidad de que se está creando con individuos

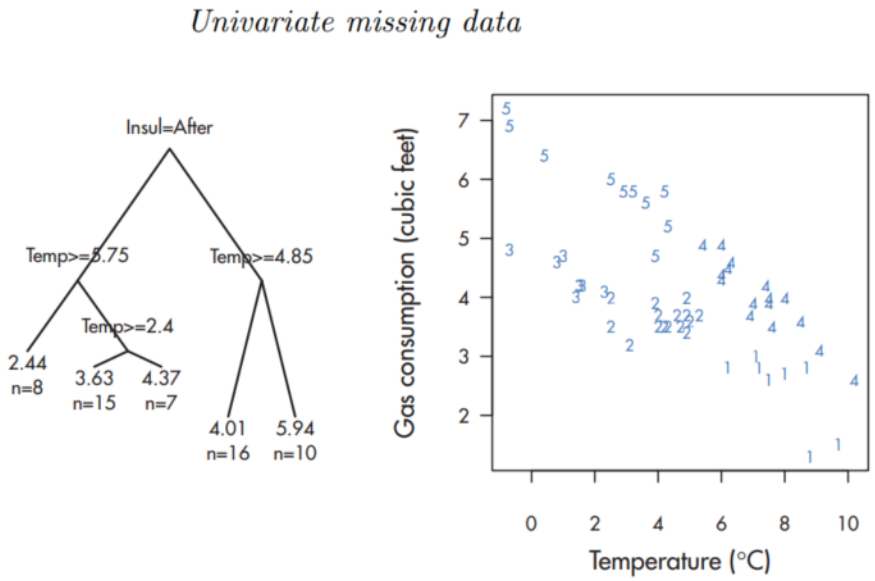


Figura 6.4: Gráfico realizado por [14] utilizando los datos `MASS::whiteside` de la paquetería estadística R. Esto es un árbol de regresión para predecir el consumo de gas. A la izquierda se muestra el árbol binario y a la derecha el gráfico que identifica a los individuos de los datos en los grupos resultantes de cada hoja de los datos. En este caso se está tomando $N = 5$ y $m = 56 = 8 + 15 + 7 + 16 + 10$ donde cada sumando representa una n_i.

obtenidos por *bootstrap* y las covariables son elegidas al azar en cada paso de construcción del árbol de decisión. Creando así un "bosque aleatorio". Lo que da una gran diversidad de árboles de decisión (Figura 6.5).

4. Finalmente, para imputar la variable objetivo, se toma al individuo que tiene a esta como `NA` y posteriormente se evalúa en cada uno de los árboles de decisión. Dependiendo de los resultados que brinden estos árboles de manera global, se elige entonces el valor de la variable de interés.

En resumen, esto significa que el método *Random forest (imputations)* (`rf`) no es más que una generalización del método *Classification and regression trees* (`cart`). En `rf` lo que se hace es crear diversos árboles de decisión (modelos) por el método `cart` pero con diversas sub-muestras tomadas vía *bootstrap*. Y finalmente el individuo a imputar se mete en todos y cada uno de estos árboles y, dependiendo si el problema es de clasificación (respuesta categórica) o de regresión (respuesta continua), se escoge un valor final para el individuo imputado generalmente tomando la respuesta más repetida para el problema de clasificación o el promedio de las respuestas en el caso de regresión.

6.1.2.3. Imputación Múltiple

De acuerdo con [13], la imputación múltiple **consiste en asignar a cada valor faltante varios valores (m), generando m conjuntos de datos completos**. En

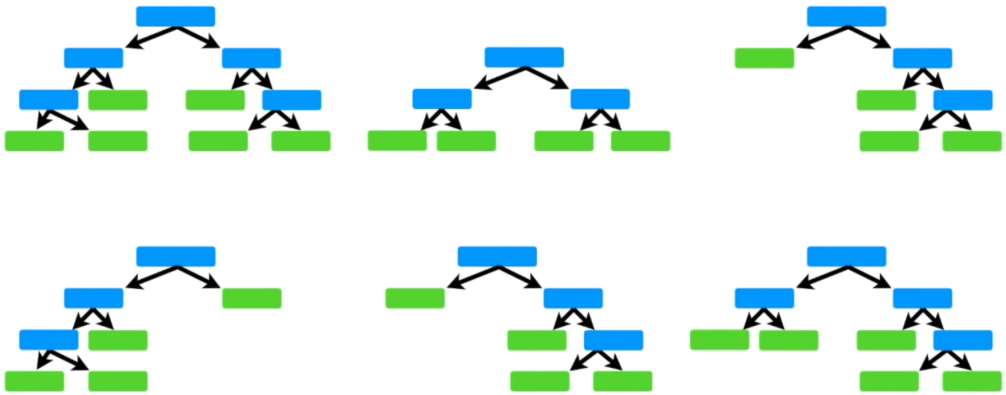


Figura 6.5: Ejemplo arbitrario de cómo se puede ver un *Random forest (imputations)* (`rf`) cuando se crean múltiples árboles de decisión obtenidos por *bootstrap* y covariables elegidas al azar.

cada conjunto de datos completo se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos.

En otras palabras, y a diferencia de los métodos anteriores, que imputan un valor único a cada dato desconocido. La imputación múltiple se basa en la imputación de más de un valor para cada valor ausente. Consiste en generar $m > 1$ valores aleatorios para cada valor `NA` de manera que se dispone de m conjuntos de datos completos. Luego, se realizan los análisis estadísticos usuales a partir de cada uno de los m conjuntos de datos generando m estimaciones. Finalmente, las distintas estimaciones son combinadas para producir una estimación con buenas propiedades estadísticas y con la posibilidad de estimar la varianza de los estimadores.

De tal manera que el método de imputación múltiple consta de 3 etapas:

1. **Imputación**. Cada valor perdido se reemplaza por un conjunto $m > 1$ valores generados por simulación, con lo que se crean m conjuntos de datos completos.

2. **Análisis**. Se aplica a cada una de ellas el método de análisis deseado.

3. **Combinación** (*Pool*). Los resultados obtenidos se combinan mediante reglas simples para producir una estimación global.

El objetivo de la imputación múltiple es hacer un uso eficiente de los datos que se han recogido, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no-respuesta introduce en la estimación de los parámetros. El proceso mencionado anteriormente se resume en la Figura 6.6 basándose en [14].

El número a elegir de conjuntos de datos completos (m) resulta ser un hiperparámetro de este algoritmo, *i.e.*, está dado por el usuario del mismo y generalmente se toma a

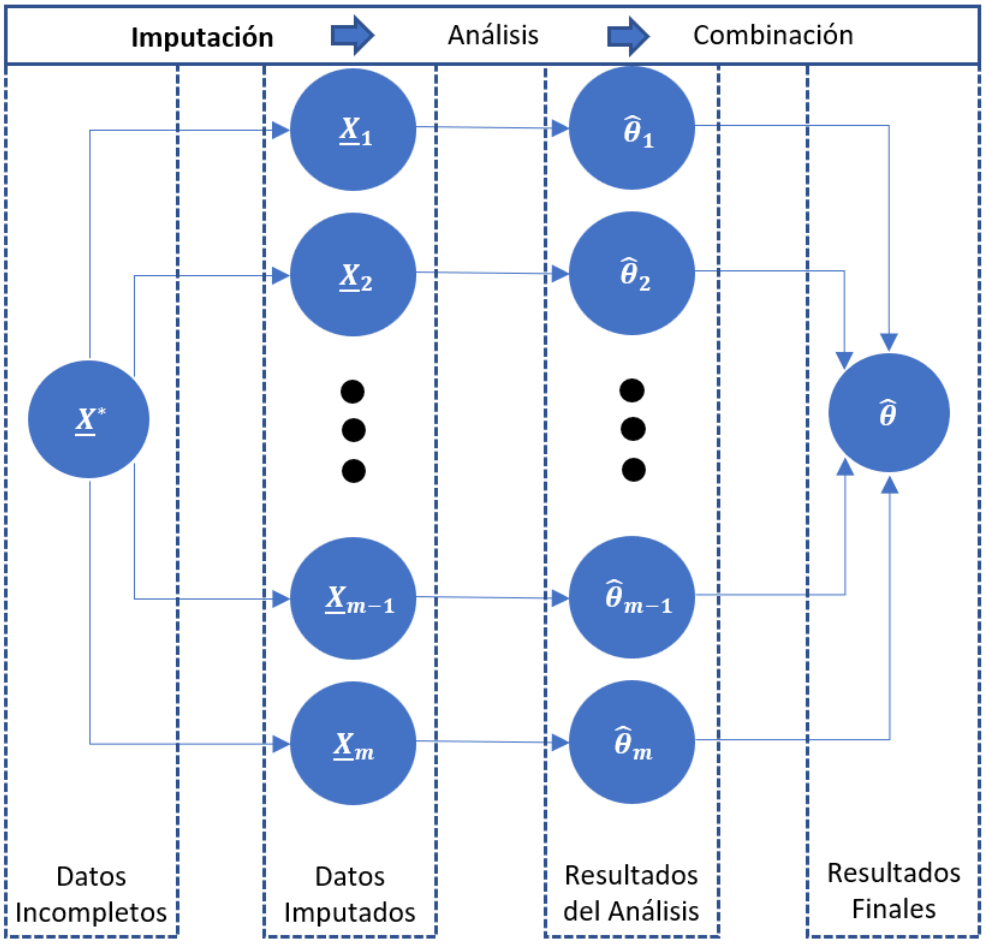


Figura 6.6: Representación visual del método de imputación múltiple por etapas y generando m conjuntos de datos completos. La notación $\underline{X}^*$ denota al conjunto de datos original (con `NA`'s), $\underline{X}_i$ denota el i-ésimo conjunto de datos completo (imputado), $\hat{\theta}_i$ el análisis realizado con su correspondiente $\underline{X}_i$ y $\hat{\theta}$ denota el análisis final dado por la combinación de las $\hat{\theta}_i$ todo esto para cada $i = 1, \ldots, m$.

consideración el porcentaje de información faltante y también del poder computacional con el que se cuente. En [12] se considera que la **mínima** m para proporcionar estimaciones válidas es, en general, $m = 3$. Cada una de las m estimaciones anteriores se pueden crear con una gran variedad de métodos, desde los más simples, como los de imputación por la media (Párrafo 6.1.2.2.1) hasta otros más complejos.

6.1.2.3.1 Ecuaciones de combinación para la imputación múltiple

De acuerdo con [12], para **combinar** las m estimaciones obtenidas se calcula la media de todas ellas. Partiendo de esto, y recordando la notación de la Figura 6.6, sean $\hat{\theta}_i$ y W_i con $i = 1, \ldots, m$ las estimaciones realizadas del parámetro de interés θ en cada conjunto de datos y las varianzas respectivas a cada estimación para un parámetro θ.

La **estimación combinada** es:

$$\overline{\theta}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \tag{6.5}$$

y la variabilidad asociada a esta estimación tiene dos componentes:

- La varianza dentro de cada imputación,

$$\overline{W}_m = \frac{1}{m}\sum_{i=1}^{m} W_i \tag{6.6}$$

- La varianza entre las imputaciones (recordando la Ecuación 6.5),

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \overline{\theta}_m\right)^2 \tag{6.7}$$

Por tanto, la variabilidad total asociada a la estimación $\overline{\theta}_m$ de la Ecuación 6.5, se obtiene de acuerdo a la Ecuación 6.8 utilizando las ecuaciones (6.6) y (6.7) como se muestra a continuación:

$$T_m = \overline{W}_m + \frac{m+1}{m}B_m \tag{6.8}$$

donde $\frac{m+1}{m}$ es el factor por corrección por ser m un número finito. Por lo tanto,

$$\hat{\gamma}_m = \frac{m+1}{m}B_m/T_m$$

es una estimación de la fracción de información sobre θ que se pierde por la falta de respuesta.

Si el parámetro $\theta \in \mathbb{R}$ (es un escalar), las estimaciones por intervalo y las pruebas de significancia siguen una distribución t de Student:

$$\frac{(\theta - \overline{\theta}_m)}{\sqrt{T_m}} \sim t_\nu, \tag{6.9}$$

donde los grados de libertad están dados por:

$$\nu = (m-1)\left(1 + \frac{\overline{W}_m}{B_m(m+1)}\right)^2.$$

En otro caso, cuando $\theta \in \mathbb{R}^k$, es decir que es un vector de k componentes, las pruebas de significancia para contrastar la hipótesis de nulidad del parámetro estimado θ deben ser realizadas a partir de las m estimaciones realizadas y no a partir de la estimación combinada.

6.1.2.3.2 Puntos a contemplar

De acuerdo con [16] en los puntos más importantes que debemos considerar para realizar una Imputación Múltiple son los siguientes:

- La imputación múltiple es un procedimiento basado en simulaciones. Su propósito no está en recrear valores faltantes individuales tan cercanos a los reales como sea posible, sino más bien manejar los datos faltantes de tal manera que los resultados de la inferencia estadística sean válidos.
- La imputación múltiple es efectiva si:
  1. El método de imputación es apropiado con respecto al mecanismo que presentan los datos faltantes (Subsubsección 6.1.1.2).
  2. El análisis con los datos completos es válido con ausencia de datos faltantes.
- Un número pequeño de imputaciones (m de 5 a 20) pueden ser suficientes cuando el porcentaje de datos faltantes es bajo. Si este porcentaje es más alto, se pudiese requerir de 100 o quizás más imputaciones. Siempre que sea viable hacerlo, [16] recomienda variar el número de imputaciones para observar el comportamiento de los resultados.
- Con un número pequeño de imputaciones, la distribución de referencia para la inferencia realizada de las imputaciones múltiples es una t de Student. Los grados de libertad dependen de m y de los cocientes de información faltante, y por lo mismo son diferentes para cada parámetro de interés.
- Con un número grande de imputaciones, la distribución de referencia para la inferencia realizada de las imputaciones múltiples es aproximadamente normal.
- Cuando el modelo de imputación es más restrictivo que el modelo de análisis, la inferencia por imputación múltiple puede ser inválida si los supuestos del modelo de imputación no son válidos. Por otra parte, cuando el modelo de análisis es más restrictivo que el modelo de imputación, los resultados por imputación múltiple serán válidos aunque hasta cierto punto conservadores si los supuestos del análisis son verdaderos. Si los supuestos del análisis son falsos, los resultados pueden ser sesgados.
- Ciertos conceptos como la verosimilitud o la devianza no tienen una interpretación clara dentro del marco de trabajo de la imputación múltiple. Por lo mismo, algunos procedimientos estadísticos (como la bondad de ajuste o pruebas por

cociente de verosimilitudes) que están basados en estos conceptos no son directamente aplicables a los resultados dados por esta metodología.

- Si los modelos de imputación y análisis son correctos entonces la imputación múltiple produce resultados consistentes y, en muestras grandes, insesgados bajo el supuesto de que los datos son MAR (Subsubsección 6.1.1.3).

6.1.2.4. Imputación simple vs. Imputación múltiple

Una de las grandes ventajas que tiene la imputación simple es que se trabaja con "bases de datos" (tablas/matrices de datos) completas, pero este método trata los valores imputados como si fueran los reales, por lo que se suele sobreestimar la precisión ya que no se tiene a consideración la volatilidad de las componentes entre las distintas imputaciones realizadas. De acuerdo con [13] existen tres **ventajas importantes de la imputación múltiple** respecto a la imputación simple:

1. **Incrementa la eficiencia de los estimadores** ya que resuelve el problema de que los errores estándares derivados de las metodologías en la Subsubsección 6.1.1.3 suelen ser muy pequeños [12].

2. **Obtiene inferencias válidas** simplemente mediante la **combinación** de las inferencias obtenidas en las matrices de datos completas.

3. **Permite estudiar la sensibilidad** directamente de las inferencias de varios modelos de no-respuesta usando los métodos de las matrices de datos completas repetidamente.

4. En [14] se menciona que la imputación múltiple resuelve el problema de tratar con la incertidumbre de la imputación aleatoria en sí misma.

Sin embargo, también existen **desventajas de la imputación múltiple**:

1. **Necesita un mayor esfuerzo** (computacional) para crearla, mayor tiempo para ejecutar el análisis y mayor espacio de almacenamiento para crear las matrices de datos imputadas. Esta desventaja no es del todo considerable cuando el número de simulaciones (m) es moderado.

2. **No produce una única respuesta**, el investigador deberá manejar múltiples bases de datos donde cada una de ellas tiene un valor posible para la observación faltante.

6.1.3. ¿Cómo seleccionar el método adecuado de imputación?

Seleccionar un método de imputación adecuado es una decisión tomada por el investigador que tendrá un gran impacto en los resultados, ya que para un conjunto de datos determinado, algunas técnicas de imputación podrían dar mejores aproximaciones a los valores verdaderos que otras. Como ya se ha mencionado anteriormente,

es deseable que el investigador tenga acceso al conocimiento de dónde provienen los datos para lograr una mejor comprensión de éstos y de esta manera seleccionar el método de imputación que considere más pertinente. Esta selección del método de imputación adecuado dependerá del tipo de datos, tamaño de los archivos, tipo de no-respuesta, patrón de datos faltantes, características específicas de la población, software disponible, distribuciones de frecuencias de cada variable, marginal, conjunta, etc. Incluso puede darse el caso en el que la técnica de imputación seleccionada sea adecuada para algunas variables pero para otras no y será decisión del investigador seleccionar el método que menos afecte a las estimaciones de las variables.

En [17] se plantea que "La técnica de imputación seleccionada debe superar las reglas de validación, cambiando lo menos posible los registros y manteniendo la frecuencia de la estructura de los datos". Teniendo en mente que la tarea de imputación varía dependiendo del tamaño del conjunto de datos, [18] resume los criterios a tomar en consideración para seleccionar un modelo de imputación adecuado:

- **La importancia de la variable a imputar**. Si la variable es de elevada importancia, es natural que se elija más cuidadosamente la técnica de imputación a aplicar.
- **Tipo de variable a imputar**. Si es continua o categórica (nominal u ordinal). Teniendo en cuenta para el primer grupo el intervalo para el cual está definida la variable y para el segundo las distintas categorías de la variable.
- **Parámetros que se desean estimar**. En el caso que solamente nos interese conocer el valor medio y el total (la suma), se pueden aplicar los métodos más sencillos. En el caso que se requiera la distribución de frecuencias de la variable, la varianza y asociaciones entre las variables, se deben emplear métodos más elaborados y analizar el conjunto de datos. Este problema se puede incrementar cuando hay una elevada tasa de no-respuesta.
- **Tasas de no respuesta**. No se debe abusar de los métodos de imputación y menos cuando se tiene una elevada tasa de no-respuesta de la cual no se conoce el mecanismo.
- **Información auxiliar disponible**. La imputación puede mejorar al emplear información auxiliar disponible. En el caso de no disponer de información auxiliar, una técnica recomendada a aplicar es la imputación aleatoria *Hot Deck* (Párrafo 6.1.2.2.4).

Estos puntos son aspectos que se deben tener a consideración para elegir un método de imputación que sea capaz de reproducir eficientemente un conjunto de datos completos, al cual se le pueda aplicar un análisis estadístico para datos completos. Dependiendo del tipo de variable que se esté considerando, [13] propone una serie de medidas para obtener un proceso de imputación adecuado, el cual debe idealmente:

1. Arrojar un valor imputado que sea lo más cercano posible al valor real.

2. Para variables numéricas o categóricas ordinales; arrojar una ordenación que relacione el valor imputado con el valor real o sea muy similar.

3. Preservar la distribución de los valores reales.

4. Producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales.

5. Conducir a valores imputados que sean plausibles.

6.2. Librería `mice` de R

6.2.1. ¿Para qué funciona?

La librería *Multivariate Imputation by Chained Equations* (`mice`) es un paquete perteneciente al lenguaje de programación estadístico R especializado en la imputación de datos, es decir, implementa computacionalmente métodos para tratar con los datos faltantes. Este paquete puede crear imputaciones múltiples (remplazamiento de valores) para datos faltantes multivariados. El algoritmo implementado en esta librería puede imputar mezclas de datos continuos, binarios, categóricos no-ordenados y categóricos ordenados. En adición, este paquete busca mantener consistencia entre imputaciones por medias y también se han implementado gráficos diagnóstico para inspeccionar la calidad de las imputaciones.

Para instalar la librería `mice` directamente de *The Comprehensive R Archive Network* (CRAN) con el siguiente fragmento de código:

```
install.packages("mice")
```

La versión más actualizada también puede ser instalada desde GitHub con el siguiente fragmento de código:

```
install.packages("devtools")
devtools::install_github(repo = "amices/mice")
```

Se puede acceder al GitHub oficial del autor de esta librería dando clic aquí[4].

Asimismo, si se desea conocer los métodos que tiene implementado la librería `mice` basta con ejecutar el siguiente código y buscar en el apartado con la leyenda "*Built-in univariate imputation*". Un par de estos se comentan en el Párrafo 6.1.2.2.7.

```
?mice::mice
```

[4]https://github.com/amices/mice

6.2.2. Ejemplos básicos

Para mostrar ejemplos de la librería `mice` en R estaremos utilizando una "base de datos" nativa en este lenguaje de programación. En particular se mostrarán ejemplos usando el objeto tipo `data.frame` llamado `airquality` de la librería nativa `datasets` en R. Este conjunto de datos almacena información sobre la calidad del aire medida diariamente en Nueva York, de mayo a septiembre de 1973. Un fragmento de este conjunto de datos lo podemos ver en la Tabla 6.1. Este conjunto de datos tiene 153 observaciones/tuplas/renglones y 6 variables/columnas con un total de 44 `NA`'s los cuales están distribuidos únicamente en las variables `Ozone` (37 `NA`'s) y `Solar.R` (7 `NA`'s). La descripción de las variables se muestra a continuación:

- `Ozone`: Ozono promedio en partes por miles de millones de las 13:00 a las 15:00 horas en *Roosevelt Island*.
- `Solar.R`: Radiación solar en Langleys en una banda de frecuencia 4,000-7,700 Angstroms de las 08:00 a las 12:00 horas en Central Park.
- `Wind`: Velocidad promedio del viento en millas por hora entre las 07:00 y 10:00 horas en el aeropuerto La Guardia.
- `Temp`: Temperatura diaria máxima en grados Fahrenheit en el aeropuerto La Guardia.
- `Month`: Mes en el que se tomó la media.
- `Day`: Día en el que se tomó la medida.

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6

Tabla 6.1: Fragmento del *dataset* `airquality` sin realizar ninguna imputación. Las celdas de color rojo con la leyenda "`NA`" representan datos faltantes.

Todos los ejemplos que se muestran a continuación pueden ser consultados de manera general en el GitHub del autor dando clic aquí[5]. Asimismo, las partes más importantes del código se mostrarán en este documento y en la sección correspondiente a la

[5]https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/6.%20Imputaci%C3%B3n/Ejemplos

metodología que se utiliza dentro de este apartado.

Adicionalmente y con el objetivo de mostrar este procedimiento de una forma lo más general posible, el conjunto de datos `airquality` será guardado en una variable auxiliar que permitirá hablar de forma más general de los resultados que veremos a continuación y adicionalmente se pueden invocar las librerías que pueden facilitar la ejecución de estas tareas. Esto se logra con el siguiente fragmento de código:

```
# Invocando las librerías a utilizar
library(mice)
library(dplyr)
# Los siguientes scripts están disponibles en el siguiente enlace:
# https://github.com/A1arcon/R_Actuarial/tree/main/_Edgar%20Package_
source("~/Actuaría/GitHub/R_Actuarial/_Edgar Package_/mis_funciones.R")
# Guardando los datos
datos <- airquality
```

6.2.2.1. Imputación simple

6.2.2.1.1 Imputación por media

A continuación, se estará ejemplificando el Párrafo 6.1.2.2.1 con un tratamiento de la no respuesta (Subsubsección 6.1.1.4) del tipo *Pairwise deletion*. En este sentido, simplemente se reemplazarán los valores de datos faltantes, como los vistos en el fragmento de datos de la Tabla 6.1, con la media de la variable que tiene los datos faltantes. Entonces, lo que se hace es calcular el promedio de los datos disponibles en las variables que presentan `NA`'s y reemplazarlos por su media. En este caso las medias de las variables `Ozone` y `Solar.R` son 42.13 y 185.93 respectivamente. Esto se puede puede lograr con el siguiente fragmento de código:

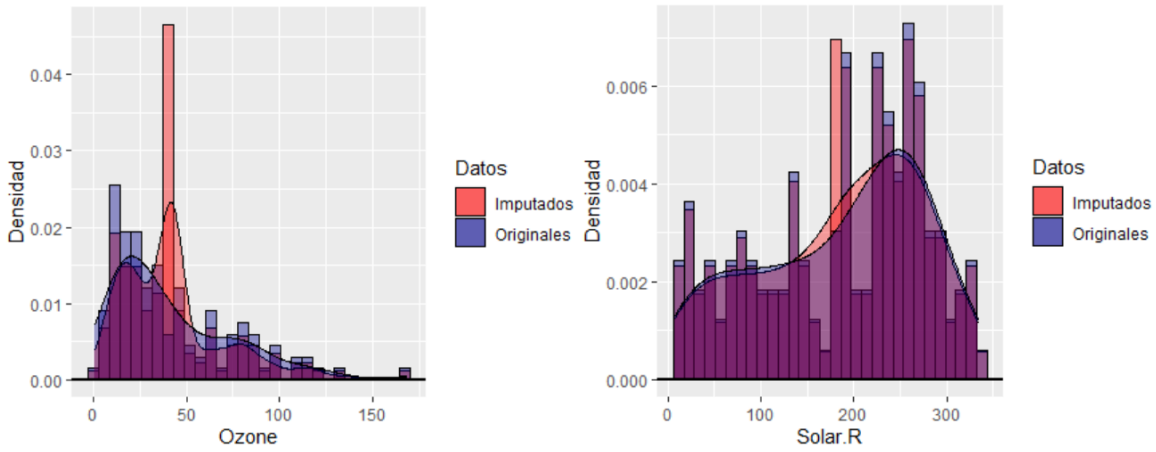
```
imp <- mice(datos, method = "mean", m = 1)
datos_imp <- imp %>%
             mice::complete(action="long") %>%
             dplyr::select(names(datos))
```

Lo cual produce una matriz de datos completos y cuyo fragmento podemos ver en la Tabla 6.2.

En la Figura 6.7 podemos ver los gráficos de la densidad suavizada y los histogramas de los datos originales (azul) y los datos imputados (rojo). Recordemos que la variable `Ozone` tiene 37 `NA`'s en contraste con la variable `Solar.R` que tiene únicamente 7 `NA`'s. Debido de esto último podemos ver la consecuencia de utilizar este método principalmente en la Figura 6.7a donde hay más datos imputados y por lo mismo la forma de la densidad se ve alterada (sesgo). Por otro lado, en la Figura 6.7b este

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
42.13	185.93	14.3	56	5	5
28	185.93	14.9	66	5	6

Tabla 6.2: Fragmento del *dataset* `airquality` habiendo realizado una imputación por la media. Las celdas verdes muestran la imputación realizada.



(a) Variable `Ozone`. (b) Variable `Solar.R`.

Figura 6.7: Densidad suavizada e Histograma de los datos originales (azul) vs. los imputados (rojo) de las variables de interés utilizando el método de imputación por media.

efecto no se ve tan pronunciado debido a la falta de datos faltantes.

En la Figura 6.8 vemos el comportamiento conjunto de estas variables, nuevamente, las observaciones que necesitaron una imputación se presentan de color rojo y las que no de color azul. Como cada variable fue imputada de manera independiente podemos ver que el comportamiento en conjunto de estas variables no presenta correlación. Este es otro de los efectos que produce este método y de los cuales se tiene que estar consiente.

6.2.2.1.2 Imputación por regresión

Ahora, se estará ejemplificando el Párrafo 6.1.2.2.5, en particular, el apartado de **imputación mediante regresión estocástica**. Debido al factor aleatorio que tiene esta metodología, será necesario imponer una semilla, misma que la función dada por

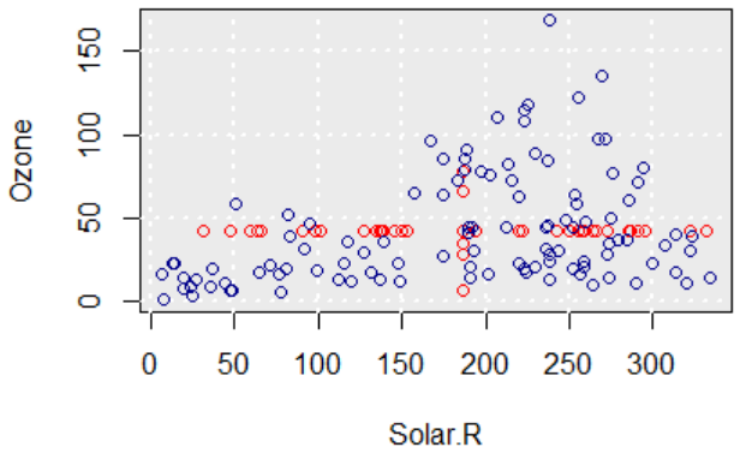


Figura 6.8: Gráfico de dispersión de los datos originales (azul) vs. los imputados (rojo) de las variables de interés utilizando el método de imputación por media.

la librería `mice` ya tiene implementada. La manera en cómo se ejecuta una imputación utilizando este paquete es similar en todos los casos, para este en particular, basta con agregar los parámetros de semilla y cambiar el método tal y como se muestra a continuación en el fragmento de código:

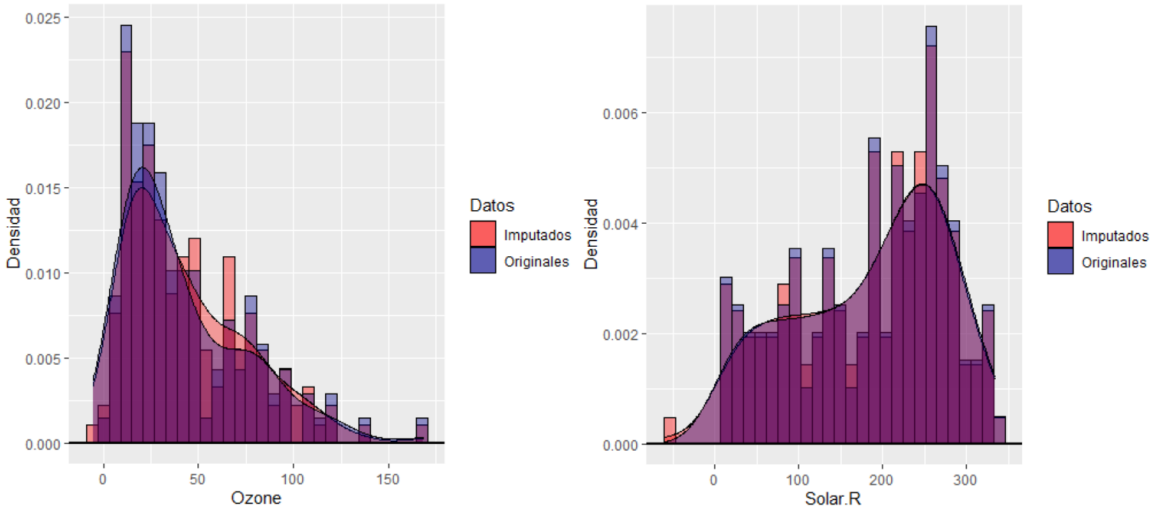
```
imp <- mice(datos, seed = 1,
            method = "norm.nob",
            m = 1, print = FALSE)
datos_imp <- imp %>%
             mice::complete(action="long") %>%
             dplyr::select(names(datos))
```

Lo cual produce una matriz de datos completos y cuyo fragmento podemos ver en la Tabla 6.3.

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
22.83	162.52	14.3	56	5	5
28	245.53	14.9	66	5	6

Tabla 6.3: Fragmento del *dataset* `airquality` habiendo realizando una imputación por regresión (estocástica). Las celdas verdes muestran la imputación realizada.

En la Figura 6.9 podemos ver los gráficos de la densidad suavizada y los histogramas de los datos originales (azul) y los datos imputados (rojo). Contrastando con lo que se observa en la Figura 6.7a, en la Figura 6.9a vemos que el sesgo ha disminuido considerablemente, esto a consecuencia de que, al haber muchos `NA`'s para esta variable, bajo esta metodología no se concentró el valor de la variable en un punto. Por otro lado, en la Figura 6.9b y comparando con su análogo en la Figura 6.7b de igual manera vemos que el sesgo desaparece, aunque por la forma de la distribución en este caso no parece ser del todo considerable, *i.e.*, este efecto no es tan notorio.



(a) Variable `Ozone`. (b) Variable `Solar.R`.

Figura 6.9: Densidad suavizada e Histograma de los datos originales (azul) vs. los imputados (rojo) de las variables de interés utilizando el método de imputación por regresión (estocástica).

En la Figura 6.10 vemos el comportamiento conjunto de estas variables, nuevamente, las observaciones que necesitaron una imputación se presentan de color rojo y las que no de color azul. En este caso y en contraste con la Figura 6.8 vemos que el comportamiento de las variables imputadas ya no es constante por variable, sino que tiene un comportamiento que incluso asemeja a los datos reales, sin embargo, puede ocurrir que existan datos atípicos o sin sentido por este método. Si se observa con detenimiento, existen ahora observaciones con valores negativos, lo cual bajo el contexto que estamos trabajando no tiene sentido y esto puede dar pie a interpretaciones que no están bien definidas.

6.2.2.1.3 Imputación *Predictive Mean Matching*

Ahora, se estará ejemplificando el Párrafo 6.1.2.2.6, para esto y al igual que el ejemplo del Párrafo 6.2.2.1.2 lo único que se debe indicar al código proporcionado por la paquetería `mice` es la semilla (al ser un método que ocupa simulaciones) y el método

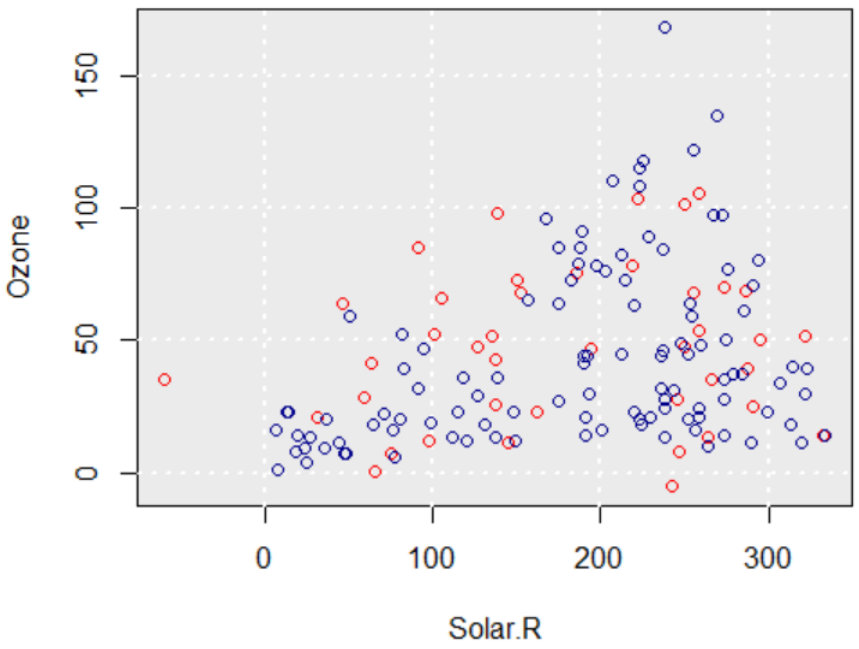


Figura 6.10: Gráfico de dispersión de los datos originales (azul) vs. los imputados (rojo) de las variables de interés utilizando el método de imputación por regresión (estocástica).

de imputación que en este caso es pmm tal y como se muestra en el siguiente fragmenteo de código:

```
imp <- mice(datos, seed = 1,
            method = "pmm",
            m = 1, print = FALSE)
datos_imp <- imp %>%
             mice::complete(action="long") %>%
             dplyr::select(names(datos))
```

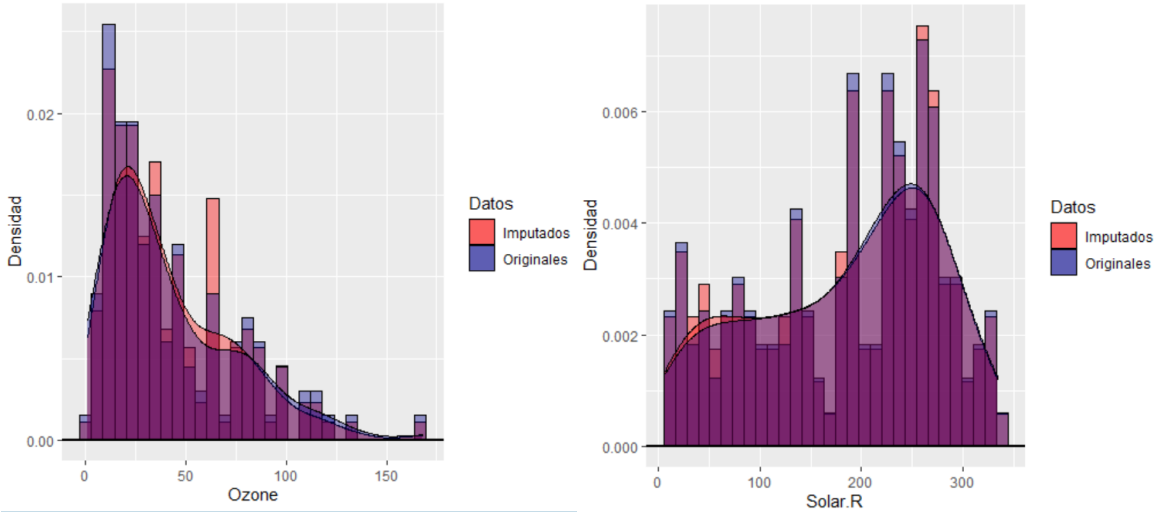
Lo cual produce una matriz de datos completos y cuyo fragmento podemos ver en la Tabla 6.4.

En la Figura 6.11 podemos ver los gráficos de la densidad suavizada y los histogramas de los datos originales (azul) y los datos imputados (rojo). A comparación de los resultados del Párrafo 6.2.2.1.2 vemos que la distribución de la densidad de los datos con y sin imputar parece mantenerse en ambos métodos. De tal manera que el sesgo se elimina a comparación de lo que sucede en el Párrafo 6.2.2.1.1 donde sí se cargan las observaciones en un punto.

En la Figura 6.12 se puede ver que los pequeños detalles de datos atípicos que se producían en el Párrafo 6.2.2.1.2 se corrigen, dando pie a observaciones que parecen ser más afines a los datos que se están observando y que no presentan un comportamiento constante por variable como en el Párrafo 6.2.2.1.1. De esta manera, el método

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
13	260	14.3	56	5	5
28	175	14.9	66	5	6

Tabla 6.4: Fragmento del *dataset* `airquality` habiendo realizando una imputación por *Predictive Mean Matching* (`pmm`). Las celdas verdes muestran la imputación realizada.



(a) Variable `Ozone`.

(b) Variable `Solar.R`.

Figura 6.11: Densidad suavizada e Histograma de los datos originales (azul) vs. los imputados (rojo) de las variables de interés utilizando el método de imputación por *Predictive Mean Matching* (`pmm`).

Predictive Mean Matching (`pmm`) parece ser el que brinda resultados más consistentes.

6.2.2.2. Imputación múltiple

Si se observan las Tablas 6.2, 6.3 y 6.4 que son las imputaciones de la Subsección 6.1.1 por la metodología de media (Párrafo 6.2.2.1.1), regresión (Párrafo 6.2.2.1.2) y `pmm` (Párrafo 6.2.2.1.3) respectivamente, vemos que los resultados pueden ser bastante diferentes entre sí. Aquí es donde entra en juego la implementación de la imputación múltiple con un objetivo en concreto.

Por ejemplo, si el objetivo final del conjunto de datos `airquality` fuese el ajustar un

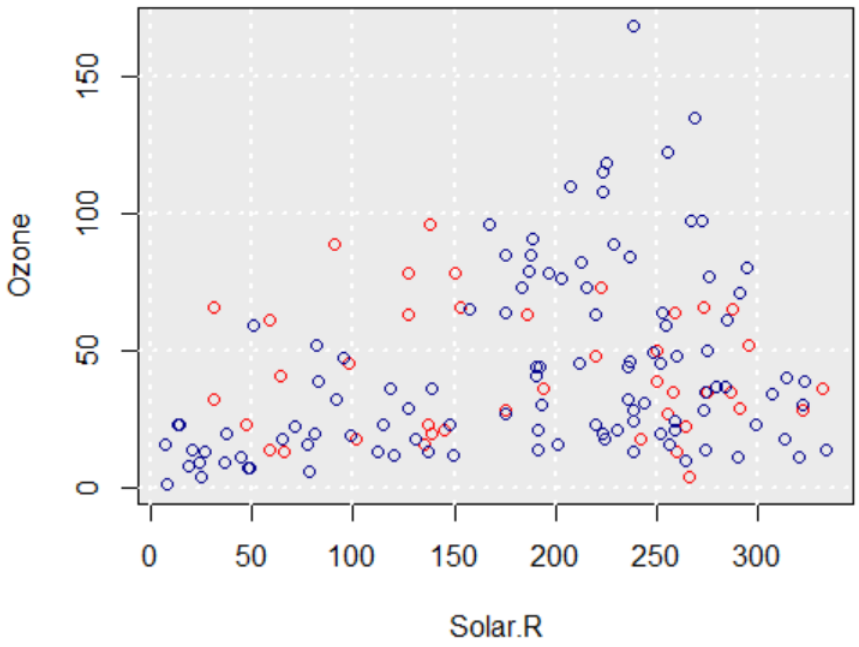


Figura 6.12: Gráfico de dispersión de los datos originales (azul) vs. los imputados (rojo) de las variables de interés utilizando el método de imputación por *Predictive Mean Matching* (`pmm`).

modelo de regresión:

$$\texttt{Ozone} \sim \beta_0 + \beta_1 \texttt{Wind} + \beta_2 \texttt{Temp} + \beta_3 \texttt{Solar.R} \tag{6.10}$$

Es decir, explicar a la variable respuesta `Ozone` en términos de las variables explicativas `Wind`, `Temp` y `Solar.R` entonces, nos estaríamos enfrentando al problema de que la variable respuesta presenta datos faltantes, por lo que existen observaciones que no nos permitirían "entrenar" al modelo. De hecho, en estos casos lo tradicional es simplemente tratar a los datos faltantes como *Listwise* (Subsubsección 6.1.1.4), algo que se puede lograr con el siguiente fragmento de código.

```
fit <- with(datos, lm(Ozone ~ Wind + Temp + Solar.R))
summary(fit)
```

Realizar el procedimiento *Listwise* nos lleva a los resultados que se observan en la Tabla 6.5.

Por otro lado, este fin se puede lograr imputando los valores faltantes, pero como comentamos al principio de esta sección, los métodos de imputación pueden dar resultados muy diferentes dependiendo de la metodología. De esta manera, entra en juego la imputación múltiple donde lo que se hace será imputar, para este ejemplo, $m = 20$ veces las variables explicativas utilizando la semilla igual a 2 y la metodología

Predictive Mean Matching (`pmm`).[6]

De esta manera, y de acuerdo con lo mencionado en la Subsubsección 6.1.2.3:

1. Se realiza el proceso de **imputación** para obtener las 20 matrices de datos completas.
2. Posteriormente se realiza el **análisis** de regresión lineal a cada una de estas matrices de datos.
3. Finalmente, se **combinan** (*pool*) los datos para así tener una estimación global de lo que en este caso se tiene como objetivo final realizar un ajuste de estimación de parámetros de acuerdo al modelo lineal establecido por la Ecuación 6.10.

Esto se puede lograr con el siguiente fragmento de código:

```
imp <- mice(datos, seed = 1, m = 20, print = FALSE) # 1. Imputación
fit <- with(imp, lm(Ozone ~ Wind + Temp + Solar.R)) # 2. Análisis
summary(pool(fit))                                  # 3. Combinación
```

Realizar el procedimiento de imputación múltiple con $m = 20$ y bajo la metodología *Predictive Mean Matching* (`pmm`) nos lleva a los resultados que se observan en la Tabla 6.5.

De la comparativa de la Tabla 6.5 podemos ver que la inferencia dada por la significancia de las pruebas realizadas a los coeficientes es la misma para ambas metodologías (algo que se comenta en la Subsubsección 6.1.2.3), además, los errores estándares en el caso de la imputación múltiple resultan un poco más grandes para algunas variables que los del procedimiento *Listwise*, lo cual es una característica deseable pues este procedimiento da pie a encontrar los parámetros de interés en un rango un tanto más amplio para obtener certidumbre.

Este ejemplo se aborda en [14] y aquí se comenta que las soluciones son casi idénticas en este caso y esto se debe al hecho de que la mayoría de los valores faltantes ocurren en la variable respuesta. Además, se menciona que la imputación múltiple es, en la mayoría de las ocasiones más eficiente que el análisis con el método *listwise*. Todo esto dependiendo de los datos y el modelo, lo cual puede llevar a resultados dramáticamente diferentes.

6.3. Imputación en el Conteo Rápido 2021

Tal y como se mencionó al principio de este capítulo y en la Subsubsección 6.1.1.2, la razón de porqué se utilizó la técnica de imputación en el Conteo Rápido de las

[6] Si a la función `mice` no se le indica la metodología, por *default* será `pmm`.

Tratamiento de Datos Faltantes	Variables	Coef	Std. Error	Statistic	$p-value$
Listwise	**(Intercept)**	**-64.342**	**23.055**	**-2.791**	**0.006**
	Wind	**-3.334**	**0.654**	**-5.094**	**0**
	Temp	**1.652**	**0.254**	**6.516**	**0**
	Solar.R	**0.06**	**0.023**	**2.58**	**0.011**
Imputación Múltiple	**(Intercept)**	**-65.878**	**23.094**	**-2.853**	**0.006**
	Wind	**-3.019**	**0.663**	**-4.557**	**0**
	Temp	**1.635**	**0.251**	**6.511**	**0**
	Solar.R	**0.059**	**0.023**	**2.585**	**0.011**

Tabla 6.5: Ajuste de una regresión lineal con los datos `airquality` tomando como modelo la Ecuación 6.10 utilizando un tratamiento de tipo de datos *Listwise* (Subsubsección 6.1.1.4) (arriba) y utilizando imputación múltiple con $m = 20$ (abajo).

elecciones federales del 2021, fue por la llegada/actualización de las remesas cada 5 minutos el día de la jornada electoral durante la tarde y noche del día 6 de Junio del 2021. Es decir, cada 5 minutos llegaba información brindada por los CAEs sobre los votos en sus casillas correspondientes tomadas de la selección de la muestra (Subsección 5.6.1) que es con lo que se estima la conformación de la Cámara de Diputados.

6.3.1. ¿Porqué se utilizó la imputación?

Si bien es cierto que el objetivo de que se tome únicamente una muestra estratificada del total es para estimar la conformación de la cámara de diputados, también es cierto que lo que se desea es que esta estimación sea lo más ágil posible, sin embargo, debido a razones humanas, aún tomando esta sub-muestra no es posible contar con el 100 % de ella en la mayoría de las ocasiones y más aún, en horas tempranas de la tarde del día de la jornada electoral, los votos siguen siendo contados en las casillas y por lo tanto su información no ha sido recibida por los miembros del COTECORA. Es por esto que se realiza una imputación de las casillas que los CAEs no han reportado basándose en la información obtenida hasta cierto momento.

Como ya se había mencionado en la Subsubsección 6.1.1.2, se pueden encontrar los archivos originales que estuvieron siendo recibidos cada 5 minutos por el COTECORA en el GitHub del autor dando clic aquí[7]. Fue con base en estos archivos que se realizaba

[7] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/6.%20Imputaci%C3%B3n/REMESAS

un proceso de imputación para estimar la conformación de la cámara de diputados cada 5 minutos cuando se contaba cada vez con más información.

6.3.2. Proceso de Imputación

Recordando que el principal objetivo del Conteo Rápido es estimar la conformación de la cámara de diputados en términos, tanto del número de escaños como del porcentaje de *Votación Válida Emitida* ($pVVE$) (Subsección 3.2.5), y estimar el porcentaje de participación ciudadana (Sección 5.2), lo que se hizo fue realizar una **imputación múltiple** que, buscaba simular las casillas de las que aún no se tenía información utilizando aquellas en las que ya se tenía, y posteriormente realizando el proceso de cálculo de la conformación de la cámara como se describe a lo largo del Capítulo 3.

Los scripts que realizan el proceso de imputación múltiple están escritos en código de R y pueden ser consultados en el GitHub del autor dando clic aquí[8], estos fueron escritos y diseñados en colaboración con el Dr. Carlos E. Rodríguez.

Antes de hacer la imputación, es necesario realizar la carga de los datos, para esto se debe tener en mente que se encuentran en archivos de texto plano (`.txt`). Para ver un ejemplo puede dar clic aquí[9]. Estos archivos fueron procesados con la función `lee_remesa_merge()` y se encuentra en el archivo `5extras_lee_y_comparte.r` la cual produce un resultado similar a lo que se mencionó en las Tablas 3.1 y 3.2. De hecho, la estructura que crea esta función sigue una estructura de variables que viene dada en la Tabla 6.6.

El proceso de imputación múltiple, basándonos en la estructura mencionada en la Subsubsección 6.1.2.3, asume la carga y procesamiento de las remesas en el objeto `res_lee` (que es el resultado de `lee_remesa_merge()`), y está compuesto por los siguientes pasos para cada remesa recibida. Además, tiene como script base el archivo de nombre `estima_imput_2021.r` del cual mostramos los fragmentos referentes a este proceso:

1. **Imputación**. Para esto, se realizaron $m = 15 =$ `m0` imputaciones por el método *Predictive Mean Matching* (`pmm`) (Párrafo 6.1.2.2.6) utilizando la remesa cargada en el objeto `res_lee`. En este caso, el modelo lineal que se utilizaba para predecir las variables que se deseaban imputar seguían, en general el modelo dado por la Ecuación 6.11:

$$\text{Variable} \sim \texttt{LISTA\_NOMINAL} + \texttt{PAN} + \texttt{PRI} + \texttt{PRD} + \texttt{MORENA} \tag{6.11}$$

[8] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/6.%20Imputaci%C3%B3n/ESTIMA_2021/ESTIMACION

[9] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/6.%20Imputaci%C3%B3n/REMESAS/REMESAS0400070210.txt

Esto fue con el objetivo de tomar a consideración el comportamiento de los partidos más populares, y se le solicitaba a la función `mice` con el objeto `pred` que es una matriz indicadora de cómo imputar las variables, resultado de la función `crea_pred_mat3()` que se encuentra en el archivo `3extras_imputacion.r`.

```
# IMPUTACIÓN
pred <- crea_pred_mat3(REMESA = res_lee$REMESA)
m0 <- 15
res_imp <- mice(res_lee$REMESA, pred = pred,
                method = "pmm", m = m0, print = FALSE)
```

Si se desea explorar a detalle con base en qué covariables se imputó cada variable, se puede observar el objeto `res_imp$formulas`.

2. **Análisis**. Para cada una de las m imputaciones anteriores, lo que se hizo fue realizar el procedimiento descrito a lo largo del Capítulo 3 (véase `conf_completa.r` y `2extras_conformacion.r`) para tener así $CONF$, la conformación de la cámara de diputados (Conformación en (3.1)) por cada una de las m imputaciones, solo que a través de una metodología *bootstrap* como en el Capítulo 5 - Subsección 5.2.1 para construir intervalos de confianza.

A diferencia del ejemplo de intervalos de confianza *bootstrap* lo que aquí se encontró fue, de cada estadística ($CONF$ y $pVVE$) asociada a cada partido político (10 en total) y al agregado de candidatos independientes, así como para la participación ($PART$) la media y varianza *bootstrap*. Este procedimiento se realizó con el siguiente fragmento de código:

```
# ESTIMO CON CADA UNA DE LAS m0 BASES
cores <- 8              # CORES QUE VOY A USAR
B <- 300                # NUMERO DE REPLICAS BOOTSTRAP
res_est_bases <- estima_bases_imput(cores, B, m0, res_imp,
                                    my_contendientes, my_partidos,
                                    INFO_COAL, INFO_BOOT_IMPUT)
```

Nótese que este proceso ocupa los núcleos (*cores*) de la computadora. Es decir, se está ocupando de una técnica computacional de programación en paralelo para así analizar y procesar más rápidamente las tablas de datos. En este caso y tras varios ensayos de eficiencia y precisión, se eligió utilizar 8 núcleos para realizar este proceso y un tamaño de remuestreo *bootstrap* de $B = 300$.

Además, los objetos `my_contendientes`, `my_partidos`, `INFO_COAL` y, finalmente, `INFO_BOOT_IMPUT` almacenan los nombres de las entidades políticas que participan, los nombres de los partidos políticos, los acuerdos de coalición (como en las Tablas 3.3 y 3.7 de la Sección 3.3) y la cantidad de casillas por estrato

(N_h y n_h de la Sección 4.2), respectivamente.

Más detalles de la función `estima_bases_imput()` se encuentran en el archivo `3extras_imputacion.r`.

3. **Combinación** (*Pool*). Finalmente, una vez teniendo las medias y varianzas *bootstrap* de cada una de las estadísticas antes mencionadas para los m conjuntos de datos completos. Se procede a realizar una **estimación combinada** (Párrafo 6.1.2.3.1) de cada una de éstas con (6.5), dando también un intervalo de confianza a través de la distribución t de Student con (6.9). Esto se logra con el siguiente fragmento de código:

```
aalpha <- 0.05 # NIVEL DE SIGNIFICANCIA
combina_imputaciones(aalpha, res_est_bases, 1, my_contendientes)
```

Nótese que `aalpha` denota en este caso el nivel de significancia que requerimos para los intervalos calculados por (6.9), en este caso se tomó del 5 %. Más detalles de la función `combina_imputaciones()` se encuentran en el archivo `3extras_imputacion.r`.

Todo este proceso se repetía cada 5 minutos, en una dinámica que implicaba desde la lectura de datos hasta la imputación múltiple (Capítulo 6) llevando dentro el cálculo de la conformación de la cámara de diputados (Capítulo 3) y re-muestreo *bootstrap* (Capítulo 4) derivado de la muestra seleccionada (Capítulo 5) para así lograr una estimación de los intervalos de confianza para las estadísticas objetivo ($CONF$, $pVVE$ y $PART$) que se deben presentar por el *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA).

Función	Variable	Descripción
Identificadores de Casilla	TIPO_SECCION	Identificador de sección
	ID_EDO_DIST	Identificador de estado - distrito
Votos otorgados directamente a Partidos Políticos	PAN	Partido Acción Nacional
	PRI	Partido Revolucionario Institucional
	PRD	Partido de la Revolución Democrática
	PVEM	Partido Verde Ecologista de México
	PT	Partido del Trabajo
	MC	Movimiento Ciudadano
	MORENA	Movimiento Regeneración Nacional
	PES	Partido Encuentro Solidario
	RSP	Redes Sociales Progresistas
	FPM	Fuerza por México
Votos otorgados directamente a Coaliciones	PAN_PRI_PRD	Va por México
	PAN_PRI	
	PAN_PRD	
	PRI_PRD	
	PVEM_PT_MORENA	Juntos Hacemos Historia
	PVEM_PT	
	PVEM_MORENA	
	PT_MORENA	
Votos otorgados directamente a Candidatos Independientes	CI1	Candidato independiente i
Votos realizados ajenos a la VVE	CNR	Candidatos No Registrados
	NULOS	Votos Nulos

Tabla 6.6: Variables de la Base de datos dada por el procesamiento de las remesas a través de la función `lee_remesa_merge()` para estimar la conformación de la cámara de diputados en el 2021.

Capítulo 7

Día de la elección: 6 de junio de 2021

7.1. El Instituto Nacional Electoral (INE)

En esta sección se comentarán algunos de los retos a los que se tuvo que enfrentar el *Instituto Nacional Electoral* (INE) para llevar acabo el Conteo Rápido de las elecciones federales del 6 de junio del 2021, así como las medidas que se tomaron en respuesta a estas problemáticas, en virtud de salvaguardar la integridad del personal de esta institución incluyendo a los miembros del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA).

7.1.1. Ubicación geográfica de puntos clave

Las oficinas centrales del INE donde se llevó a cabo de manera presencial todo lo referente al Conteo Rápido en los días cercanos a la Jornada Electoral, se encuentran en Viad. Tlalpan 100, Arenal Tepepan, Tlalpan, 14610 Ciudad de México, de acuerdo con la aplicación de Google Maps. Se puede explorar de manera interactiva dentro de las instalaciones y a sus alrededores dando clic aquí[1].

Prácticamente todas las reuniones del COTECORA se realizaron de manera virtual, salvo algunos simulacros referentes a la estimación de la cámara de diputados, la selección de la muestra y el ejercicio de estas dos de manera oficial el día de la jornada electoral. En las Figuras 7.1 y 7.2 se muestran mapas satelitales de los alrededores y de la zona interior de las oficinas centrales del INE, respectivamente. Asimismo, en la Tabla 7.1 se da una breve descripción de estas ubicaciones.

En particular, en la Figura 7.1 se muestran algunos de los puntos geográficos más importantes que rodean a las oficinas centrales del INE y que son ubicaciones que fueron estratégicas, no solo para poder ingresar y hacer uso de las instalaciones, sino también para satisfacer con las labores sanitarias debido a la pandemia por *Coronavirus disease 2019* (COVID-19). El perímetro punteado de color rojo que rodea el punto 1 de esta figura son como tal las oficinas centrales donde se realizó todo lo

[1] https://www.google.com/maps/search/ine/@19.2881396,-99.15192,238m/data=!3m1!1e3?hl=en

referente a las elecciones. En el punto 2, se tiene el Instituto Nacional de Medicina Genómica, donde realizaron pruebas COVID-19 a los integrantes del COTECORA además de también funcionar como estacionamiento de los colaboradores el día de la jornada electoral. Asimismo, los puntos 3, 4 y 5 fungieron como entradas auxiliares a las oficinas centrales del INE y estacionamientos con el objetivo de optimizar espacios en los simulacros y el mismo día de la jornada electoral.

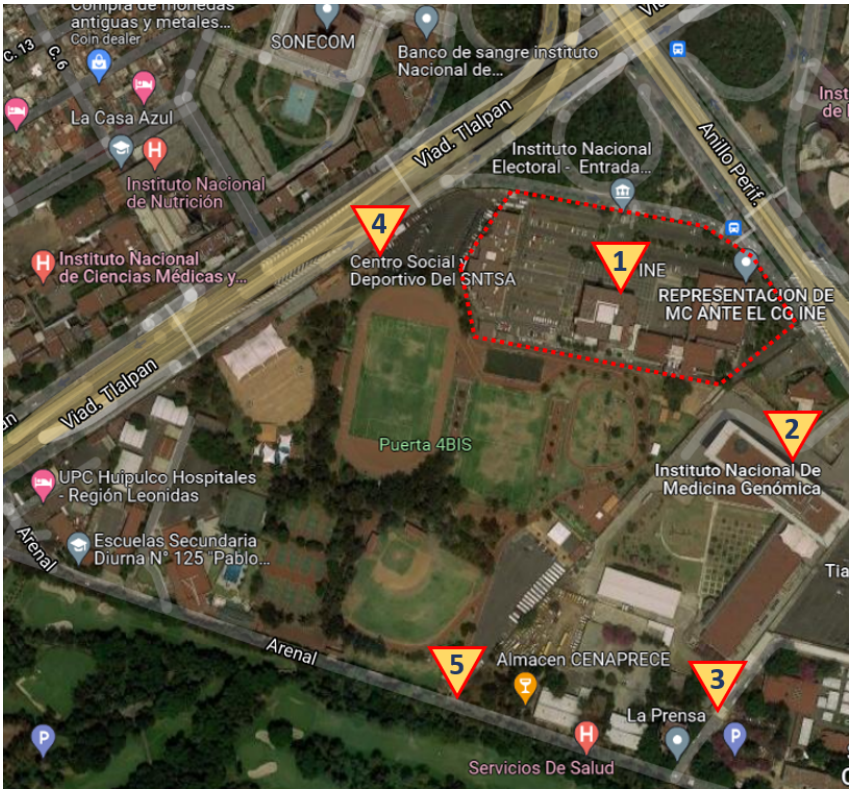


Figura 7.1: Mapa satelital de los alrededores a las oficinas centrales del *Instituto Nacional Electoral* (INE). Se puede acceder al mapa a través de Google Maps dando clic aquí.

De la misma manera en la Figura 7.2 se muestran ciertas ubicaciones dentro de las oficinas centrales del INE que fueron relevantes para llevar a cabo el Conteo Rápido. Todos estos puntos se describen en la Tabla 7.1 y entre éstos se destacan los que mencionaremos a continuación:

- En el punto 6 tenemos el edificio principal del INE, mismo que se muestra en la Figura 7.3 acompañado de una fotografía del autor de esta tesis haciendo uso de un cubrebocas, lo cual refleja la situación mundial a consecuencia de la pandemia de COVID-19 por la que se estaba atravesando en esos momentos.

- En el punto 7 se encontraba la ubicación del búnker donde los miembros del COTECORA llevarían a cabo el Conteo Rápido de las elecciones de ese año.

Una fotografía que muestra el interior de este búnker se puede encontrar en la Figura 7.6.

- El punto 8 se utilizó para que los miembros del COTECORA realizaran pruebas COVID-19 antes de los simulacros y el ejercicio del Conteo Rápido. Fue uno de los puntos preventivos que necesitaban ser implementados para salvaguardar la integridad de los miembros del Comité.
- En el punto 9 se realizó en vivo la selección de la muestra para el Conteo Rápido (Subsección 5.6.1). En la Figura 7.4 se muestra una fotografía capturada durante uno de los simulacros para llevar a cabo este proceso.
- En el punto 10 se encuentra la sala de conferencias del consejo general del INE (Figura 7.7) aquí es donde se celebran los acuerdos y decisiones tomadas por los miembros del consejo.
- Los demás puntos sirven como accesos y estacionamientos que, dependiendo de la actividad que se tenía programada, se debía acceder a través de estos puntos.

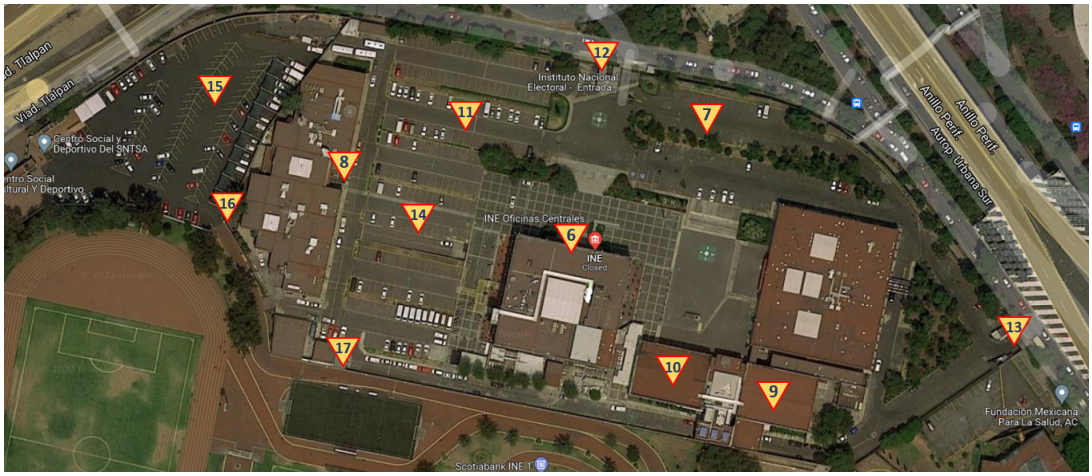


Figura 7.2: Mapa satelital del interior de la zona de las oficinas centrales del *Instituto Nacional Electoral* (INE). Se puede acceder al mapa a través de Google Maps dando clic aquí.

7.1.2. Preparativos para el Conteo Rápido

Con el objetivo de llevar a cabo el Conteo Rápido de manera eficiente y segura es necesario que el INE implemente ciertas medidas tales como la instalación de un búnker en sus oficinas centrales, que es un espacio donde los integrantes del COTECORA estarán realizando las estimaciones. A consecuencia de esta convivencia presencial y por la emergencia sanitaria de estos tiempos, también fue requerido realizar pruebas de COVID-19 cada vez que se realizara un simulacro presencial y también previo al Conteo Rápido oficial.

Mapa	ID	Descripción
Figura 7.1 (Alrededor del INE)	(1)	Oficinas centrales del *Instituto Nacional Electoral* (INE).
	(2)	Instituto Nacional de Medicina Genómica.
	(3)	Acceso a estacionamiento del Instituto Nacional de Medicina Genómica.
	(4)	Acceso a estacionamiento del Centro Social y Deportivo del SNTSA.
	(5)	Arenal 4 Bis - Puerta 4.
Figura 7.2 (Dentro del INE)	(6)	Edificio principal del INE.
	(7)	Ubicación del búnker para el COTECORA.
	(8)	Zona designada para pruebas COVID-19.
	(9)	Sala de conferencias para la selección de la muestra.
	(10)	Sala conferencias del consejo general.
	(11)	Zona designada para medios de comunicación.
	(12)	Entrada peatonal principal a las oficinas del INE.
	(13)	Entrada de automóviles a las oficinas INE.
	(14)	Estacionamiento principal del INE.
	(15)	Estacionamiento auxiliar del INE.
	(16)	Entrada peatonal auxiliar 1 a las oficinas del INE.
	(17)	Entrada peatonal auxiliar 2 a las oficinas del INE.

Tabla 7.1: Descripción de las ubicaciones geográficas mostradas en las Figuras 7.1 y 7.2.

(a) Vista de la fachada. (b) Evidencia de la presencia del autor.

Figura 7.3: Vista de frente del edificio principal de las oficinas centrales del *Instituto Nacional Electoral* (INE). Esto se encuentra en la ubicación 6 de la Figura 7.2.

Como parte de las actividades a desarrollar, el COTECORA realiza múltiples simulacros, no solo para realizar el Conteo Rápido (Figura 7.6), sino también para la selección de la muestra (Figura 7.4). Estos simulacros se realizan semanas y días antes del 6 de junio de 2021, en particular la selección de la muestra oficial se llevó a cabo el 4 de junio de 2021 (Subsección 5.6.1) con el objetivo de que las casillas seleccionaras fueran notificadas anticipadamente de su participación en el Conteo Rápido por medio de los CAEs.

En cuanto al equipo de trabajo, el INE ponía a la disposición el uso de una computadora portátil (laptop) con la que los integrantes del COTECORA se podían apoyar para el cálculo de las estimaciones. Sin embargo, los integrantes del Comité podían hacer uso de su equipo de cómputo personal, tomando a consideración que no podrán disponer de conexión a internet y que se tendrá que configurar la red interna del INE para poder recibir los datos proporcionados por los CAEs y presentar las estimaciones.

Cada vez que se realiza el Conteo Rápido el INE monta un búnker donde, durante la tarde del día de la jornada electoral, los miembros del COTECORA estarán realizando las estimaciones. Este es un espacio en donde se prohíben los televisores, aparatos de radio, ver noticias o el uso de dispositivos móviles en los momentos en los que se estén haciendo las estimaciones; para lo que los integrantes del Comité son despojados de aparatos electrónicos de ésta índole y la comunicación en el exterior está monitoreada por el INE y restringida únicamente para ciertas personas dentro del búnker, con el objetivo de minimizar las posibilidades de un ataque cibernético. El 15 de junio del 2018, el sitio web "*Eje Central*" publicó una imagen que muestra algunas especificaciones del búnker que fue utilizado para las elecciones de ese año.

El búnker para las elecciones del 2021 tuvo que ser acondicionado para que se respetaran las medidas sanitarias a consecuencia de la pandemia por COVID-19. En este espacio se cuenta también con aire acondicionado, servicio de alimentos, café, agua,

Figura 7.4: Evidencia de la presencia del autor en los simulacros para la selección de la muestra (Subsección 5.6.1). Esto se encuentra en la ubicación 9 de la Figura 7.2.

etc. Incluyendo servicios tales como baños para evitar que los integrantes del comité tengan necesidad de abandonar el espacio. Este es un trabajo que finalmente debe ser impoluto y apegado a principios científicos y sin ninguna influencia del ruido que pueda ser generado, por ejemplo, de los medios de comunicación.

7.2. Las semanas Previas a la Elección

En fechas cercanas al día de la Jornada Electoral (6 de junio del 2021) salió a debate las reglas dictadas por el *Instituto Nacional Electoral* (INE) para limitar los curules en el Congreso y evitar la sobrerrepresentación, fijada por la (CPEUM, [4]) en un 8 %.

Lo que se debatía era un cambio en dichas reglas que conforman el cálculo de los escaños otorgado por el principio de *Representación Proporcional* (RP). El mecanismo que determinaría la asignación de curules por RP para las elecciones en cuestión está fundamentado en el acuerdo INE/CG193/2021 del consejo general del INE.

Este debate podría causar un impacto importante en la estimación realizada por el Conteo Rápido, ya que las metodologías implementadas hasta este entonces fueron implementadas tal y como se menciona en el Capítulo 3, de tal manera que se tendría que hacer un cambio en el desarrollo computacional de estos algoritmos para adap-

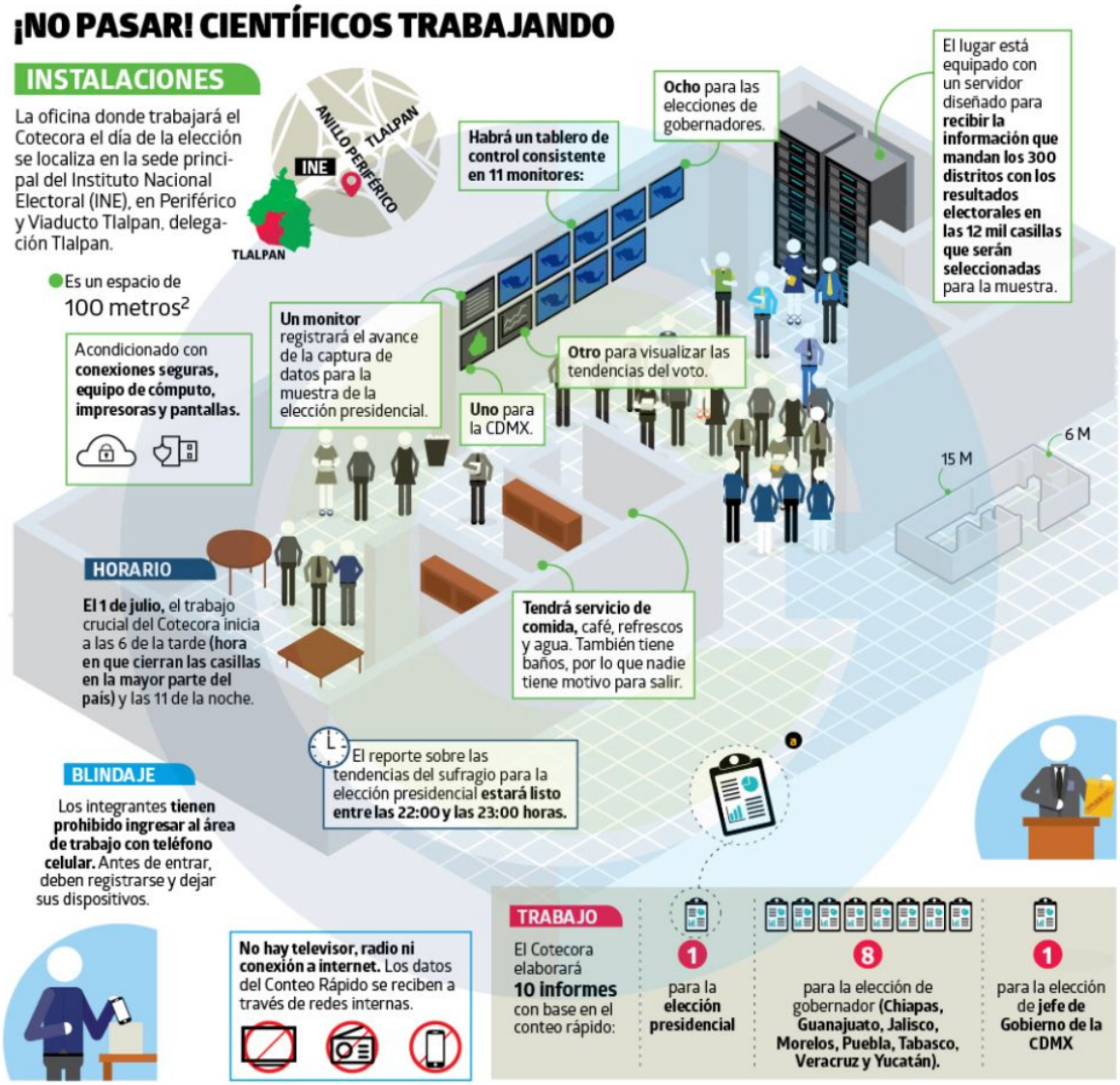


Figura 7.5: Descripción gráfica del búnker para los miembros del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA) para las elecciones del 2018. Imagen publicada por el sitio web "*www.ejecentral.com.mx*" el día 15 de junio del 2018. Esta nota puede ser consultada dando clic aquí.

tarse al posible cambio que este debate generaría.

La razón principal del porqué surgió esta polémica fueron por los puntos mencionados en la Subsección 3.3.2, en particular sobre lo comentado derivado de las Tablas 3.2 y 3.9. Lo que dio pie a la implementación de la *Afiliación Efectiva* (Subsección 3.2.6).

El 22 de Marzo del 2021 el INE publicó en su página de internet "Central Electoral" un artículo multimedia con la participación del Consejero Electoral del INE Ciro Murayama, donde se explica públicamente qué es la sobrerrepresentación y las circunstancias

Figura 7.6: Fotografía del interior del búnker en donde los miembros del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA) llevaron a cabo el Conteo Rápido. Esta imagen fue capturada durante uno de los simulacros. El búnker se encontraba en la ubicación 7 de la Figura 7.2.

que habían dado pie a que este principio fuese "distorsionado artificialmente". Dicho artículo se puede explorar dando clic aquí[2].

7.2.1. Notas en periódicos

Diversos medios de comunicación cubrieron la nota que giraba alrededor de un posible cambio que sufriría el principio de *Representación Proporcional* (RP) con el objetivo de cubrir la no sobrerrepresentación, lo cual sería una modificación a la (CPEUM, [4]) a fechas muy cercanas del día de la Jornada Electoral y puso en alerta a los militantes de los partidos políticos participantes.

De acuerdo con [19], que fue publicado el día 27 de Abril del 2021:

> "*El mecanismo aprobado por los siete magistrados de la Sala Superior busca frenar el trasvase de candidatos entre fuerzas que se presentan en coalición y así evitar desvirtuar los grupos parlamentarios. El acuerdo fue*

[2]https://centralelectoral.ine.mx/2021/03/22/con-la-aprobacion-de-criterios-que-buscan-evitar-la-sobrerrepresentacion-en-la-camara-de-diputados-se-trata-de-garantizar-que-la-votacion-de-la-gente-se-traduzca-de-manera-nitida-en-asientos-en-san-la/

Figura 7.7: Sala de conferencias del consejo general del *Instituto Nacional Electoral* (INE). Esto se encuentra en la ubicación 10 de la Figura 7.2.

> *adoptado por el INE en marzo con vistas a las elecciones que el 6 de junio renovarán la Cámara de los Diputados, además de las gubernaturas de 15 de las 32 entidades federativas. Fue impugnado por Morena, el PAN y Encuentro Solidario. El presidente, Andrés Manuel López Obrador, estalló ante la medida y habló de complot de la autoridad electoral.*
>
> *...*
>
> ***El nuevo mecanismo tiene mucha trascendencia y supone un cambio estructural de gran calado**. De las 500 curules de la Cámara, 300 se votan de forma directa en las urnas -los diputados uninominales- y 200 -los plurinominales- se designan mediante criterio de representación proporcional. **Este nuevo esquema incorpora la figura de "afiliación efectiva"**. Las formaciones, como Morena, que concurran a las elecciones bajo fórmulas de coalición deberán presentar candidatos que puedan acreditar su militancia. El objetivo, señala el tribunal, es que no haya "préstamo" de candidatos de un partido mayoritario a uno minoritario.*
>
> *La sentencia recuerda que, según la Constitución, "en ningún caso un partido político podrá contar con un número de diputados [...] que representen un porcentaje del total de la Cámara que exceda en ocho puntos a su porcentaje de la votación nacional emitida". Con esta premisa, según los magistrados, el Instituto Nacional Electoral (INE) no vulneró "los dere-*

chos político-electorales de los candidatos, además de que el acuerdo no fue emitido de manera extemporánea, como argumentaban los partidos quejosos". ***El presidente nacional de Morena, Mario Delgado, llegó a calificar este mecanismo de "maniobra oscura y vergonzosa". Y el vocero presidencial, Jesús Ramírez, consideró, en conversación con EL PAÍS, que se trataba de un intento de "cambiar las reglas del juego a mitad del partido"****."*

Tiempo después y derivado de estos acontecimientos, se comenzó la especulación sobre el impacto que tendría la *Afiliación Efectiva* sobre el partido político Morena. El 16 de Mayo del 2021 se publicó [20] donde se mencionaba:

"*Morena, el partido del Gobierno mexicano, dejará de tener la mayoría absoluta tras las elecciones del próximo 6 de junio y necesitará pactar con partidos aliados para controlar el Congreso, según una encuesta de SIMO Consulting para EL PAÍS. La composición de la Cámara de Diputados definirá el futuro político de México para los próximos tres años. Las votaciones pondrán a prueba, principalmente al Movimiento de Regeneración Nacional (Morena) y los planes del presidente, Andrés Manuel López Obrador, y su llamada Cuarta Transformación.*

...

Hay que subrayar la incertidumbre en estas proyecciones. ***A tres semanas de una elección que está movilizando a todo el país, aún hay espacio para vaivenes. Algunos podrían ser en la intención de voto, pero otros pueden darse en la propia afiliación efectiva de los candidatos****. El Instituto Nacional Electoral (INE) estableció en marzo una serie de normas destinadas a garantizar la regla constitucional de proporcionalidad, fijada en un 8 %. Estas normas podrían terminar afectando el número final de curules a disfrutar por el grupo parlamentario de cada partido. La proyección presentada por SIMO tiene en cuenta tanto las normas como la incertidumbre asociada, y por eso ofrece intervalos amplios. Pero ni siquiera dentro de dichos intervalos está Morena, a día de hoy, en condiciones de revalidar su mayoría absoluta en la Cámara."*

A su vez, el mismo 16 de Mayo del 2021 se publicó [21] donde se mencionaba lo siguiente:

"*En la conformación de la Cámara baja de la actual Legislatura, en 2018, la coalición "Juntos Haremos Historia"(Morena, PT y Encuentro Social) logró el 45.9 % de los sufragios, pero se le asignó el 61.6 % de la Cámara, lo que significó una sobrerrepresentación de 15.7 por ciento. De acuerdo con el consejero electoral, Ciro Murayama, esto superó el límite constitucional casi por partida doble.*

Situaciones similares ocurrieron en años de elecciones legislativas anteriores. En 2012, la coalición PRI-PVEM obtuvo el 40.0 % de los votos y el 48.2 % de los escaños: rebasando en apenas 0.2 % el límite constitucional. En 2015, la misma coalición del PRI-PVEM recibió 40.3 % de los votos y el 50 % de los diputados, 9.7 % más.

Derivado de estas experiencias y de cara a las elecciones de este año, el Instituto Nacional Electoral (INE) decidió tomar medidas adicionales a favor de proteger la representatividad del voto popular en la Cámara de Diputados.

...

En la fracción V del artículo 54 dispone que al momento de repartir las diputaciones por representación proporcional (plurinominales) y sumarlas a los escaños ganados por mayoría, un partido no puede tener 8 % más legisladores de lo que sacó de porcentaje de votación.

"En ningún caso un partido podrá contar con un número de diputados por ambos principios que representen un porcentaje del total de la Cámara que exceda en ocho puntos su porcentaje de votación nacional emitida"

Es decir, si un partido consiguió el 30 % de los votos, no podrá tener más del 38 % del total de los diputados.

En 2012, de la coalición formada por el Partido Revolucionario Institucional (PRI) y el Partido Verde Ecologista de México (PVEM), resultaron triunfadores cinco candidatos registrados a nombre del PVEM, pero que en realidad eran priistas. Con este mecanismo, el PRI evitó el tope de sobrerrepresentación; la coalición, que obtuvo 40 % de los votos, sumó 241 diputaciones (48.2 % de la Cámara, un lugar arriba del tope constitucional). En 2015 estos dos partidos, coaligados, aplicaron la misma fórmula.

En las elecciones de 2018, la coalición entre Morena, el PT y el PES adoptó el mismo mecanismo. En 292 distritos participaron en coalición y en ocho distritos lo hicieron por separado. De 220 distritos de mayoría relativa ganados por estos partidos, 106 fueron postulados por Morena, 58 por el Partido del Trabajo (PT) y 56 por el PES. En todos ellos, sin excepción, los votos fueron mayoritariamente para Morena, mientras que la votación del PT o del PES no habría alcanzado para ganar en ningún distrito.

...

> *Apenas fueron declarados electos, muchos de aquellos candidatos prestados al PT y al PES volvieron a su partido de origen para sumarse a la fracción parlamentaria de Morena. En tanto, cinco diputados del PVEM se agregaron a la fracción morenista y así le aseguraron a ésta la mayoría absoluta de la Cámara (requisito legal para presidir la Junta de Coordinación Política por los tres años de la legislatura, sin tener que alternarla con otros partidos).*
>
> *La representación política también se puede distorsionar cuando un diputado renuncia a su bancada original para cambiarse a otra o permanecer en calidad de "sin partido". **Esta maniobra parlamentaria permite formar mayorías artificiales**, con la finalidad de controlar organismos internos del Congreso como la Junta de Coordinación Política o la Mesa Directiva.*
>
> *...*
>
> *El Instituto detalló que considera "**afiliación efectiva**" a aquella que esté vigente al momento del registro de la candidatura.*
>
> *Así, el triunfo será contabilizado a favor del partido con el cual la persona ganadora tenga una "**afiliación efectiva**".*
>
> *El INE acordó que los diputados electos deben **contar para el partido al que realmente han pertenecido** y no al que dicen pertenecer al momento de registrarse para la candidatura. "*

Así se publicaba la estrategia de los partidos políticos más influyentes para buscar burlar el sistema de no sobrerrepresentación establecido en el principio de *Representación Proporcional* (RP).

Las acciones a tomar con respecto al cálculo de curules bajo el principio de *Representación Proporcional* (RP) por parte del Consejo General del INE eran inciertas, lo cual causó intriga a los miembros del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA) principalmente por las fechas tan cercanas al día de la jornada electoral.

7.2.2. La determinación del Consejo General y la última impugnación a unos días de la Elección

Durante el día 20 de Mayo del 2021 se publica el acuerdo INE/CG467/2021 por el consejo general del INE. En éste, se da respuesta a la consulta formulada por el secretario técnico del COTECORA en relación con el mecanismo de asignación de los curules por el principio de *Representación Proporcional* (RP), aprobado mediante al

acuerdo INE/CG193/2021.

A lo largo del acuerdo publicado ese día, se repasa **el cálculo de la conformación de la cámara de diputados como se menciona en el Capítulo 3**. Donde finalmente se decide así **aprobarlo** en la sesión extraordinaria del Consejo General celebrada el 20 de mayo de 2021, por **siete votos a favor** de los Consejeros Electorales, Maestra Norma Irene De La Cruz Magaña, Doctor Uuc-kib Espadas Ancona, Doctora Adriana Margarita Favela Herrera, Maestro José Martín Fernando Faz Mora, Carla Astrid Humphrey Jordán, Maestra Dania Paola Ravel Cuevas y Doctor José Roberto Ruiz Saldaña, y **cuatro votos en contra** de los Consejeros Electorales, Doctor Ciro Murayama Rendón, Maestro Jaime Rivera Velázquez, Maestra Beatriz Claudia Zavala Pérez y el Consejero Presidente, Doctor Lorenzo Córdova Vianello.

Esto da pie a que los cálculos para conformar la cámara de diputados queden bien definidos y establecidos, de tal manera que el COTECORA pudiese realizar las estimaciones pertinentes el día de la Jornada Electoral fundamentadas con los acuerdos del Consejo General del INE.

7.3. El día de la jornada electoral

En esta sección se narran los acontecimientos desde el punto de vista del autor de cómo se vivió el día de la Jornada Electoral, 6 de junio de 2021. Una vez desarrollados todos los cómputos e implementados en el lenguaje de programación R, esta parte fue realizada de manera prácticamente automática para que las estimaciones fuesen algo que únicamente se debía observar el día de la Jornada Electoral sin la necesidad de estar operando el programa.

7.3.1. Previo al confinamiento en el búnker

Durante el día de la mañana de este día los miembros del COTECORA podían realizar sus actividades habituales libremente. Al igual que todos los mexicanos, se tiene la libertad de elegir si ejercer el sufragio, esto sin influir en ningún sentido los resultados que se presentan en el Conteo Rápido.

Esto sucede ya que los votos apenas están siendo emitidos por el electorado, y no pueden ser contados hasta que las casillas sean cerradas en la tarde de este mismo día. En consecuencia, existe una importancia fundamental en cuanto a los tiempos de captura de la información por parte de los CAEs, ya que México cuenta con distintos husos horarios y por lo mismo, las casillas cierran a distintas horas dependiendo de su ubicación geográfica. De hecho, esto es algo que se debe de contemplar en el momento de la selección de la muestra (Subsección 5.6.1), para que se cuente con una cantidad de casillas apropiada para el final del día de cada uno de los distritos federales electorales, y las estimaciones con estos datos sean lo más confiables posible y el Conteo

Rápido sea así relevante.

Cercana la tarde de este día, los miembros del COTECORA deben reunirse en las oficinas centrales del INE. Tomando en cuenta las ubicaciones geográficas de las Figuras 7.1 y 7.2. Primero para realizar una prueba COVID-19 y posteriormente para entrar al búnker y así comenzar las estimaciones siguiendo los protocolos de seguridad y en este caso de sanidad (Figura 7.8).

Figura 7.8: Fotografía del interior del búnker en donde los miembros del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA) llevaron a cabo el Conteo Rápido. Esta imagen fue capturada el día de la jornada electorial previo al aislamiento. El búnker se encontraba en la ubicación 7 de la Figura 7.2.

7.3.2. Durante el Conteo Rápido

Al comenzar el proceso de estimación de la conformación de la cámara de diputados empezó también el confinamiento de los integrantes del COTECORA en el búnker. Por lo que en estos puntos el uso de dispositivos móviles quedaba prohibido y también la salida de este espacio.

La muestra comenzaba a llegar en forma de los archivos que se denominan como *Remesas* (Subsubsección 6.1.1.2). Esta muestra era compartida con los miembros del COTECORA y que provenía de los CAEs a través de la red INE siendo actualizada cada 5 minutos.

Los programas automatizados fueron implementados utilizando el lenguaje computacional R con el apoyo visual de una herramienta conocida como *Shiny* que es capaz de crear aplicaciones cimentadas en este lenguaje de programación. Estos programas, ya previamente puestos a prueba desde los simulacros, fueron activados desde que la información comenzó a llegar.

7.3.2.1. Aplicación en R Shiny

La aplicación en R *Shiny* (Figura 7.9) que fue utilizada para el Conteo Rápido puede ser consultada en el GitHub del autor dando clic aquí[3]. En este enlace se encuentra un archivo de nombre `LEEME.txt` en donde se exhiben las instrucciones para ejecutar los programas que se implementaron para la estimación de la Cámara de Diputados en colaboración con el Dr. Carlos E. Rodríguez, y que parte de su aplicación el día de la Jornada Electoral se puede ver en la Figura 7.8. Más adelante se muestran los resultados que desplegaban estos programas en diferentes momentos de la tarde y noche de la Jornada Electoral.

Esta aplicación está compuesta por múltiples secciones/pestañas (marcas "1" a la "4"), en particular, primero se explicará la parte principal en la Figura 7.9 con el apoyo de la Tabla 7.2, donde se muestran diversos gráficos (marca "7") que dependen de la remesa más actual que haya llegado y se actualizan automáticamente de acuerdo a la llegada de éstas, también se muestra una tabla (marca "8") en la parte inferior que indica los valores numéricos de los gráficos (marca "7"), que se están mostrando en la parte superior para cada partido político. Esta aplicación se puede manipular (marcas "9" y "10") de tal manera que los resultados que se estén observando sean únicamente los solicitados, en este caso, en términos de partidos políticos (marca "9"), *i.e.*, si por ejemplo se desea únicamente ver los intervalos de los partidos `MORENA` y `PAN` se pueden alterar las casillas de la izquierda para que se desplieguen únicamente los dos intervalos relacionados a estos partidos políticos. También, es posible alterar el gráfico con base en la hora de llegada de las remesas (marca"10") con el objetivo de analizar única-

[3]https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/7.%20Jornada%20Electoral

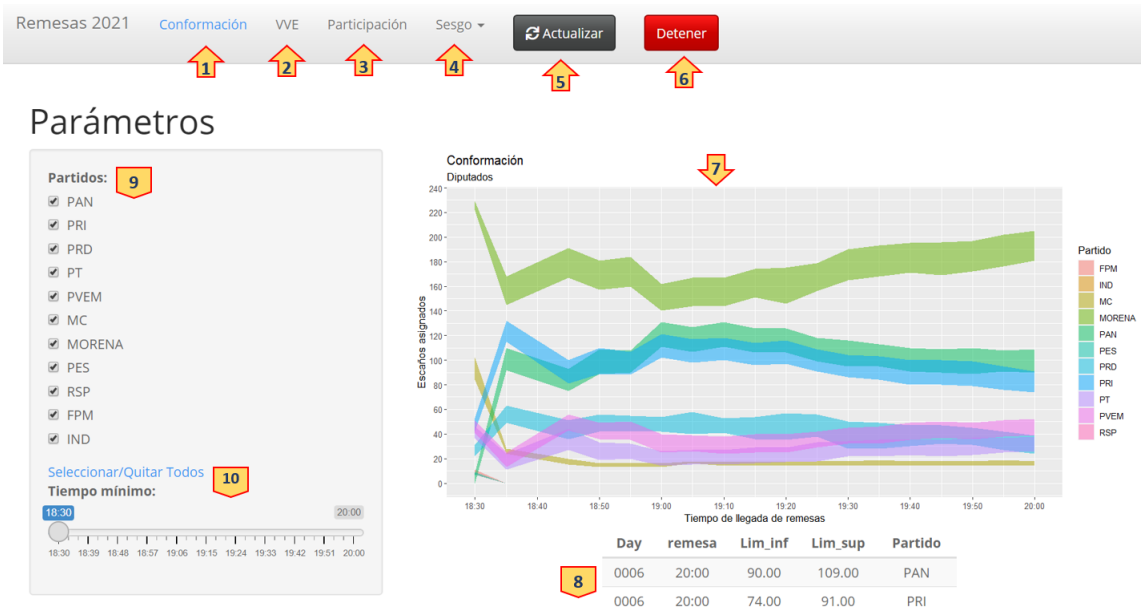


Figura 7.9: Interfaz principal de la aplicación en *Shiny* que mostraba en tiempo real, con base en las remesas actualizadas cada 5 minutos, los resultados de las estimaciones por la metodología implementada.

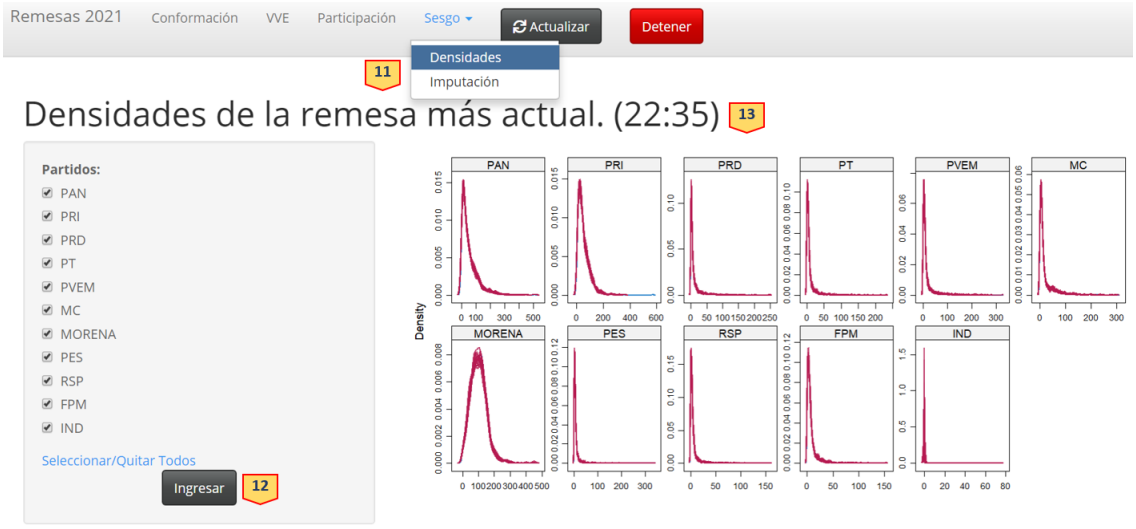


Figura 7.10: Interfaz secundaria de la aplicación en *Shiny*. Aquí se mostraba el comportamiento del sesgo en las estimaciones derivadas del proceso de imputación múltiple (Subsubsección 6.1.2.3).

mente los intervalos a horas más avanzadas del desarrollo de la llegada de las remesas.

En la interfaz secundaria de esta aplicación (Figura 7.10) se tiene la exploración del sesgo de las estimaciones a la hora de la remesa más actual. De tal manera que, para cada estimación, se puede observar el comportamiento del efecto en el sesgo que está teniendo el proceso de imputación múltiple (Subsubsección 6.1.2.3). La marca

Interfaz	Marca	Funcionamiento
Principal (Figura 7.9)	1	Pestaña de los resultados de $CONF$ (Ecuación 3.1)
	2	Pestaña de los resultados de $pVVE$ (Ecuación 3.8)
	3	Pestaña de los resultados de $PART$ (Ecuación 5.3)
	4	Entrada a la Interfaz Secundaria (Sesgo)
	5	Actualizar manualmente la aplicación
	6	Detener manualmente la aplicación
	7	Gráfico de la pestaña actual
	8	Tabla del gráfico de la pestaña actual
	9	Partidos políticos que se desean observar
	10	Tiempo mínimo a mostrar en el gráfico ("eje x")
Secundaria (Figura 7.10)	11	Opciones de la interfaz secundaria
	12	Gráficos para explorar el sesgo
	13	Botón para ejecutar la exploración del sesgo con los partidos ingresados.

Tabla 7.2: Descripción de las marcas en la aplicación *Shiny* (Figuras 7.9 y 7.10).

"11" muestra un pequeño sub-menú donde se pueden cambiar los gráficos a observar, y para observar los gráficos de densidades es necesario presionar el botón indicado en la marca "12", para que finalmente los gráficos sean desplegados debajo de donde se muestra la marca "13" a un lado de la cual se indica la hora de la remesa más actual.

Una versión *beta* de la aplicación interactiva que utiliza datos simulados y se diseñó para el Conteo Rápido, puede ser manipulada en cualquier dispositivo con navegador a internet haciendo clic aquí[4].

Cada vez que se realizaba una estimación, ésta se guardaba en un archivo del tipo `.csv` (archivo separado por comas) y con un nombre análogo a `rodriguez00061915.csv` en donde el apellido indicaba que era una estimación realizada por el equipo del investigador Carlos E. Rodríguez y los 6 últimos caracteres previos a la extensión indicaban el día y la hora (en formato de 24 hrs.) en la que la estimación había sido realizada proveniente de la remesa de esa misma hora. Las estimaciones eran realizadas ya de forma automática cada cierto tiempo por el Script en código de nombre `0lazy_run.r`, utilizando los demás scripts que pueden ser consultados dando

[4]https://edgar-alarcon.shinyapps.io/CONF_2021/

clic aquí[5].

Las estimaciones de la Conformación de la Cámara de Diputados, así como del correspondiente porcentaje de *Votación Válida Emitida* (VVE) para cada partido político (Subsección 3.2.5) y de la participación (Sección 5.2) fueron construidas con la teoría mencionada en la Sección 6.3 y eran compartidas a cartografía (Subsubsección 7.3.2.2) por red INE utilizando el Script de R de nombre `5extras_lee_y_comparte`. Estas estimaciones pueden ser consultadas en dando clic aquí[6], en particular las estimaciones de la Conformación de la Cámara de Diputados (Ecuación 3.1), se encuentran en la carpeta `diputaciones_pef` y las estimaciones de la $pVVE$ (Ecuación 3.8) junto con la Participación (Ecuación 5.3) se encuentran en la carpeta `pef`.

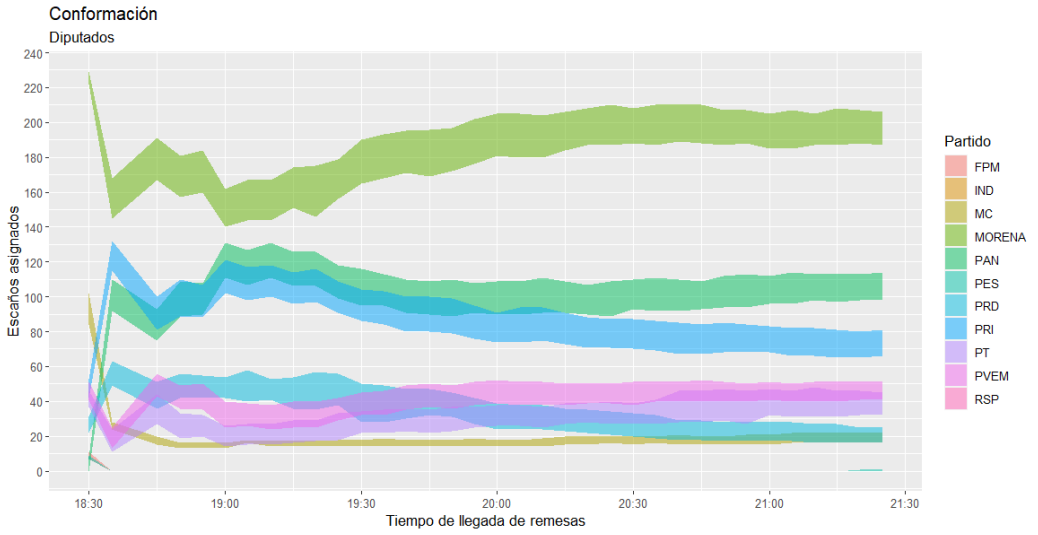
Los gráficos de la remesa de las 20:00 que fueron procesados por la aplicación en R *Shiny* pueden ser consultados en el anexo en las Figuras 9.3 (referente a la Conformación de la Cámara de Diputados y el porcentaje de VVE), 9.4 (referente a la participación ciudadana) y 9.5 (referente a la imputación múltiple). Sin embargo, las dos primeras figuras se encuentran contenidas en los gráficos que mostraremos a continuación, y que hemos agregado para dar una idea cómo es que fueron dándose los cambios en los gráficos a lo largo del tiempo.

Las horas pasaban y los gráficos producidos por la aplicación mostraban el avance de las estimaciones, en un principio los intervalos parecían moverse mucho y esto debido a la escasa cantidad de muestra con la que se contaba, pero cada vez que se contaba con más muestra y esto mostraba a la larga un comportamiento asintótico que describían los gráficos. Esto finalmente estaba poco a poco conformando un intervalo de confianza con el cual ya el valor real de la muestra estaría contenido en estos intervalos. A continuación explicaremos cada uno de los gráficos y mencionaremos qué interpretación tenían estos bajo el contexto del Conteo Rápido.

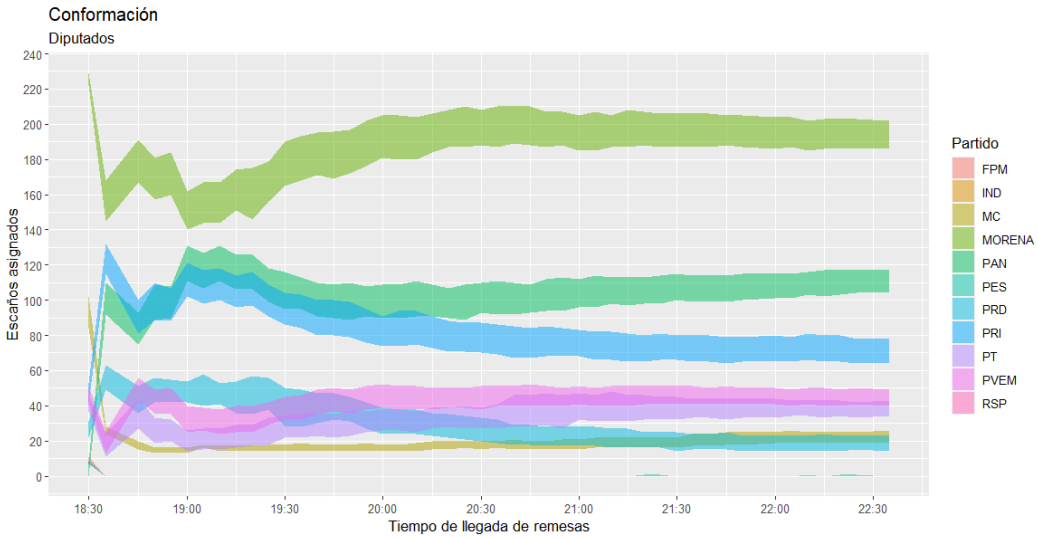
En la Figura 7.11 podemos ver el avance en las estimaciones que hay como tal en la conformación de la cámara de diputados. Es decir el número de escaños que se estaba asignando a cada partido político. Durante las primeras horas de las estimaciones se puede observar la volatilidad en los intervalos, es decir que estaban siendo muy cambiantes e incluso con el avance de las horas algunos de los que parecían estar superpuestos poco a poco se fueron separando. Eso pasa, por ejemplo, con los partidos Políticos PRI y PAN que son precisamente parte de las potencias políticas que hay en México. Sin embargo el partido que siempre se mostró dominante fue MORENA ya que de igual manera por contexto reciente, el presidente de México en estas fechas es precisamente de este partido político y ha estado recibiendo mucho apoyo por parte de la sociedad mexicana. Por otro lado, los demás partidos parecen no tener mucha relevancia y su presencia en la cámara de diputados es bastante nula a excepción de

[5] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/7.%20Jornada%20Electoral/R_Apps/TESIS_INE_2021/ESTIMACION

[6] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/7.%20Jornada%20Electoral/R_Apps/TESIS_INE_2021/rodriguez_sample



(a) 21:25

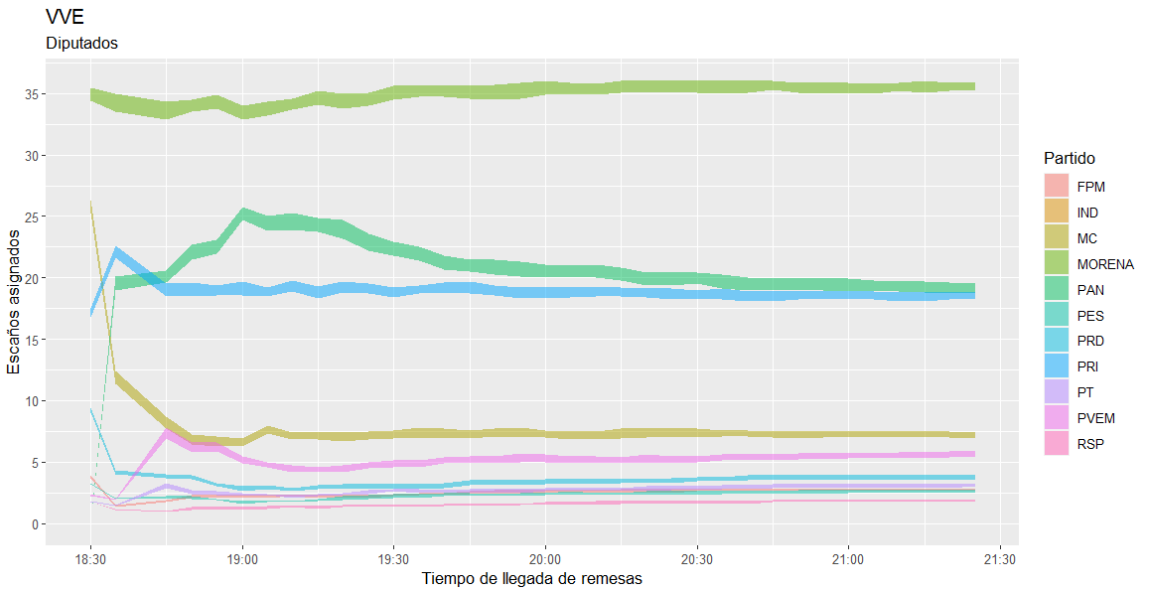


(b) 22:35

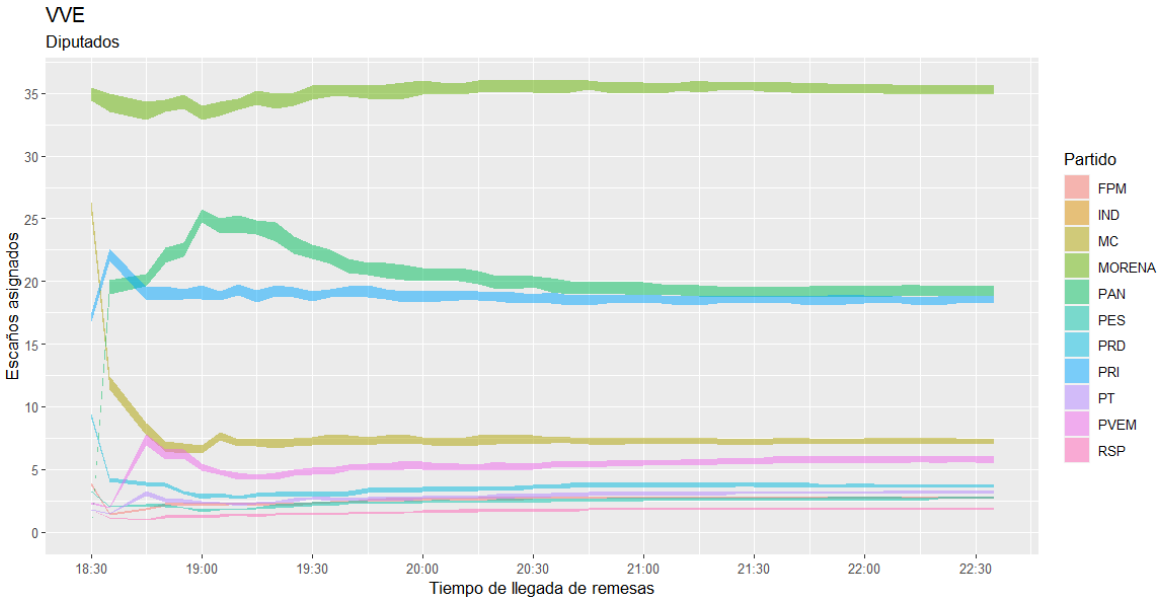
Figura 7.11: Gráfico de las estimaciones de los intervalos de la Conformación (Ecuación 3.1) de las 18:30 a la hora indicada del día de la Jornada Electoral. Un gráfico ampliado de esto se puede ver en la Figura 9.3a.

los partidos coaligados con MORENA que son el PVEM y el PT pero sin sobrepasar a las otras dos potencias políticas también coaligadas (ver Tabla 6.6).

En la Figura 7.12 podemos ver el avance en el porcentaje de *Votación Válida Emitida* (VVE) el cual es de suma importancia ya que basándonos en esto y de acuerdo a lo mencionado en la Subsubsección 3.2.2.3 si algún partido cae debajo del 3 % en este porcentaje perderá su registro a nivel nacional. De tal manera que para los partidos con poca relevancia política este dato es de suma importancia porque estará indicando cuáles partidos se mantienen vigentes y cuáles deberán salir que son usualmente



(a) 21:25

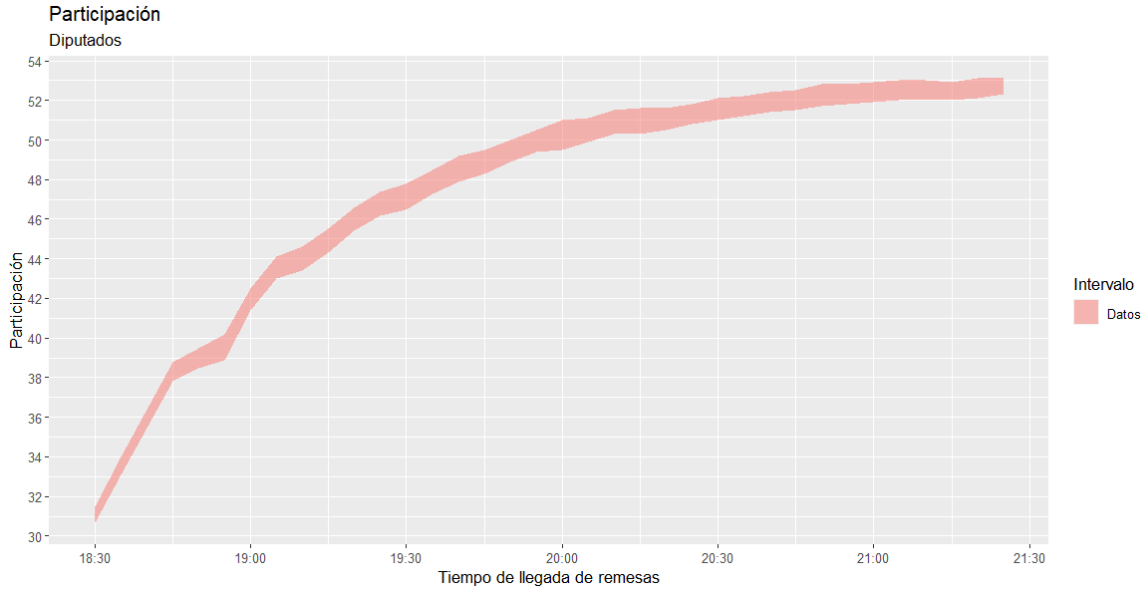


(b) 22:35

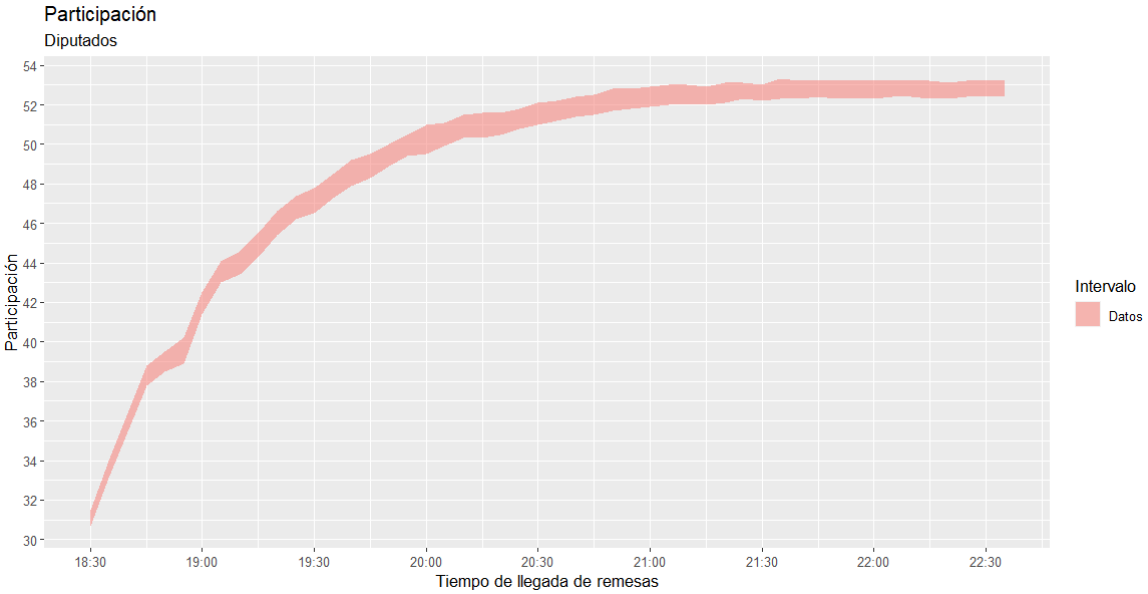
Figura 7.12: Gráfico de las estimaciones de los intervalos del porcentaje de *Votación Válida Emitida* (VVE) (Ecuación 3.8) de las 18:30 a la hora indicada del día de la Jornada Electoral. Un gráfico ampliado de esto se puede ver en la Figura 9.3b.

partidos con muy poca antigüedad. En los resultados finales (Subsección 7.3.3) mencionaremos cuáles fueron los pronósticos del Conteo Rápido sobre los partidos que estarían perdiendo su registro en estas elecciones.

De la Figura 7.13 se observan las predicciones para la Participación ciudadana, que en este caso, estos intervalos no hicieron más que crecer de manera logarítmica. Esto



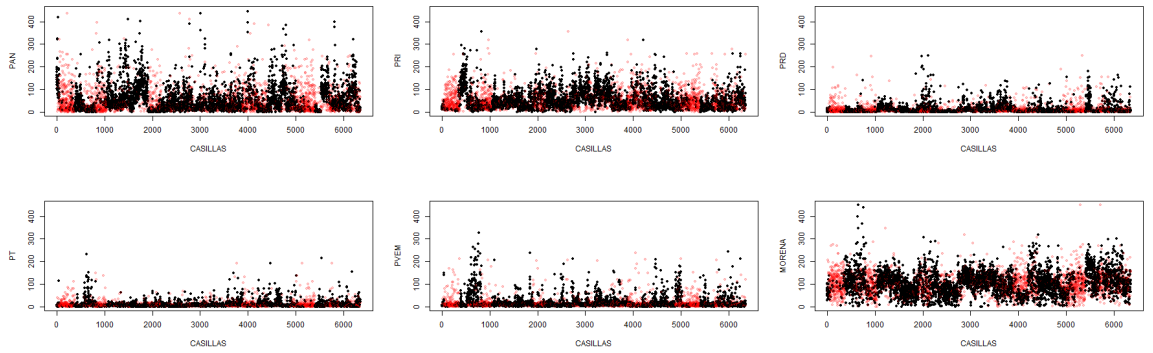
(a) 21:25



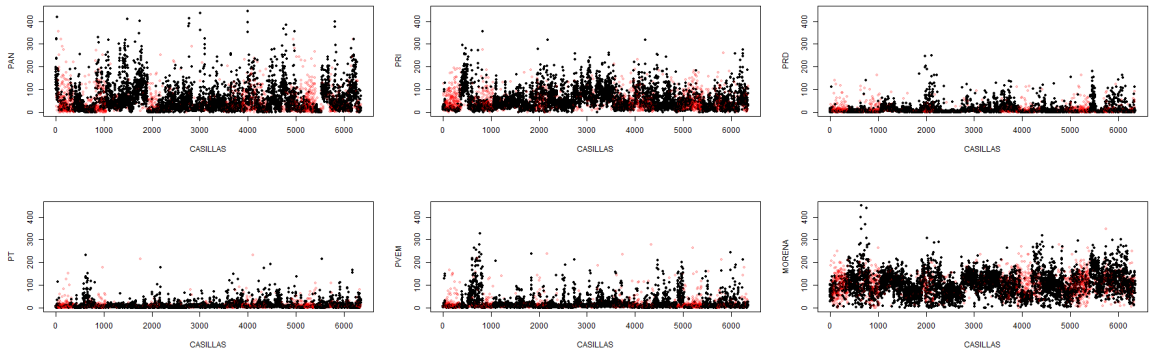
(b) 22:35

Figura 7.13: Gráfico de las estimaciones de los intervalos de la Participación (Ecuación 5.3) de las 18:30 a la hora indicada del día de la Jornada Electoral. Un gráfico de las 20:00 hrs. de esto se puede ver en la Figura 9.4.

daba pie a que en efecto la participación ciudadana fuera una de las más importantes históricamente hablando, pues ya cercanas las 20:00 horas se empezaba a predecir que la participación estaría arriba del 50 % que era algo nunca antes visto y reflejaba tanto soberanía nacional como el desarrollo de la ejecución del sufragio por parte de la población.



(a) 21:25



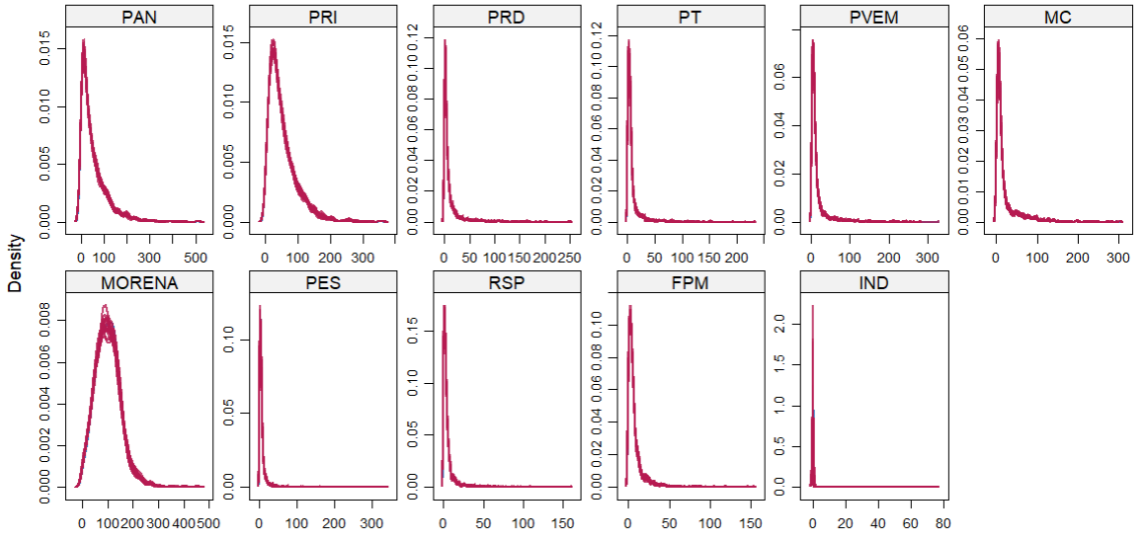
(b) 22:35

Figura 7.14: Gráfico de las casillas imputadas en el proceso de Imputación Múltiple (ver Subsección 6.2.2) a la hora indicada del día de la Jornada Electoral. Un gráfico de las 20:00 hrs. de esto se puede ver en la Figura 9.5b.

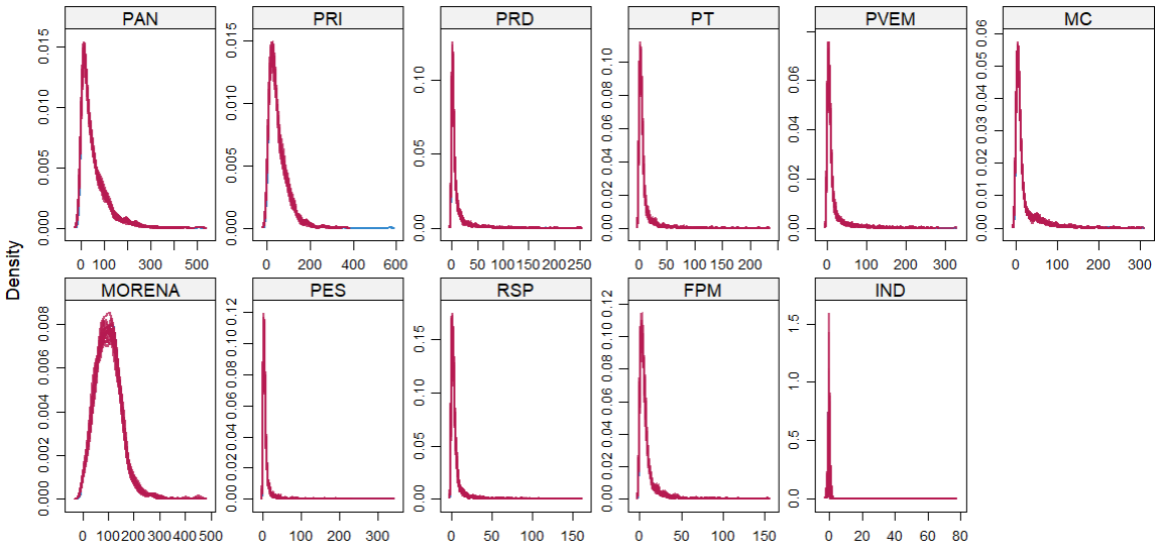
Por otro lado, en las Figuras 7.14 y 7.15 se muestra lo referente al sesgo causado por el proceso de Imputación Múltiple. En este caso y a diferencia de los gráficos anteriores, éstos no están conteniendo del todo a los anteriores, ya que van cambiando su comportamiento poco a poco a cada hora y por lo mismo aquí se muestra un mayor contraste con respecto a las Figuras 9.5b y 9.5a.

Tomando como ejemplo el gráfico de las casillas imputadas (Figuras 7.14 y 9.5b) estos gráficos muestran la cantidad de votos que estaban siendo imputados (marcados de rojo) con base en las remesas (marcadas de negro) para cada uno de los partidos políticos mostrados. Se observa que entre más temprana la hora el día de la jornada electoral, las casillas imputadas son más que las que las de horas más tarde lo cual refleja la llegada de la muestra y que el comportamiento realizado por las imputaciones parecía ser razonable. Además, los gráficos relacionados con la densidad de las imputaciones (Figuras 7.15 y 9.5a) siempre parecían tener el mismo comportamiento unimodal, de tal manera que no estaba existiendo un sesgo con las estimaciones realizadas.

Estos fueron los gráficos que mostraba la aplicación *Shiny* que funcionaban para



(a) 21:25



(b) 22:35

Figura 7.15: Gráfico de las densidades de las estimaciones por Imputación Múltiple (ver Subsección 6.2.2) para observar el sesgo a la hora indicada del día de la Jornada Electoral. Los puntos rojos indican imputaciones y los negros observaciones de la cantidad de votos en la casilla indicada. Un gráfico ampliado de esto se puede ver en la Figura 9.5a.

supervisar el comportamiento de los resultados que se estaban teniendo a lo largo de la tarde de la Jornada Electoral. Estos resultados eran al mismo tiempo mandados por red INE a Cartografía (Subsubsección 7.3.2.2) para que mostrara a los demás miembros del COTECORA el comportamiento de las estimaciones que se estaba teniendo en tiempo real. Finalmente, **a las 22:35 de la noche se terminó de realizar las estimaciones para que fueran publicadas por cadena nacional** (Subsección 7.3.3).

7.3.2.2. Resultados finales de Cartografía

A lo largo de la jornada electoral un equipo dentro del búnker se encargaba de recibir los resultados e ir proyectando a todos los miembros de COTECORA la llegada de las remesas a nivel nacional. De igual manera, a través de la red INE, podía ser consultado también en las computadoras personales de los integrantes del COTECORA y por medio de la cual estarían compartiendo las estimaciones realizadas. Con la ventaja de que dentro de esta red se podían explorar a voluntad las estimaciones que cada equipo del COTECORA estaba realizando.

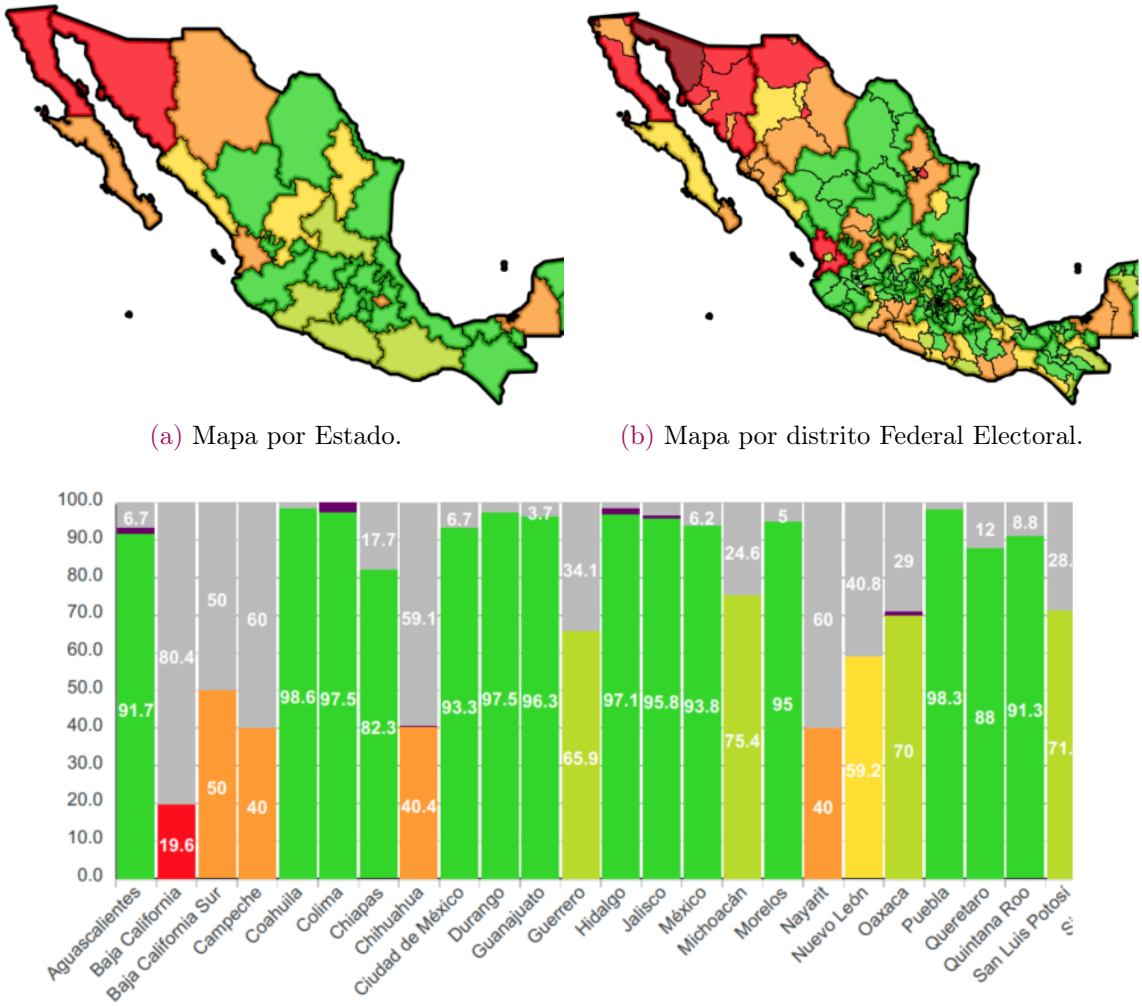
El funcionamiento de estos gráficos fue realizado por un equipo técnico del INE a petición de diseño del COTECORA en algunas de las sesiones que se tienen previo al día de la Jornada Electoral.

En la Figura 7.16 se muestra lo que Cartografía estaba proyectando a todos los integrantes del COTECORA, esto es un Mapa geopolítico de México que mostraba al país en 7.16a dividido por estados y en 7.16b dividido por distrito federal electoral. Recordando que la selección de la muestra (Subsección 5.6.1) se hizo precisamente con base en los distritos federales electorales que a su vez son una partición de cada estado. Este mapa comenzó de color rojo, lo cuál indicaba que aún no se contaba con una cantidad suficiente de muestra y poco a poco fue cambiando de color hasta que se tornó a los que se pueden apreciar, lo cual, como se indica en 7.16c, es el avance de la muestra recibida en cada uno de los estados, hablado en cuanto cantidad de remesas. Este es un tema relevante porque desde que se seleccionó el tamaño de muestra (Sección 5.3) era importante detectar el efecto del horario en México.

La Figura 7.17 muestra los distintos husos horarios que tiene México. Tomando como referencia el "Tiempo de Centro" que es donde se encuentra la Ciudad de México y, por ende, la ubicación de los miembros del COTECORA, así como también el horario con mayor terreno en el país, los estados que se encuentran al noroeste de México tienen un horario más temprano, lo cual implica que habrá un retraso en la llegada de las remesas de estos estados debido a este efecto.

Es por esto que fue necesario, ya de manera anticipada, fijar la cantidad de remesas en estos estados. De hecho, este efecto sí se ve reflejado al momento en que se terminó de hacer las estimaciones para la publicación de los resultados a las 22:35 del día de la Jornada Electoral (Figura 7.16). Fue debido a este retraso de las remesas en los estados al noroeste del país, que el Conteo Rápido tuvo que terminar a esta hora.

Por parte de las estimaciones correspondientes a la conformación de la cámara de diputados y la participación ciudadana, los gráficos de las estimaciones se pueden ver en la Figura 7.18. Donde se ve el compulsado de las estimaciones por parte de los equipos del Dr. Rodríguez y el Dr. Nieto (Sección 5.6) y que es, de hecho, uno de los resultados que se desplegaban a través de la red INE por parte de Cartografía.



(a) Mapa por Estado.

(b) Mapa por distrito Federal Electoral.

(c) Gráfico de barras por Estado.

Figura 7.16: Fragmento del mapa y gráfico de barras proporcionado por Cartografía a las 22:25 hrs. que mostraba la llegada de las remesas dividido geopolíticamente. La escala de color medía el arribo de información e iba de rojo (muy poco), naranja (poco), amarillo (intermedio), lima (casi suficiente) y verde (suficiente).

En particular, la Figura 7.18a mostraba el compulsado de la conformación de la cámara de diputados, donde se unían las metodologías que cada equipo estaba realizando en el intervalo. En la Figura 7.18b se muestra el compulsado del porcentaje de VVE análogo al anterior.

En estos dos últimos gráficos se ven los intervalos de estas estadísticas dada la remesa más actual, es decir, con la mayor cantidad de información disponible. Finalmente en la Figura 7.18c se muestra el intervalo de la participación ciudadana a diferentes horas de las remesas.

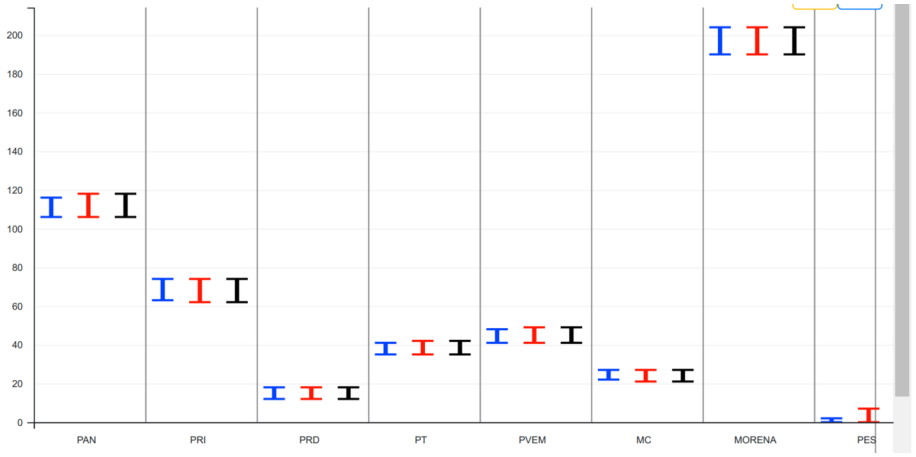
Figura 7.17: Husos horarios de México. Dependiendo de la época del año las horas pueden variar en cada uno de estos.

En todos los casos, se puede apreciar que existe una convergencia a los mismos resultados por parte de ambos equipos. Es decir, se puede apreciar que los intervalos de color azul y rojo que fueron los dados por ambos equipos de manera individual y unificados en este compulsado, tienen un comportamiento similar a pesar de que cada equipo utilizó diferentes enfoques estadísticos para las estimaciones. Sin embargo, y con el objetivo de tener una certeza mayor a las estimaciones, lo que se reporta finalmente son los intervalos de color negro que es precisamente la unión de los intervalos dada por cada equipo.

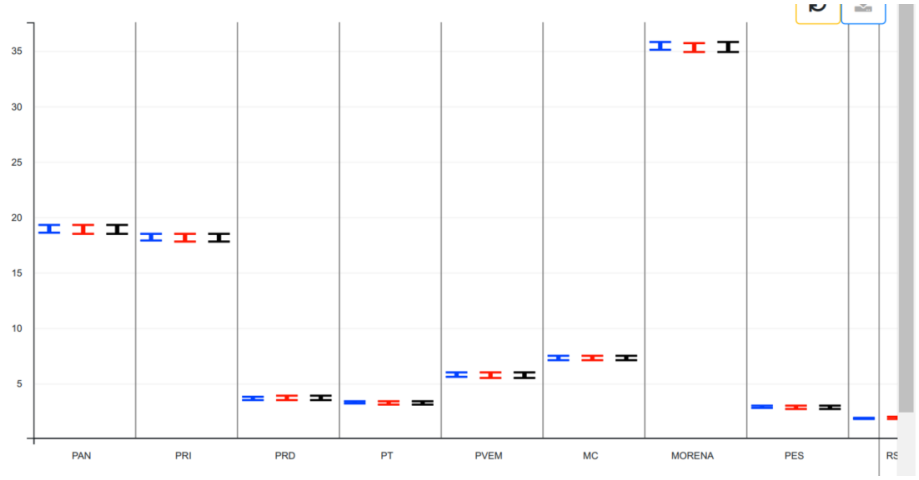
No está demás mencionar que para estas elecciones los resultados giraban alrededor de los partidos más populares, los cuales, presentaban resultados contrastantes contra los que no lo son. En este caso MORENA se mantuvo prácticamente todo el tiempo en la delantera y de hecho resultó el partido con un intervalo que reflejaba una clara ventaja contra los demás partidos políticos. Asimismo, los partidos PRI y PAN también se mostraban en la delantera con respecto al resto de los partidos y con intervalos bastante similares (ver Figuras 7.18a y 7.18b).

7.3.3. Últimos momentos del Conteo Rápido

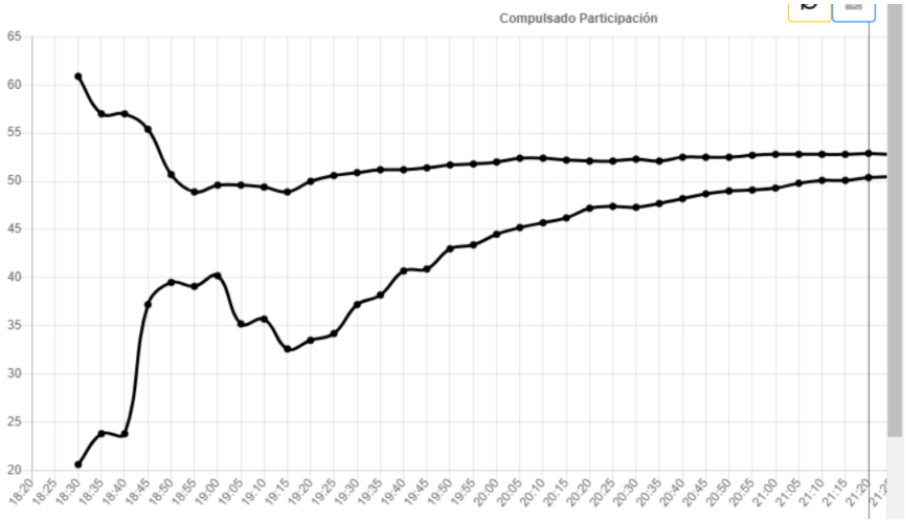
Dieron entonces las 22:35 hrs. de la hora centro del 6 de junio del 2021 y con la información de esta remesa, el COTECORA mandó a imprimir el informe final del Conteo Rápido para las elecciones federales de este año.



(a) Escaños asignados. (3.1)



(b) Porcentaje de VVE (3.8)



(c) Participación (5.3)

Figura 7.18: Fragmento de los gráficos proporcionados por cartografía a las 22:35 de compulsado (Sección 5.6) de la Conformación de la Cámara de Diputados por partido político y Participación Ciudadana (Subsección 3.2.5). Las líneas de colores azules y rojas eran dadas de manera individual por los equipos del Dr. Rodríguez y el Dr. Nieto, mientras que las de color negro eran ya el compulsado de sus metodologías.

Partido político / Candidatos Independientes		Porcentaje de votación válida emitida		Número de Diputados	
		Mínimo	Máximo	Mínimo	Máximo
PAN	Partido Acción Nacional	18.5	19.3	106	117
PRI	Partido Revolucionario Institucional	17.8	18.5	63	75
PRD	Partido de la Revolución Democrática	3.5	3.9	12	21
VERDE	Partido Verde Ecologista de México	5.5	6.0	40	48
PT	Partido del Trabajo	3.1	3.5	35	41
MC	Movimiento Ciudadano	7.1	7.5	20	27
morena	Morena	34.9	35.8	190	203
PES	Partido Encuentro Solidario	2.7	3.0	0	6
RSP	Redes Sociales Progresistas	1.8	2.0	0	0
FXM	Fuerza por México	2.6	2.8	0	0
	Candidatos Independientes	0.1	0.3	0	0

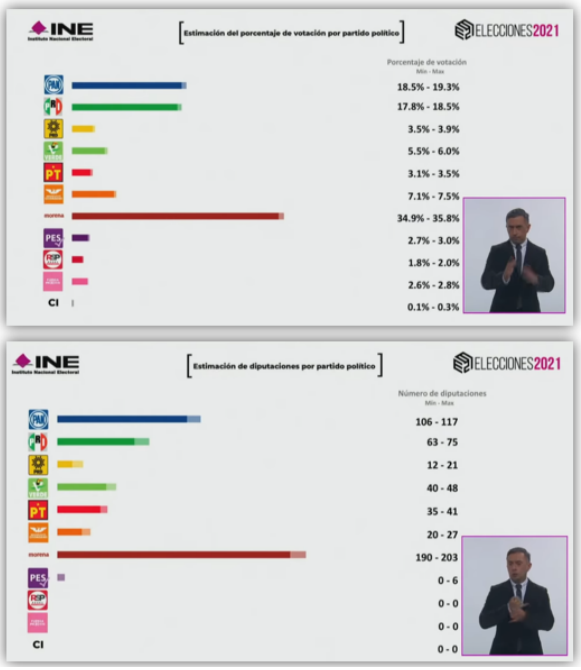


Figura 7.19: Informe final que el *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA) proporcionó al *Instituto Nacional Electoral* (INE) como resultados del Conteo Rápido de las Elecciones Federales del año 2021. La publicación de estos resultados fue un evento público y puede ser visto dando clic aquí.

El informe final se puede encontrar dando clic aquí[7]. En este documento se especifican diversos detalles sobre las estimaciones realizadas. Entre lo que más destaca es:

1. De las 6345 casillas que integran la muestra, se recibió información de 5040 casillas, las cuales representan el 79.3 % de la muestra total.

2. De los 300 estratos considerados para definir el diseño muestral (Sección 5.6), se contó con información de todos.

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

3. La participación ciudadana se encuentra entre 51.7 % y 52.5 %.

4. Las estimaciones para el porcentaje de *Votación Válida Emitida* (VVE) y Conformación de la Cámara de Diputados se presentan en la Figura 7.19.

De estos resultados, se pronosticaba desde estos momentos que partidos como PES, RSP y FM (Tabla 6.6) estarían perdiendo su registro. Por otro lado, en términos de Coaliciones, la mayoría de la Cámara de Diputados ya no estaría conformada por

[7] https://github.com/A1arcon/R_Actuarial/tree/main/Conteo%20R%C3%A1pido%20(INE)/7.%20Jornada%20Electoral/Informe_conteo_rapido_diputados_2021.pdf

partidos de la coalición "Juntos Haremos Historia", sin embargo tendrían aún una gran cantidad de curules asignados.

Apoyado de este informe, durante la noche del día de la Jornada Electoral, el Consejero Presidente de INE, el Dr. Lorenzo Córdova Vianello dio a conocer los resultados de las estimaciones. Este fue un evento transmtido en vivo en cadena nacional y puede ser consultado dando clic aquí[8].

[8]https://www.youtube.com/watch?v=eQfz5pVP-aU

Figura 7.20: Fotografía tomada al final del día de la Jornada Electoral de todos los integrantes del *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA).

Capítulo 8

Conclusiones

Las estimaciones realizadas el día de la Jornada Electoral (Capítulo 7) utilizaron toda la teoría desarrollada en esta tesis. Comenzando por la implementación desde la ley hasta su forma teórica y computacional del cálculo de la Conformación de la Cámara de Diputados (Capítulo 3). Luego, pasando con el diseño de la muestra que se seleccionó para realizar las estimaciones (Capítulo 5) y cuya justificación está cimentada por la teoría del muestreo probabilístico (Capítulo 4). Tomando a consideración que el día de la Jornada Electoral, la muestra llegaría poco a poco, y, utilizando la ya observada, se realizaría un proceso de Imputación Múltiple (Capítulo 6).

Los resultados finales del proceso de asignación de diputados para las elecciones federales del año 2021, no eran conocidas el día de la Jornada Electoral, 6 de Junio del 2021, cuando se publicaron las estimaciones. Fue hasta el 29 de Agosto del 2021 que el sitio web de la Cámara de Diputados dio a conocer los resultados finales de la Conformación de la Cámara, en un reporte que puede ser consultado dando clic aquí[1].

Estos resultados finales se pueden comparar contra las estimaciones mostradas en el informe final que publicó el INE el mismo día de la Jornada Electoral (Subsección 7.3.3) en la Figura 8.1. Donde se puede apreciar que todos los intervalos dados por el COTECORA contuvieron al valor real. Esto significa que el ejercicio estadístico y la teoría desarrollada para lograrlo fueron correctos y, en este caso, cumplieron plenamente el propósito de dar a conocer una idea de cómo estaría conformada la Cámara de Diputados de manera oportuna. Esto incluye el caso de los partidos con menor antigüedad como "Partido Encuentro Solidario", "Redes Sociales Progresistas" y "Fuerza por México", donde se estimaba que perderían su registro y además no tendrían ni un escaño asignado en la Cámara de diputados.

En conclusión, esta tesis expone y justifica una metodología que fue puesta en práctica para el Conteo Rápido de las elecciones federales del año 2021 con el propósito de dar un pronóstico oportuno de la conformación de la cámara de diputados. Misma que, en este caso, brindó resultados con una precisión efectiva del 100 % y por lo mismo,

[1] https://comunicacionnoticias.diputados.gob.mx/comunicacion/index.php/boletines/informa-secretaria-general-la-composicion-inicial-de-la-lxv-legislatura-de-la-camara-de-diputados-gsc.tab=0

demuestra su validez teórica en la fase de experimentación del método científico. Cabe mencionar que este es un proceso que resulta ser útil para todo México en el sentido político, y que da tanto credibilidad como certidumbre a los procesos democráticos.

Partido político / Candidatos Independientes		Porcentaje de votación válida emitida		Número de Diputados	
		Mínimo	Máximo	Mínimo	Máximo
PAN	Partido Acción Nacional	18.5	19.3	106	117
PRI	Partido Revolucionario Institucional	17.8	18.5	63	75
PRD	Partido de la Revolución Democrática	3.5	3.9	12	21
VERDE	Partido Verde Ecologista de México	5.5	6.0	40	48
PT	Partido del Trabajo	3.1	3.5	35	41
	Movimiento Ciudadano	7.1	7.5	20	27
morena	Morena	34.9	35.8	190	203
PES	Partido Encuentro Solidario	2.7	3.0	0	6
RSP	Redes Sociales Progresistas	1.8	2.0	0	0
FUERZA MÉXICO	Fuerza por México	2.6	2.8	0	0
	Candidatos Independientes	0.1	0.3	0	0

Figura 8.1: Comparativa del informe final que el *Comité Técnico Asesor de los Conteos Rápidos* (COTECORA) proporcionó al *Instituto Nacional Electoral* (INE) como resultados del Conteo Rápido de las Elecciones Federales del año 2021 (izquierda) contra los resultados reales obtenidos una vez finalizado el proceso de elección (derecha).

Capítulo 9

Anexos

9.1. Detalles matemáticos

Esta sección está destinada para mostrar algunas pruebas matemáticas de diversos resultados que se mencionan a lo largo de esta tesis, así como brindar más detalles sobre algunos de los temas mencionados a lo largo de este documento.

9.1.1. Intervalos de Confianza para la función de distribución acumulada empírica

Nótese que por construcción de (4.26) vista como variable aleatoria, se tiene que

$$nF_n(x) = \sum_{i=1}^{n} \mathbb{1}\left(X_i \leq x\right) \sim Binomial(n, p = \mathbb{P}\left[X_i \leq x\right] = F(x)).$$

De donde se sigue que

$$\begin{aligned} \mathbb{E}\left[F_n(x)\right] &= F(x), \text{ y,} \\ V(F_n(x)) &= \frac{F(x)(1-F(x))}{n} \underset{n\uparrow\infty}{\longrightarrow} 0. \end{aligned}$$

En [11], se enuncia dos resultados importantes que tiene la ECDF el primero es derivado de una aplicación del Teorema de Glivenko-Cantelli y es que[1]

$$\underset{x}{sup}|F_n(x) - F(x)| \xrightarrow{P} 0.$$

El segundo resultado sirve para construir intervalos de confianza para la ECDF y es un resultado del la Desigualdad de Dvoretzky-Kiefer-Wolfowitz; considerando cualquier $\varepsilon > 0$,

[1]Siendo precisos, $\underset{x}{sup}|F_n(x) - F(x)|$ converge a 0 casi seguramente.

$$\mathbb{P}\left[\sup_x |F_n(x) - F(x)| > \varepsilon\right] \leq 2e^{-2n\varepsilon^2}. \tag{9.1}$$

De donde, como resultado de (9.1), se puede construir un intervalo de confianza no paramétrico del $(1-\alpha)\times 100\,\%$ para cualquier CDF F como se muestra a continuación:

$$\mathbb{P}\left[L(x) \leq F(x) \leq U(x), \forall x\right] \geq 1 - \alpha, \tag{9.2}$$

donde

$$L(x) = \text{máx}\{F_n(x) - \varepsilon_n, 0\}, \quad U(x) = \text{mín}\{F_n(x) + \varepsilon_n, 1\}, \quad \text{y}, \quad \varepsilon_n = \sqrt{\frac{1}{2n} log\left(\frac{2}{\alpha}\right)}.$$

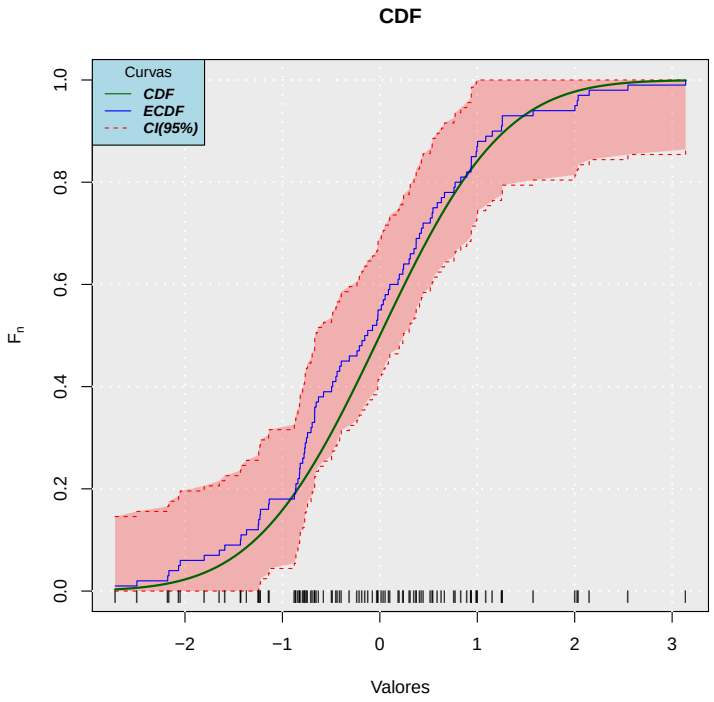


Figura 9.1: Aplicación de (4.26) como la ECDF de 100 simulaciones de una distribución Normal Estándar y su Intervalo de Confianza no-paramétrico dado por (9.2).

En la Figura 9.1 vemos un ejemplo aplicado de (4.26) y (9.2). En la Subsección 9.2.1 se muestra el código para crear esta gráfica en lenguaje R.

9.1.2. Ejemplo de Intervalos de Confianza Bootstrap

Una aplicación de estas metodologías para obtener intervalos de confianza bootstrap (Subsubsección 4.3.2.3) aplicando el algoritmo visto en la Subsubsección 4.3.2.2, la podemos ver en un ejemplo donde se simularán datos que siguen una distribución normal y se buscará encontrar un intervalo al $(1-\alpha)\times 100\,\% = 95\,\%$ de confianza para la mediana de estos datos. La implementación en código de R se puede encontrar en la Sección 9.2 y los resultados de este experimento se pueden visualizar en la Figura 9.2, donde se observa un histograma de las estadísticas $T^*_{n,b}$ con $b \in \{1,\ldots,B\}$ y los intervalos de confianza obtenidos por las metodologías en cuestión. Adicionalmente, en la Tabla 9.1, se reportan los valores numéricos de la mediana muestral original (T_n) y los intervalos de confianza correspondientes a la Figura 9.2.

Estadística/Intervalo	Valor(es)
Mediana (T_n)	**0.016**
Normal	**(0.014 , 0.019)**
Pivotal	**(0.014 , 0.018)**
Percentil	**(0.015 , 0.019)**

Tabla 9.1: Valores de las estadísticas mostradas en la Figura 9.2 correspondientes al ejemplo computacional aplicado de la estimación de la mediana de simulaciones de una Normal Estándar.

9.2. Códigos de R

Esta sección está destinada para mostrar los códigos en lenguaje R necesarios para la elaboración de gráficas e implementaciones computacionales para este documento. Versiones generalizadas de muchos de los códigos mostrados a continuación y otros tantos complementarios, se podrán encontrar en el GitHub [2] del autor de este documento.

9.2.1. Programación de la función de distribución acumulada y función de distribución acumulada empírica

Código para la creación de la CDF y la ECDF con intervalos de confianza en la Figura 9.1. Este código está inspirado en una construcción hecha por SwampThing-Paul/AnalystHelper.

[2] https://github.com/A1arcon/R_Actuarial

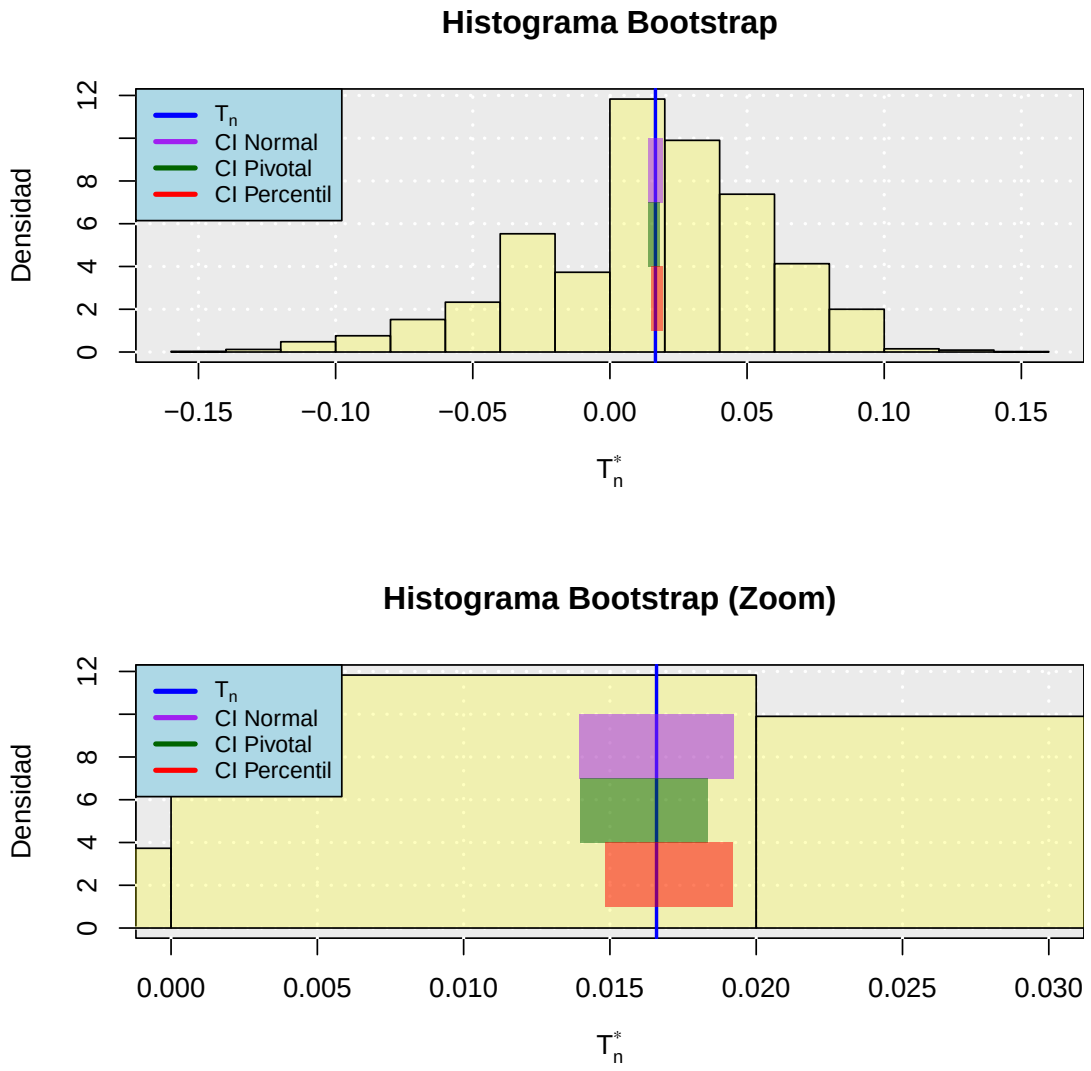


Figura 9.2: Histograma de bootstrap con intervalos de confianza (CI) por diferentes métodos para estimar la mediana (T_n) de una muestra aleatoria de observaciones Normales Estándar.

```r
# Implementación de la función de distribución empírica
ecdf_fun=function(x,CI=TRUE,CI.interval=0.95){
  x <- sort(x)
  n <- length(x)
  vals <- unique(x)
  rval <- approxfun(vals, cumsum(tabulate(match(x, vals)))/n,
                    method = "constant", yleft = 0,
                    yright = 1, f = 0, ties = "ordered")
  class(rval) <- c("ecdf", "stepfun", class(rval))
  assign("nobs", n, envir = environment(rval))
```

```
  attr(rval, "call") <- sys.call()
  rval
  x.val=environment(rval)$x
  y.val=environment(rval)$y

  if(CI==TRUE){
    alpha=1-CI.interval
    eps=sqrt(log(2/alpha)/(2*n))
    ll=pmax(y.val-eps,0)
    uu=pmin(y.val+eps,1)
    return(data.frame(value=x.val,proportion=y.val,
                      lwr.CI=ll,upr.CI=uu))
  }else{
    return(data.frame(value=x.val,proportion=y.val))
  }
}
# Datos iniciales
set.seed(2012)
sim <- rnorm(100)
Fn <- ecdf_fun(sim)
# Formato del gráfico
plot(Fn$proportion~Fn$value,ylim=c(0,1),
     ylab=latex2exp::TeX("$F_n$"),xlab="Valores",
     main = "CDF")
# Fondo
rect(par("usr")[1], par("usr")[3],
     par("usr")[2], par("usr")[4],
     col = "#ebebeb")
grid(col="white",lwd=2)
# Gráfico de los intervalos de confianza
with(Fn,polygon(c(value,rev(value)), c(lwr.CI,rev(upr.CI)),
                col = adjustcolor("red",alpha.f=0.25) ,
                border=NA))
with(Fn,lines(lwr.CI~value,type="s",lty=2,col="red"),lwd=2)
with(Fn,lines(upr.CI~value,type="s",lty=2,col="red"),lwd=2)
#Gráfico de la CDF
curve(expr = pnorm(x),col="darkgreen",add = TRUE,lwd=2)
# Gráfico de la ECDF
with(Fn,lines(proportion~value,type="s",col="blue"))
# Observaciones
with(Fn,points(rep(-0.01,length(value))~value,
               pch="|",lty=2,col="black",cex=0.75))
# Leyenda
legend("topleft", legend=c("CDF", "ECDF","CI(95%)"),
```

```
       col=c("darkgreen", "blue","red"), lty=c(1,1,2), cex=0.8,
       title="Curvas", text.font=4, bg='lightblue')
```

9.2.2. Intervalos de confianza bootstrap

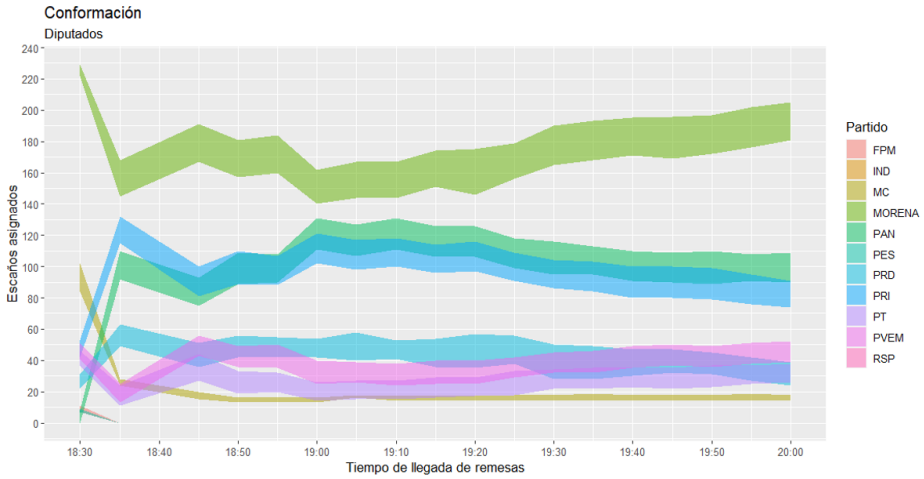
Código de la Subsección 9.1.2 para la creación del ejemplo mostrado en la Figura 9.2 de un histograma para estimar la mediana de observaciones Normales Estándar vía bootstrap.

```
# Implementación del algoritmo con datos similados
set.seed(20)
X <- rnorm(1000)    # Datos simulados
Tn <- median(X)      # Mediana real de los datos
B = 5000           # Repeticiones bootstrap
# Función que calcula Tn*
Tstar <- function(datos,i){median(datos[i])}
# Procedimiento bootstrap
Tboot <- boot::boot(data = X,statistic = Tstar,R = B)$t[,1]
# Error estándar
se_boot <- sqrt(var(Tboot))
# Nivel de confianza para los intervalos
alpha <- 0.95
# Intervalo Normal
CI_Nor <- Tn + qnorm(1-alpha/2)*se_boot*c(-1,1)
# Intervalo Pivotal
CI_Piv <- 2*Tn-quantile(Tboot,probs = c(1-alpha/2,alpha/2))
# Intervalo por Percentiles
CI_Per <- quantile(Tboot,probs = c(alpha/2,1-alpha/2))
# Dividimos en dos gráficos
par(mfrow=c(2,1))
# Gráfico 1 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# Gráfico de la simulaciones
hist(Tboot,probability = TRUE,
     main="Histograma Bootstrap",
     ylab="Densidad",xlab = latex2exp::TeX("$T^*_{n}$"))
# Fondo
rect(par("usr")[1], par("usr")[3],
     par("usr")[2], par("usr")[4],
     col = "#ebebeb")
grid(col="white",lwd=2)
# Histograma
hist(Tboot,
     col=adjustcolor("yellow",alpha.f=0.25),
     probability = TRUE,
     add=TRUE)
abline(v = Tn, col = "blue",lwd=2)
polygon(y=c(7, 10,  10, 7),x= rep(CI_Nor,each=2),
        col=adjustcolor("purple",alpha.f = 0.5),
```

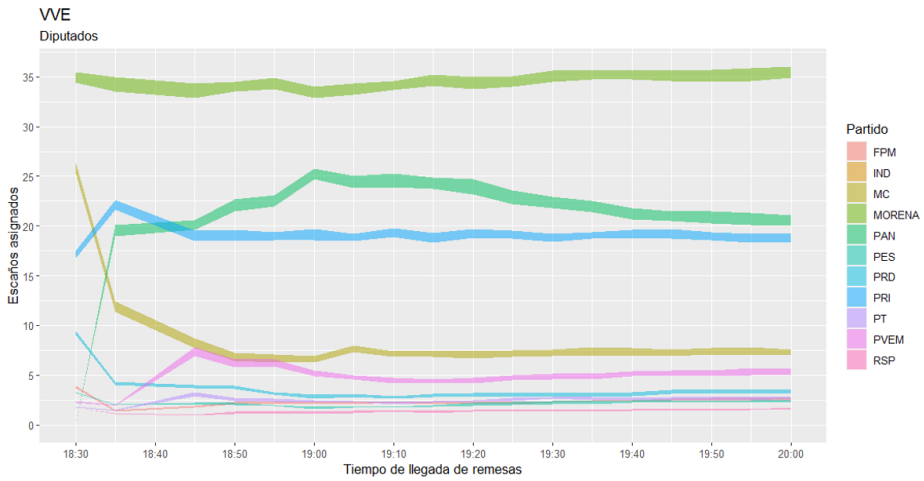
```
        border=NA)
polygon(y=c(4, 7,  7, 4),x= rep(CI_Piv,each=2),
        col=adjustcolor("darkgreen",alpha.f = 0.5),
        border=NA)
polygon(y=c(1, 4,  4, 1),x= rep(CI_Per,each=2),
        col=adjustcolor("red",alpha.f = 0.5),
        border=NA)
# Leyenda
legend("topleft", legend=c(latex2exp::TeX("$T_n$"),
                           "CI Normal",
                           "CI Pivotal",
                           "CI Percentil"),
       col=c("blue", "purple","darkgreen","red"),
       cex=0.8, lwd=3,
       text.font=4, bg='lightblue')
# Gráfico 2 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# Gráfico de la simulaciones
hist(Tboot,probability = TRUE,
     main="Histograma Bootstrap (Zoom)",
     xlim = c(0,0.03),
     ylab="Densidad",xlab = latex2exp::TeX("$T^*_{n}$"))
# Fondo
rect(par("usr")[1], par("usr")[3],
     par("usr")[2], par("usr")[4],
     col = "#ebebeb")
grid(col="white",lwd=2)
# Histograma
hist(Tboot,
     col=adjustcolor("yellow",alpha.f=0.25),
     probability = TRUE,
     add=TRUE)
abline(v = Tn, col = "blue",lwd=2)
polygon(y=c(7, 10,  10, 7),x= rep(CI_Nor,each=2),
        col=adjustcolor("purple",alpha.f = 0.5),
        border=NA)
polygon(y=c(4, 7,  7, 4),x= rep(CI_Piv,each=2),
        col=adjustcolor("darkgreen",alpha.f = 0.5),
        border=NA)
polygon(y=c(1, 4,  4, 1),x= rep(CI_Per,each=2),
        col=adjustcolor("red",alpha.f = 0.5),
        border=NA)
# Leyenda
legend("topleft", legend=c(latex2exp::TeX("$T_n$"),
                           "CI Normal",
```

```
                          "CI Pivotal",
                          "CI Percentil"),
       col=c("blue", "purple","darkgreen","red"),
       cex=0.8, lwd=3,
       text.font=4, bg='lightblue')
```

9.3. Gráficos del Conteo Rápido



(a) $CONF$ en la Ecuación 3.1.



(b) $pVVE$ en la Ecuación 3.8.

Figura 9.3: Gráfico de las estimaciones de los intervalos de la Conformación y el porcentaje de *Votación Válida Emitida* (VVE) de las 18:30 a las 20:00 horas el día de la Jornada Electoral. Si se desea ver el desarrollo de estos gráficos a lo largo del tiempo pueden verse las Figuras 7.11 y 7.12.

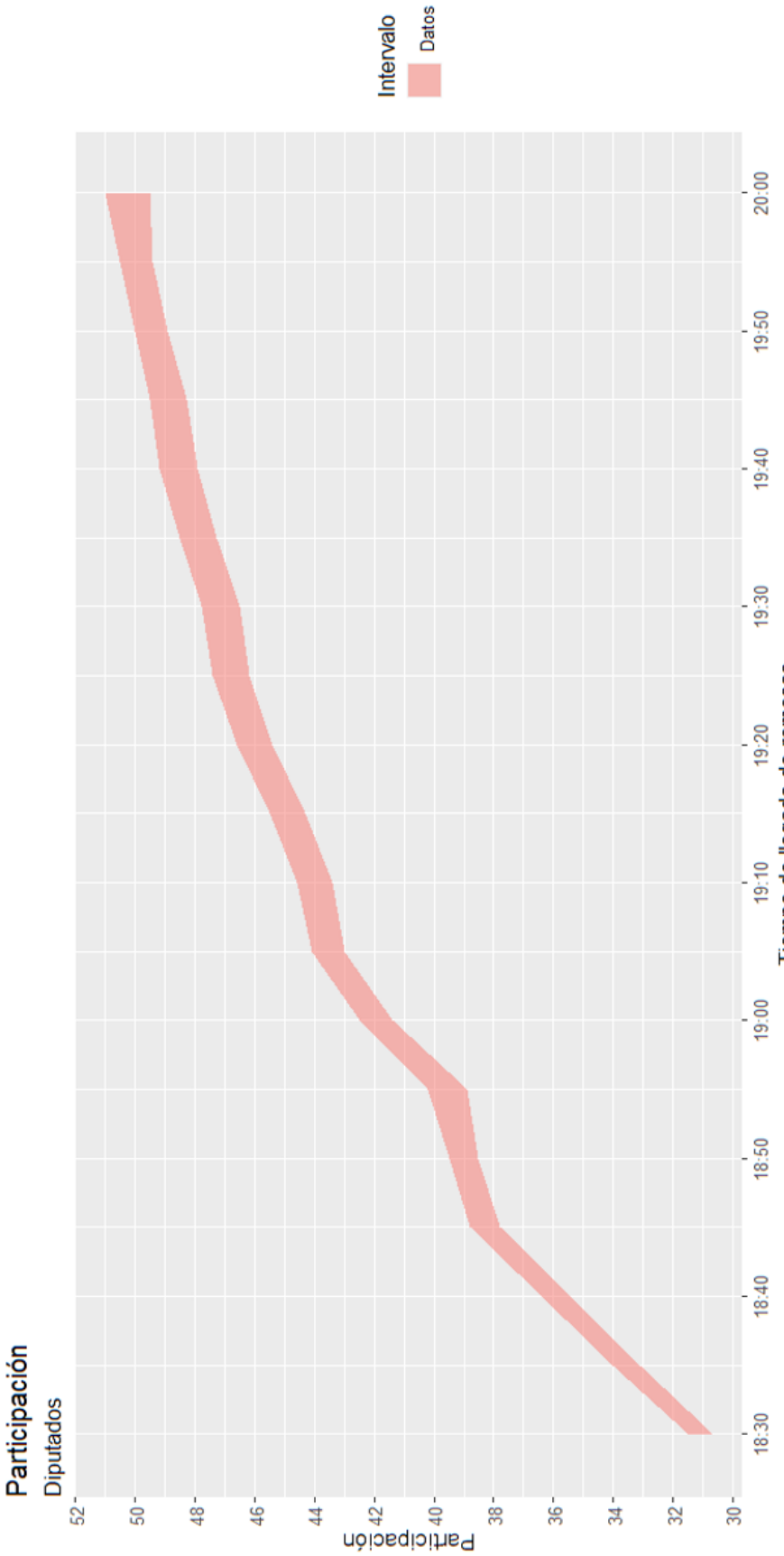
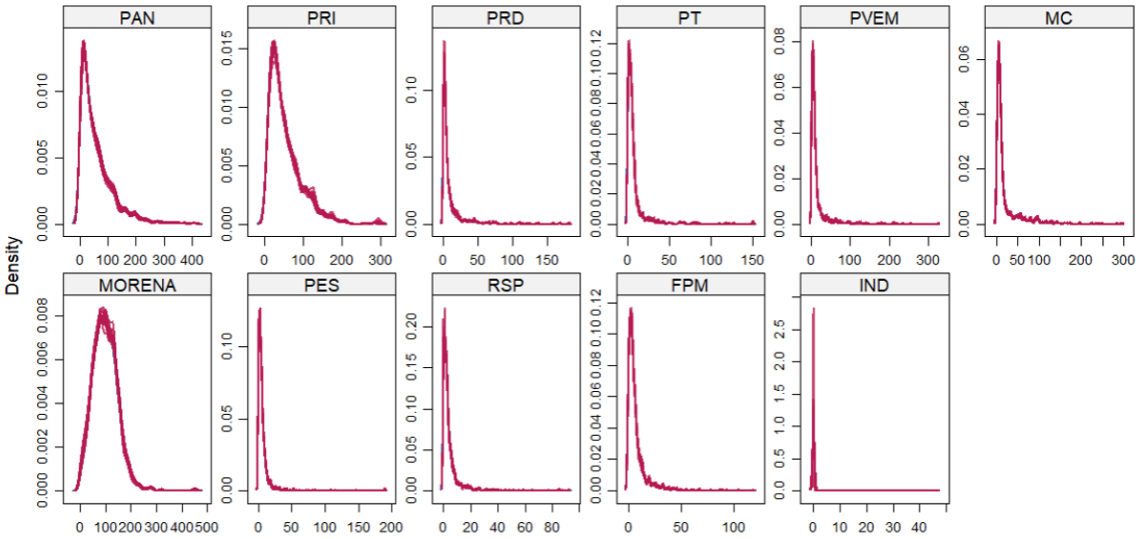
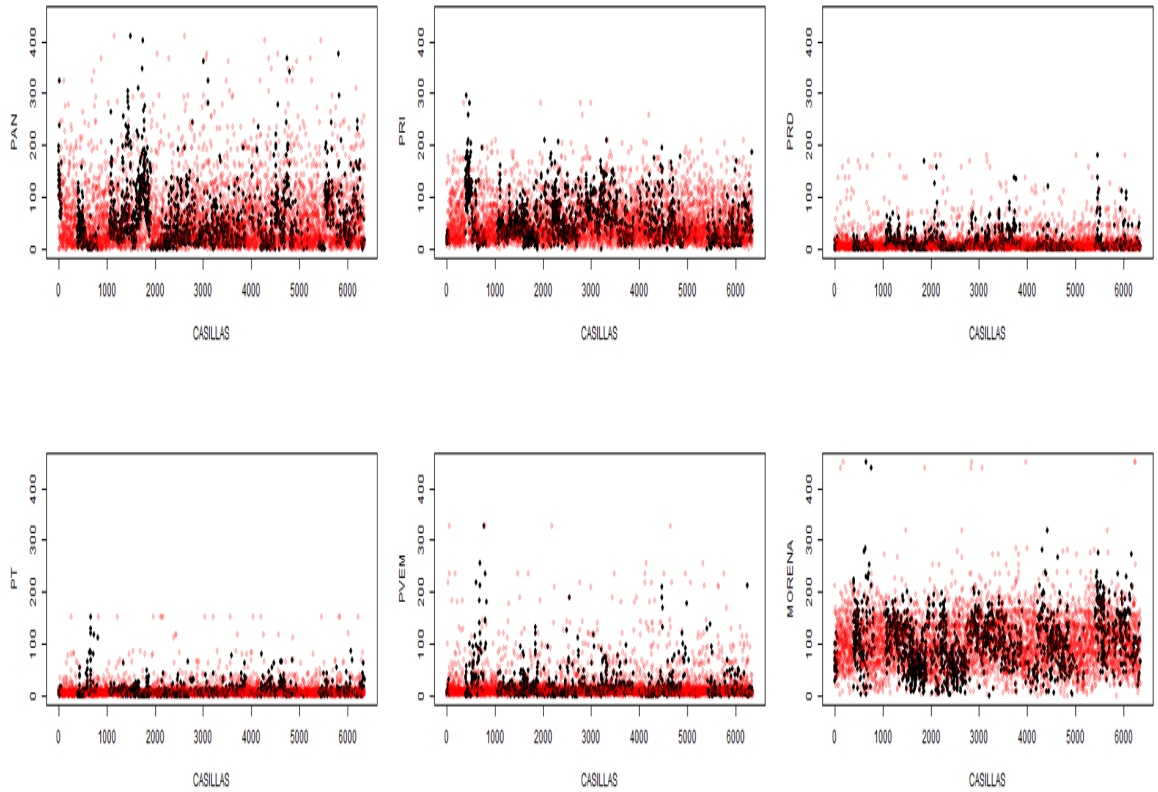


Figura 9.4: Gráfico de las estimaciones del intervalo de la Participación (Ecuación 5.3) de las 18:30 a las 20:00 horas el día de la Jornada Electoral. Si se desea ver el desarrollo de estos gráficos a lo largo del tiempo puede verse la Figura 7.13.



(a) Densidad de las imputaciones para observar el sesgo en las estimaciones.



(b) Contraste de casillas imputadas (rojas) contra las observadas (negras).

Figura 9.5: Comportamiento de las estimaciones para ver el sesgo por Imputación Múltiple (ver Subsección 6.2.2) a las 20:00 horas del día de la Jornada Electoral. Si se desea ver el desarrollo de estos gráficos a lo largo del tiempo pueden verse las Figuras 7.14 y 7.15.

Siglas y Notaciones Matemáticas

Siglas

CAE *Capacitador Asistente Electoral.*

`cart` *Classification and regression trees.*

CDF *Función de distribución acumulada.*

CN *Cociente Natural.*

COTECORA *Comité Técnico Asesor de los Conteos Rápidos.*

COVID-19 *Coronavirus disease 2019.*

CPEUM *Constitución Política de los Estados Unidos Mexicanos.*

CPF *corrección por población finita.*

CR *Conteo Rápido.*

CRAN *The Comprehensive R Archive Network.*

ECDF *Función de distribución acumulada empírica.*

IFE *Instituto Federal Electoral.*

INE *Instituto Nacional Electoral.*

INEGI *Instituto Nacional de Estadística y Geografía.*

LEGIPE *Ley General de Instituciones y Procedimientos Electorales.*

LGPP *Ley General de Partidos Políticos.*

LN *Lista Nominal.*

LOCGEUM *Ley Orgánica del Congreso General de los Estados Unidos Mexicanos.*

MAE *Muestro Aleatorio Estratificado.*

MAR *Missing At Random.*

MAS *Muestro Aleatorio Simple.*

MASSR *Muestro Aleatorio Simple Sin Remplazo.*

MCAR *Missing Completely At Random.*

MEMA *Máximo de Escaños Mal Asignados.*

`mice` *Multivariate Imputation by Chained Equations.*

MR *Mayoría Relativa.*

`NA` *Not Available.*

NCN *Nuevo Cociente Natural.*

NMAR *Not Missing At Random.*

OPLE *Organismo Público Local Electoral.*

`pmm` *Predictive Mean Matching.*

PREP *Programa de Resultados Electorales Preliminares.*

RCD *Reglamento de la Cámara de Diputados del H. Congreso de la Unión.*

`rf` *Random forest (imputations).*

RP *Representación Proporcional.*

SCJN *Suprema Corte de Justicia de la Nación.*

TEPJF *Tribunal Electoral del Poder Judicial de la Federación.*

TLC *Teorema del Límite Central.*

v.a.i.i.d. *Variables Aleatorias Independientes e Idénticamente Distribuidas.*

VNE *Votación Nacional Emitida.*

VNE$_f$ *Votación Nacional Efectiva.*

VTE *Votación Total Emitida.*

VVE *Votación Válida Emitida.*

Notación Matemática

Actualizar En este documento, cuando decimos que *actualizamos* a x con $x = f(x)$ donde f no es necesariamente la función identidad, significa que x tiene algún valor inicial y éste irá cambiando de tal manera que le apliquemos f y dicho resultado se volverá a llamar x sin distinción del valor anterior que tenía.

$\mathcal{C}_i$ Conjunto de partidos que forman una coalición arbitraria en el distrito i.

$|A|$ **con** A **un conjunto.** Denotaremos de esta forma la cardinalidad del conjunto A.

CONF Conformación de la cámara de diputados (ver Ecuación 3.1).

$\mathbb{E}[X]$ Valor esperado de la variable aleatoria X.

$\mathbb{1}(A)$ Función indicadora. Bajo esta notación, si A es un evento verdadero esta función toma el valor 1, en otro caso toma el valor 0.

$Ker(f)$ Kernel de la función f. Es el conjunto de valores x en el dominio de f tales que $f(x) = 0$.

$\lfloor x \rfloor$ Función mayor entero menor o igual de $x \in \mathbb{R}$.

$\mathbb{N}$ Conjunto de Números Naturales (incluye al cero).

$\mathbb{P}[A]$ Probabilidad del evento A.

Φ Función de distribución acumulada de una normal estándar.

$\mathbb{R}$ Conjunto de Números Reales.

sii Si y sólo si.

$V(X)$ Varianza matemática de la variable aleatoria X.

Glosario

Antinomia Contradicción u oposición entre dos conceptos o principios.

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

Circunscripción plurinominal Área geográfica integrada por un grupo de entidades federativas, que sirve de base para la elección de los 200 diputados y 32 senadores electos por el principio de representación proporcional.

Coalición (Política) Pacto entre dos o más partidos políticos, normalmente de ideas afines, para gobernar un país, una región u otra entidad administrativa.

Coaligado Que forma parte de una coalición, en este caso, entre partidos políticos.

Conteo Rápido Es un procedimiento estadístico diseñado con la finalidad de estimar con oportunidad las tendencias de los resultados finales de una elección. Está basado en las actas de escrutinio y cómputo de casilla a fin de conocer las tendencias de los resultados de la jornada electoral.

Diputación Ejercicio/Duración/Quehacer del cargo de un diputado.

Distrito Federal Electoral Se refiere a la división geográfica en que se organiza el territorio con fines electorales. Se ocupan para organizar la elección de diputados federales. México está dividido en 300 distritos electorales federales.

Escaño Puesto representativo en una cámara electiva. Sinónimo de curul.

Escrutinio Reconocimiento y cómputo de los votos en las elecciones o en otro acto análogo.

Jurisprudencia Compuesta por los vocablos "*juris*" que significa derecho y "*prudentia*" que quiere decir conocimiento, ciencia. Se define como el conjunto de tesis que constituyen valioso material de orientación y enseñanza, que señalan a los jueces la solución de la multiplicidad de cuestiones jurídicas que contemplan; que suplen las lagunas y deficiencias del orden jurídico positivo; que guían al legislador en el sendero de su obra futura. En el caso de México, la jurisprudencia judicial es la interpretación de la ley, firme, reiterada y de observancia

obligatoria, que emana de las ejecutorias pronunciadas por la Suprema Corte de Justicia de la Nación, funcionando en pleno o por salas, y por los Tribunales Colegiados de Circuito.

Lista Nominal Es un documento que contiene el nombre y la foto de la ciudadanía que cuenta con credencial para votar vigente, es decir, es la población que podrá ejercer su derecho al voto.

Militante Persona que pertenece a determinada ideología, grupo o partido político.

Parlamentario Perteneciente o relativo al Parlamento judicial o político.

Plurinominal* Referente a los curules a otorgar por el principio de *Representación Proporcional* (RP).

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

Sobrerrepresentación (legislativa) Por sobrerrepresentación se entiende que un partido político obtiene, en función de determinados mecanismos electorales, un porcentaje de curules superior al porcentaje de votos obtenidos o permitidos por la ley.

Sufragar Dar el voto a un candidato, una propuesta, dictamen, etc.

Bibliografía

[1] J. H. Mendoza, "Esquemas de muestreo para el conteo rápido bajo restricciones operativas," 2019. https://drive.google.com/file/d/18MQBTuntjQ2ky_u5MbWuFeDDO1ogdax_/view?usp=sharing.

[2] COTECORA, "Criterios científicos, logísticos y operativos para la realización de los conteos rápidos y protocolo para la selección de las muestras," 2020-2021. https://repositoriodocumental.ine.mx/xmlui/bitstream/handle/123456789/119626/CGor202104-28-ap-15-Anexo.pdf.

[3] LOCGEUM, "Ley orgánica del congreso general de los estados unidos mexicanos," Última reforma publicada DOF 08-05-2019. http://www.diputados.gob.mx/LeyesBiblio/pdf/168_080519.pdf.

[4] CPEUM, "Constitución políca de los estados unidos mexicanos," 2021. http://www.diputados.gob.mx/LeyesBiblio/pdf_mov/Constitucion_Politica.pdf.

[5] RCD, "Reglamento de la cámara de diputados del h. congreso de la unión," Última reforma publicada DOF 27-04-2021. http://www.diputados.gob.mx/LeyesBiblio/pdf/Reg_Diputados_270421.pdf.

[6] LEGIPE, "Ley general de instituciones y procedimientos electorales," 2017. https://www.ine.mx/wp-content/uploads/2020/07/Despen-LEGIPE-NormaINE.pdf.

[7] F. de la Mata Pizaña, "Caso: Garantía a los límites de sobrerrepresentación en la cámara de diputados," 29/04/2021. https://www.te.gob.mx/blog/delamata/front/openJustice/article/228.

[8] S. L. Lohr, "Sampling: Design and analysis," 2010. https://drive.google.com/file/d/18QN65cxy8YzbbaIMGA4DGQNTQnSHgl14/view?usp=sharing.

[9] W. Cochran, "Sampling techniques," 1977. https://drive.google.com/file/d/18ERsPOMYGma583Z6Tnl00oqBgHPHHeef/view?usp=sharing.

[10] R. A. Sudgen, "Cochran's rule for simple random sampling," 2000. https://drive.google.com/file/d/1vKp04dvvz-dMwLN1MXBAFZJcIovlrTXT/view.

[11] L. Wasserman, "All of statistics. a concise course in statistical inference," 2003. https://drive.google.com/file/d/1XIQYZvmw1W21qDH0KWC1rQA8jbSdIBO5/view?usp=sharing.

[12] D. B. Rubin, "Multiple imputation for nonresponse in surveys," 1987. https://drive.google.com/file/d/1XHU9GGKuA5vFLGckdM9AK_ikqagTDrBO/view?usp=sharing.

[13] D. O. García, "Imputación de datos faltantes en un sistema de información sobre conductas de riesgo," 2011. https://drive.google.com/file/d/1zRvIoPmU53xXybr2OstXozMuVZsb1_AK/view?usp=sharing.

[14] S. van Buuren, "Flexible imputation of missing data," 2018. Versión más actual: https://stefvanbuuren.name/fimd/.

[15] J. Schork, "Predictive mean matching imputation (theory example in r)," 2022. https://statisticsglobe.com/predictive-mean-matching-imputation-method/.

[16] StataCorp, "Stata multiple-imputation reference manual," 2021. Versión más actual: https://www.stata.com/manuals/mi.pdf.

[17] I. P. Fellegi and D. Holt, "A systematic approach to automatic editing and imputation," 2014. https://drive.google.com/file/d/1Y2a1eC4_he1LfX3wTSgQk0dRGXVlSDF7/view?usp=sharing.

[18] A. P. Goicoechea, "Imputación basada en árboles de clasificación," 2002. https://www.eustat.eus/documentos/datos/ct_04_c.pdf.

[19] F. Manetto, "El tribunal electoral respalda al ine frente a la presión de morena por la elección de la cámara," 27 de Abril del 2021. https://tinyurl.com/bdhjn7p4.

[20] S. Corina, "Morena pierde la mayoría absoluta y necesita de aliados para controlar el congreso," 16 de Mayo del 2021. https://tinyurl.com/49sxnjm8.

[21] K. G. Baray, "Elecciones 2021: ¿qué es la sobrerrepresentación en la cámara de diputados; y cómo el ine busca evitarla?," 16 de Mayo del 2021. https://tinyurl.com/44bueshn.

[22] LGPP, "Ley general de partidos políticos," 2021. http://www.diputados.gob.mx/LeyesBiblio/pdf/LGPP_130420.pdf.