Artículo Original
Revista Varianza - Año 2022, Nro. 19
ISSN 2789-3510 versión impresa
Páginas: 17-26

Una aproximación al diseño muestral óptimo

An approach to the optimal sampling design

Ronal Edwin Condori Huanca[1]
Universidad Mayor de San Andrés, La Paz- Bolivia
✉ *ronal.c.huanca@gmail.com*

Artículo recibido: 2022-02-18
Artículo aceptado: 2022-03-28

Resumen

Ante la ausencia del desarrollo y cuantificación de una mayor diversidad de medidas de precisión para los diseños de muestreo en el contexto Boliviano. El presente documento aborda el desarrollo del objetivo experimental de comparabilidad y elección de diferentes diseños de muestreo potenciales a ser aplicados a conteos rápidos para fines electorales, se utilizó información de las elecciones generales de 2020 en Bolivia, seleccionando hasta dieciocho escenarios posibles de muestreo, y evaluando sus errores estándar, coeficientes de variación logarítmica, y otras medidas más. Considerando al estimador de razón, se encontró que el diseño con estratificación en base al tamaño de recinto con una selección de tipo secuencial, es el diseño más óptimo con un coeficiente logarítmico de 0.042, errores estándar cercanos al 0.018, y otras medidas de precisión.

Palabras clave: Conteo Rápido, Coeficiente de variación logarítmico, Diseño de muestreo optimo, Elecciones.

Abstract

In the absence of the development and quantification of a greater diversity of precision measures for proven designs in the Bolivian context. This document addresses the development of the experimental objective of comparability and choice of different potential sampling designs to be applied to quick counts for electoral purposes, for the same information from the 2020 general elections in Bolivia was used, selecting up to eighteen possible proven scenarios, and evaluating their standard errors, coefficients of logarithmic variation, and other measures. Considering the ratio estimator, it was found that the design with stratification based on enclosure size with a sequential type selection is the most optimal design with a logarithmic coefficient of 0.042, standard errors close to 0.018, and other more accurate measures.

Keywords: Quick Count, Logarithmic coefficient of variation, Optimal sampling design, Elections.

1. Introducción

En el contexto Boliviano, para la mayoría de las encuestas probabilísticas, la elección del diseño de muestra puede ser considerada como una decisión muy subjetiva, y se basa principalmente los siguientes aspectos:

[1] Profesional/consultor en Estadística(s), candidato a Doctor en Políticas Públicas de la UMSA: Ha ejercido laboralmente en el área de estadística en entidades privadas y en entidades públicas como el INE-Bolivia, Ministerios de Salud, Ministerio de Desarrollo Productivo, Ministerio de Educación, Ministerio de Economía, etc. https://orcid.org/0000-0003-2557-7079

- Información disponible: encuestas previas, marcos muestrales, cartografía, etc.
- Limitantes de planificación y ejecución: presupuesto y personal disponible, fechas comprometidas para entrega de resultados.
- Aspectos Logísticos: organización, desplazamiento en campo, limitantes urbano rural, externalidades (clima, conflictos políticos, fricciones socioculturales en los informantes, etc.).

Este análisis suele estar a cargo del personal que elabora el diseño de muestreo, en coordinación y aprobación con los financiadores, planificadores y demás miembros del equipo temático a cargo de la encuesta en la institución. Y bajo esta última etapa de diseño y consolidación metodológica, suelen ignorarse aspectos como, el tamaño de los conglomerados, estratificación, la reducción mínima de los márgenes de error de muestreo, sesgos posibles, varianzas mínimas, y otros.

1.1. Antecedentes

En los diferentes actos electorales que se celebran en cualquier país, el recuento de votos por muestreo, o conteo de votos, es una técnica ampliamente utilizada para dar un resultado preciso, confiable, y rápido, de manera de proporcionar un panorama de tranquilidad a los diferentes actores políticos, instituciones y público en general.

En la región latinoamericana el uso de esta técnica en los actos electorales ha ido desapareciendo, debido a la digitalización de los actos electorales, o de los procesos utilizados en el recuento de votos oficial que se aplican, pudiendo así entregar sus resultados incluso en menos de 48 horas posteriores al día del acto electoral (RNEC, 2019), (TEP, 2021) y otras aún con postergaciones mayores (Garzón-Sherdek, 2021), (SERVEL, 2021), (TSJE, 2021).

En los últimos años en Bolivia, estas operaciones estadísticas fueron utilizadas en las elecciones sub nacionales de 2021 (CIESMORI, 2021), (FOCALIZA, 2021), en las elecciones generales de 2020 (Página-Siete, 2020) y las de 2019 (ViaCiencia, 2019).

Según toda esta información, las operaciones estadísticas señaladas anteriormente, utilizan diseños de muestreo que en su mayoría solo mencionan el error de muestreo teórico usado en el tamaño de muestra, sin cuantificar otras medidas de precisión. Sin embargo para 2020, ya se incluyen el error estándar y los intervalos de confianza (FOCALIZA, 2021), conforme señala el último reglamento para la elaboración de estos estudios (OEP, 2020).

1.2. Problemática

En el medio Boliviano, no existe una práctica o hábito en las encuestas por muestreo y su diseño, sobre como sustentar cuantitativamente el proceso de elección del diseño de muestreo más apropiado. Incluso en las mismas encuestas multipropósito que realiza el Instituto Nacional de Estadística de Bolivia (INE), no se detallan públicamente pormenores de este tipo (INE, 2020).

Esta acefalía en la práctica para elaboración de estos procesos estadísticos (que incluso no es cubierta por la oficina nacional de estadística), plantea la necesidad de poder aplicar y desarrollar en el contexto Boliviano un conjunto de recomendaciones sobre la elección cuantitativa de diseños de muestreo (Deville, Sarndal, & Sautory, 1993), entre las diferentes alternativas que puedan ser utilizadas de forma posterior.

Cabe aclarar que no está en discusión la técnica de distribución del tamaño de muestra, la cual verifica por ejemplo que para un estimador estratificado de la media $\bar{y}_{st}$, establece que la afijación óptima de Neyman posee la mínima varianza respecto a una afijación proporcional y un muestreo aleatorio simple: $V_{opt}(\bar{y}_{st}) \leq V_{prop}(\bar{y}_{st}) \leq V_{ran}(\bar{y})$ (Cochran, 1977), o la disyuntiva sobre cual método de estimación es más preciso entre los métodos directos (totales, promedios o proporciones), en comparación con los métodos indirectos (razón, regresión o por diferencias), esta comparación ya está verificada teóricamente, pudiéndose expresar como: $V(\bar{y}_{reg}) \leq V(\bar{y}_R) \leq V(\bar{y})$ (Perez, 2005).

Lo que está en discusión, es la comparación de las alternativas de diseño de muestreo que se pueden aplicar a una encuesta por muestreo en específico.

1.3. Hipótesis

Es posible desarrollar una comparación de diseños de muestreo de manera cuantitativa, y así elegir una estrategia aproximadamente óptima para el diseño de muestreo aplicado en las elecciones generales de 2020.

2. Material y Métodos

2.1. El contexto de la estimación

Como se mencionó anteriormente, la aplicación de la investigación se realiza en las elecciones generales 2020, en la cual el objetivo del diseño de muestreo es el de *estimar la distribución de votos válidos por partido político, a nivel nacional, y a nivel departamental.* Este hecho plantea definir nueve (9) muestras independientes en cada uno de los departamentos, para la asignación de escaños, tanto en diputados como en senadores (Ley026, 2010).

Otro aspecto que siempre se consideran en encuestas electorales es el del contraste urbano vs. rural, o en algunos casos, capital vs. resto, o también una agrupación de las diferentes ciudades, comunidades o localidades, según su tamaño poblacional.

2.2. Los marcos de muestreo

Los marcos de muestreo utilizados están compuestos por el padrón electoral oficial para las elecciones generales de 2020 y los resultados oficiales para el mismo evento. En algunos casos se utilizaron como información auxiliar los tamaños de muestra elaborados por las instituciones citadas concretamente.

Cabe aclarar que se utilizaron varios marcos de muestreo, ya que, al modificar los estratos, conglomerados u otra etapa, estos marcos son diferentes entre sí, aunque en algunos casos pueden ser comparables a nivel de unidades primaras de muestreo (UPM).

2.3. Tipos de estratificación

Se planteó el uso de estratificación en cada uno de los departamentos del siguiente tipo:

1. Urbano y Rural
2. Capital y Resto
3. Por tamaño del recinto

El primer caso asumió la clasificación que el INE maneja sobre los centros poblados clasificados como urbanos en base al último Censo de Población y Vivienda de 2012. El segundo caso consideró al municipio capital como un estrato único dentro del departamento y el resto fue el complemento de los municipios dentro de este. Sólo en el departamento de La Paz, consideró al municipio de El Alto como un estrato separado. El último tipo de estratificación, creo estratos para los recintos de menos de 6 mesas, y otro de 6 o más mesas.

Estos tipos de estratificación son ampliamente usados, aunque existan otros métodos para fines electorales que permiten mejorar las estimaciones usando variables electorales como tal (Condori, 2021), (PEW R.C., 2021), estos no serán abordados.

2.4. Tipos de selección

Los tipos de selección utilizados en la 1ra etapa fueron:

- Aleatoria simple de UPM
- Sistemática simple de UPM
- Secuencial simple de UPM
- Aleatoria con probabilidad proporcional al tamaño de la UPM
- Sistemática con probabilidad proporcional al tamaño de la UPM
- Secuencial con probabilidad proporcional al tamaño de la UPM

Los tipos de selección variaron solo en la primera etapa, en la segunda etapa solo se seleccionaron una única mesa dentro de cada recinto, la que se realizó de forma aleatoria simple para todos los diseños. Para las selecciones proporcionales al tamaño, se utilizó como medida de tamaño de selección al número de mesas dentro del recinto.

Las selección aleatoria simple y sistemática son ampliamente conocidas, pero la selección secuencial no es muy difundida, sin embargo se la puede describir como una simbiosis de ambas, planteando un escenario intermedio, entre saltos sistemáticos con una componente aleatoria (Chromy, 1979).

2.5. Tamaños y distribución de muestra

El tamaño y distribución de muestra es tal vez la parte importante en los diseños, sin embargo, para no recaer en usar “muestras muy grandes” e inoperables, se utilizó el tamaño de muestra total y del recuento de votos de CiesMori de 2020, que es bastante modesto y más ágil en su aplicación. El tamaño de la muestra fue de 269 recintos con un total aproximado de 151 inscritos (ERBOL, 2020), y asumiendo que regularmente hay 200 inscritos por mesa, se tuvieron que haber visitado un aproximado de 755 mesas.

Por otro lado, el estudio de la iniciativa TuVotocuenta, organizado por la UMSA y la Fundación Jubileo, (LosTiempos, 2020), muestrearon 4.711 mesas en 1.200 recintos.

Para el presente estudio se aplicó el tamaño de CiesMori de 269 recintos, y en cada recinto se elegió una sola mesa. Por limitantes de información, la ficha técnica no estaba publicada oficialmente en la página oficial del Órgano Electoral Plurinacional (OEP), con lo cual se optó por una distribución potencial con un α=0.5 entre los departamentos, y posteriormente dentro de cada departamento se aplicó una distribución proporcional.

2.6. Metodología e indicadores para la comparación de diseños de muestreo y sus estimaciones

Entre los indicadores de calidad posterior a estimación de un muestreo las métricas fueron:

- Error estándar (ee)
- Error de muestreo o coeficiente de variación (cv)
- Efecto de diseño (eff)
- Coeficiente de variación logarítmico (cvl)

Este último elemento es una medida propuesta por algunos autores, pero poco difundida (SAMHSA, 2007), debido a que en diversas ocasiones cuando los estimadores de tipo proporción $\hat{p}$, suelen tener errores de muestreo muy altos cuando las estimaciones de $\hat{p}$ presentan valores muy bajos, pero contradictoriamente para el estimador de su complemento $\hat{q} = 1 - \hat{p}$ el error de muestreo estimado es muy bajo. Esta medida de error de muestreo se expresa como:

$$cv(\hat{L}) = -cv(\hat{p})/log(1-\hat{p})$$

Dicha expresión se deduce de la aproximación de Taylor de primer orden para el estimador logarítmico $\hat{L} = -log(1-\hat{p})$ (Gutierrez, Fuentes, & Mancero, 2020). Para el presente estudio se aplicó al estimador de razón $\hat{r}$ para el porcentaje de votos recibidos a cada partido: *cv(L̂)=-cv(r̂)/log(1-r̂)*.

El resto de estadísticas fueron definidas convencionalmente $ee(\hat{r}) = \sqrt{V(\hat{r})}$ para el error estándar, $ee(\hat{r}) = \sqrt{V(\hat{r})}$ para el coeficiente de variación, y $eff(\hat{r}) = V(\hat{r})/V_{MAS}(\hat{r})$ para el efecto de diseño, donde el último termino es la varianza del estimador bajo un muestreo aleatorio simple .

2.7. Diseño de la investigación

Dado que se planificó múltiples escenarios, y se compararon entre sí, se considera esta investigación de tipo experimental, ya que se están evaluaron diferentes diseños de muestreo a ser aplicados.

2.8. Implementación de los diseños de muestreo

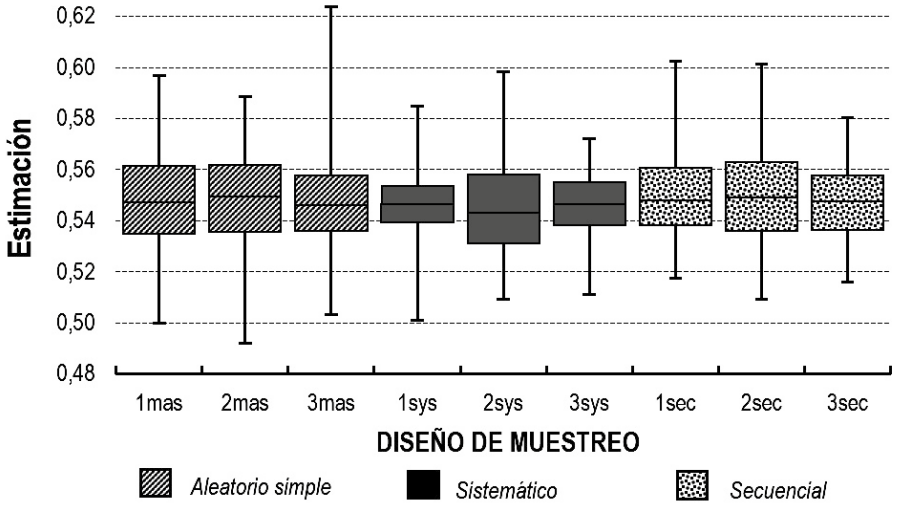
Para poder implementar todos los diseños de muestreo y sus estimaciones, se simularon en base a un conjunto de 100 semillas aleatorias, a 100 muestras independientes para cada uno de los diseños propuestos anteriormente.

Dado que se trabajó con 3 diseños de muestreo posibles con su estratificación respectiva y 6 tipo de selección, se tuvieron 18 escenarios planteados se limitó la comparación de los estimadores y sus medidas de error solo a los resultados a nivel nacional.

3. Resultados

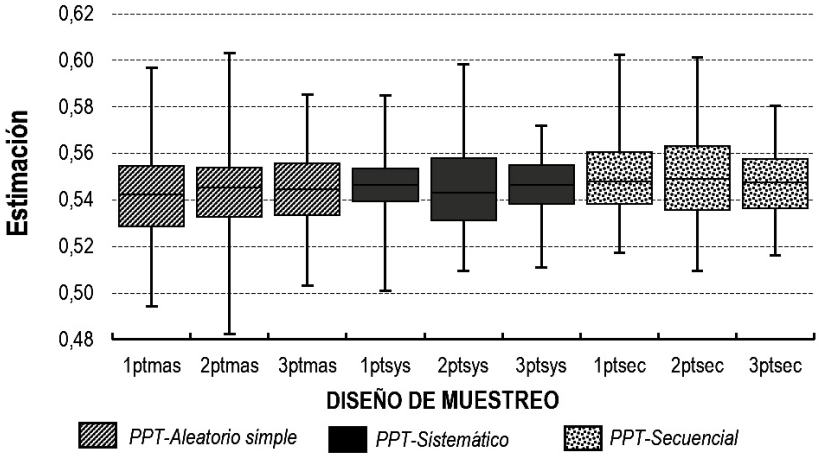
A continuación, se muestran las estimaciones del porcentaje de votos válidos, su error estándar y coeficiente de variación, margen de error respecto al parámetro y error cuadrático medio.

Figura 1.
Boxplot Estimaciones de votos al partido MAS, según los diferentes tipos de diseño



Fuente: Resultados Elecciones Generales 2020, OEP.
Elaboración: Propia.

Figura 2.
Boxplot Coeficientes de variación promedio de votos al partido MAS, según los diferentes tipos de diseño

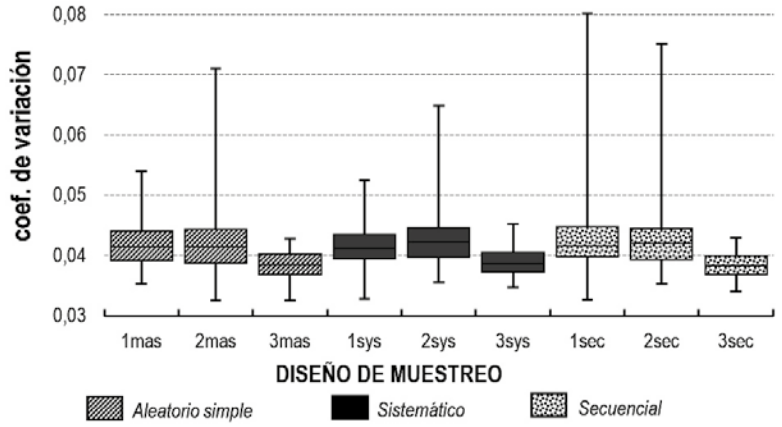


Fuente: Resultados Elecciones Generales 2020, OEP.
Elaboración: Propia.

Estas estimaciones en su mayoría bordean el resultado oficial nacional de 54.7% para el partido MAS, 14.06% a CREEMOS, y 29.16% para CC (OEP, 2020). Se puede observar en las Figuras 1 y 2, que las selecciones no son proporcionales, las que poseen menor dispersión en sus estimaciones, e incluso son más homogéneas entre sí.

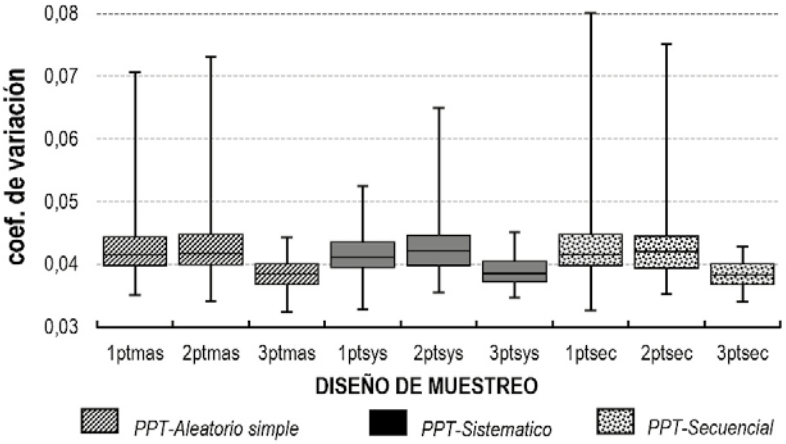
Por su lado, los coeficientes de variación, son muy semejantes entre selección PPT y selección simple (Figuras 2 y 3).

Figura 3.
Boxplot Coeficientes de variación promedio de votos al partido MAS, según los diferentes tipos de diseño no proporcional



Fuente: Resultados Elecciones Generales 2020, OEP.
Elaboración: Propia.

Figura 4.
Boxplot Coeficientes de variación promedio de votos al partido MAS, según los diferentes tipos de diseño proporcional



Fuente: Resultados Elecciones Generales 2020, OEP.
Elaboración: Propia.

A continuación, se observa la dispersión de los promedios de errores estándar y coeficientes de variación, para la misma estimación. En dicho gráfico se observa que los diseños de menor error estándar y coeficiente de variación son los de la 3ra estratificación con selección secuencial (Figura 4).

Tabla 1:
Coef. de var. logarítmicos promedios, según tipo de diseño

Selección	Estratificación		
	1ra	2da	3ra
mas	0.047	0.046	0.042
sys	0.046	0.048	0.043
sec	0.048	0.047	0.042
pptmas	0.048	0.049	0.043
pptsys	0.046	0.048	0.043
pptsec	0.048	0.047	0.042

Fuente: Resultados Elecciones Generales 2020, OEP.
Elaboración: Propia

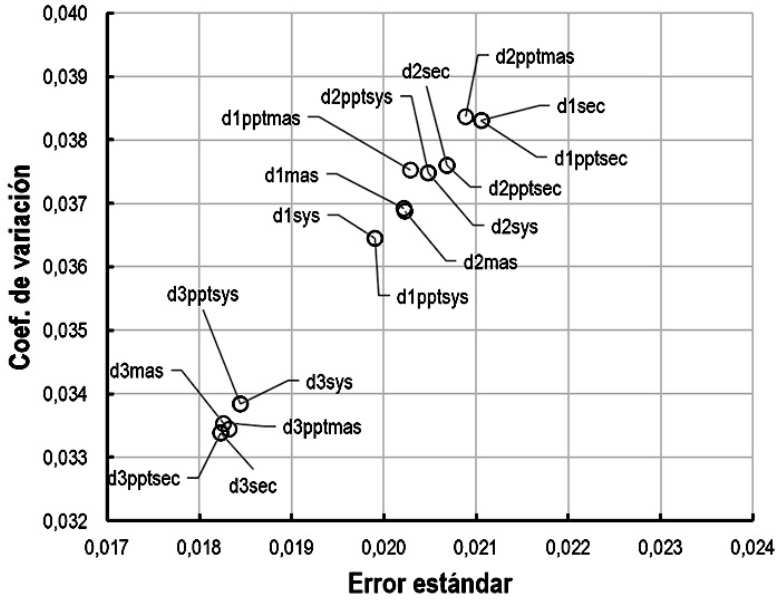
Este último hecho se contrasta con el coeficiente de variación logarítmico (Tabla 1), y el efecto de diseño (Tabla 2) el cual también señala a al mismo tipo de diseño, como el de menor medición, junto con la selección aleatoria simple bajo la 3ra estratificación.

Tabla 2:
Efectos de diseño promedio, según tipo de diseño

Selección	Estratificación		
	1ra	2da	3ra
mas	1.408	1.414	1.138
sys	1.359	1.461	1.158
sec	1.539	1.471	1.141
pptmas	1.406	1.518	1.136
pptsys	1.359	1.461	1.158
pptsec	1.539	1.471	1.141

Fuente: Resultados Elecciones Generales 2020, OEP.
Elaboración: Propia

Figura 5.
Coeficientes de variación y errores estándar promedio, según los diferentes tipos de diseño



Fuente: Resultados Elecciones Generales 2020, OEP.
Elaboración: Propia.

4. Discusión

Si bien los resultados se inclinan a señalar que el mejor diseño de muestreo fue el de estratificación según el tamaño de recinto (3ro), con una selección secuencial. Este diseño aún debe ser evaluado por la componente aleatoria pura, la cual encierra a los errores no muestrales como son la tasa de no respuesta, inaccesibilidad potencial de algunos unidades seleccionadas en la muestra, los sesgos de información (datos mal transcritos), u otros más, que deben ser incluidos en las simulaciones antes planteadas.

Lastimosamente no se puede comparar los errores estándar, coeficientes de variación y demás medidas aquí cuantificadas, dado que el estudio guía utilizado no publicó de manera oficial dicha información.

Se debe ampliar la comparación sobre las estimaciones en la composición de las cámaras de senadores y diputados para el caso Boliviano, esto de manera semejante al análisis de Stoker sobre los efectos multinivel en las medidas de precisión de muestreo (Stoker y Bowers, 2002).

5. Conclusiones

Es posible aplicar la metodología de comparación de diseños de muestreo en la temática electoral, pudiendo ostentar otras alternativas a solo observar el error teórico en función al tamaño de muestra.

Estos ejercicios permitirían elegir estrategias de muestreo con una mayor experiencia de los posibles escenarios, o incluso la posibilidad de reducir los tamaños de muestra en las encuestas oficiales que el INE sin incrementar los errores de muestreo y agilizando el procesamiento de estas.

Referencias Bibliográficas

Chromy, J. (1979). Sequential Sample Selection Methods. *Research Triangle Institute*, 401-406, http://www.asasrms.org/Proceedings/papers/1979_081.pdf

CIESMORI. (marzo de 2021 de 2021). *Informe Conteo Rapido Elecciones Subnacionales del 7 de marzo 2021*. Obtenido de https://www.oep.org.bo/wp-content/uploads/2021/04/Informe_Conteo_Rapido_CIESMORI_EDRM_2021.pdf

Cochran, W. (1977). *Sampling Techniques*, 3rd Edition. New York: Jhon Wiley & Sons.

Condori, R. (2021). Estratificacion Asimetrica en Encuestas Electorales. *Varianza Nro 18, UMSA,* 9-19, https://drive.google.com/file/d/1lDzdNUQlg1A68UFVed96Fp3W1V9fCVTS/view

Deville, J.-C., Sarndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association, Vol. 88, No. 423*, 1013-1020, https://doi.org/10.2307/2290793

ERBOL. (19 de octubre de 2020). *ERBOL Educacion Radiofonica de Bolivia.* Obtenido de Periodico ERBOL: https://erbol.com.bo/el-%C3%A1nfora-1/conteo-de-ciesmori-proyecta-victoria-de-luis-arce-en-primera-vuelta

FOCALIZA. (2021). *Informe Técnico Recuento Rápido Bolivia: Elecciones Sub Nacionales 2021*. Santa Cruz de la Sierra, https://www.oep.org.bo/wp-content/uploads/2021/04/Informe_Conteo_Rapido_Focaliza_EDRM_2021.pdf

Garzón-Sherdek, K. (2021). Ecuador Elecciones Generales 2021 (Segunda vuelta). *Análisis de Elecciones 2021, Observatorio de Reformas Políticas en América Latina, IIJ-UNAM y Organización de los Estados Americanos (OEA)*., 1-15, https://reformaspoliticas.org/wp-content/uploads/2021/05/Analisis-Elecciones-Ecuador-segunda-vuelta-1.pdf

Gutierrez, A., Fuentes, A., & Mancero, X. (2020). *Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: Una aplicacion a la migracion internacional.* Santiago: Comisión Económica para América Latina y el Caribe (CEPAL).

INE. (2020). *Diseño de muestra Encuesta de Hogares 2020*. La Paz: Instituto Nacional de Estadistica, http://anda.ine.gob.bo/index.php/catalog/88/download/922

Ley026. (30 de junio de 2010). Ley del Régimen Electoral. *Estado Plurinacional de Bolivia*.

LosTiempos. (19 de octubre de 2020). *"Tu Voto Cuenta": MAS obtuvo 53% y CC 30.8%, según resultados de conteo rápido*. Obtenido de Periodico Los Tiempos: https://www.lostiempos.com/actualidad/pais/20201019/tu-voto-cuenta-mas-obtuvo-53-cc-308-resultados-conteo-rapido

OEP. (2020). *Acta de Computo Nacional Elecciones Generales 2020*. La Paz-Bolivia: Organo Electoral Plurinacional, https://www.oep.org.bo/wp-content/uploads/2020/11/Acta-de-computo-2020.pdf

OEP. (septiembre de 2020). *Reglamento de Elaboración y difusión de Estudios de Opinión en materia Electoral en Procesos Electorales*. Obtenido de Elecciones Generales 2020: https://www.oep.org.bo/wp-content/uploads/2020/10/Reg_Estudios_Opinion_EG_2020.pdf

Página-Siete. (19 de octubre de 2020). *Página Siete*. Obtenido de Arce logra más de 50%, según el conteo rápido de dos firmas: https://www.paginasiete.bo/nacional/2020/10/19/arce-logra-mas-de-50-segun-el-conteo-rapido-de-dos-firmas-272050.html

Perez, C. (2005). *Muestreo Estadistico, conceptos y problemas resueltos.* Madrid: Pearson Prentice Hall.

PEW R.C. (30 de Junio de 2021). Behind Biden's 2020 *Victory: Methodology.* Obtenido de PEW RESEARCH CENTER: https://www.pewresearch.org/politics/2021/06/30/validated-voters-methodology/

RNEC. (2019). *Registraduria Nacional del Estado Civil*. Obtenido de Calendario Electoral Autoridades Locales 2019: https://www.registraduria.gov.co/IMG/pdf/RES-14778-11-OCT-2018-CALENDARIO-ELECTORAL-AUTORIDADES-LOCALES.pdf

SAMHSA. (2007). *2006 National Survey on Drug Use and Health: National Findings*. Rockville: Office of Applied Studies, NSDUH Series H-32. Obtenido de https://files.eric.ed.gov/fulltext/ED498206.pdf

SERVEL. (Julio de 2021). *Cronograma Electoral Elecciones Presidencial, Parlamentarias y de Consejeros regionales 2021*. Obtenido de Subdirección Registro, Inscripciones y Acto Electoral: División de Procesos Electorales: https://www.servel.cl/wp-content/uploads/2021/07/CRONOGRAMA_ELECTORAL-_ELECCIONES_DEFINITIVAS_2021.pdf

Stoker, L., & Bowers, J. (2002). Designing multi-level studies: sampling voters and electoral contexts. *Electoral Studies 21*, 235–267, http://www.jakebowers.org/PAPERS/stokerbowersFinalplusErratum.pdf

TEP. (2021). *Calendario Electoral: Elecciones Federales y elecciones estatales Mexico 2021*. Obtenido de Escuela Judicial Electoral: https://www.te.gob.mx/calendario_electoral/

TSJE. (28 de Septiembre de 2021). *Tribunal Superior de Justicia Electoral - Republica del Paraguay.* Obtenido de Elecciones Municipales 10 de octubre 2021 - Dossier Informativo: https://www.tsje.gov.py/elecciones-municipales-10-de-octubre-2021---dossier-informativo.html

ViaCiencia. (2019). *Conteo Rapido ViaCiencia Elecciones Generales 2019: Ficha Técnica.* Santa Cruz, https://www.oep.org.bo/wp-content/uploads/2019/11/Conteo_Rapido_VIACIENCIA_EG_2019.pdf