TP 3 - Final Muestreo Polietápico

Gomez Vargas Andrea, Iummato Luciana, Pesce Andrea Gisele

2024-12-16

Tabla de contenidos

Paquetes de trabajo 1
Ejercicio 1A 2
Ejercicio 2: Probability vs. Nonprobability Sampling 3
Bibliografía 7

Paquetes de trabajo

```
library(tidyverse)
library(survey)
library(readxl)
library(gt)
library(sampling)
library(VIM)
library(binom)
library(openxlsx)
library(DT)
library(stratification)
library(kableExtra)
options(scipen = 999)
```

Elegiremos avanzar con el ejericicio 1A, a realizar con la base de votos de Octubre 2023, trabajando con voto a presidente, en el cual compararemos el efecto del tamaño de muestra en cada etapa en el CV.

Ejercicio 1A

El conjunto de mesas electorales de la elección Octubre 2023 será nuestro universo bajo estudio.

Se desea estimar el total de votos a Unión por la Patria, Juntos por el Cambio, La Libertad Avanza y FIT a Presidente y Vice a nivel nacional y proporción de votos respecto al total de votos positivos mediante una muestra aleatoria de mesas electorales.

Se compararán dos diseños, ambos bietápicos; con los circuitos electorales como Unidades de Primera Etapa (UPEs) y las mesas electorales como Unidades de Segunda Etapa (USEs).

Diseño A - Primer etapa de selección

```
Resultados_Octubre_2023_PRESIDENCIALES<-
read_csv("Resultados_Octubre_2023_PRESIDENCIALES.csv")

glimpse(Resultados_Octubre_2023_PRESIDENCIALES)
```

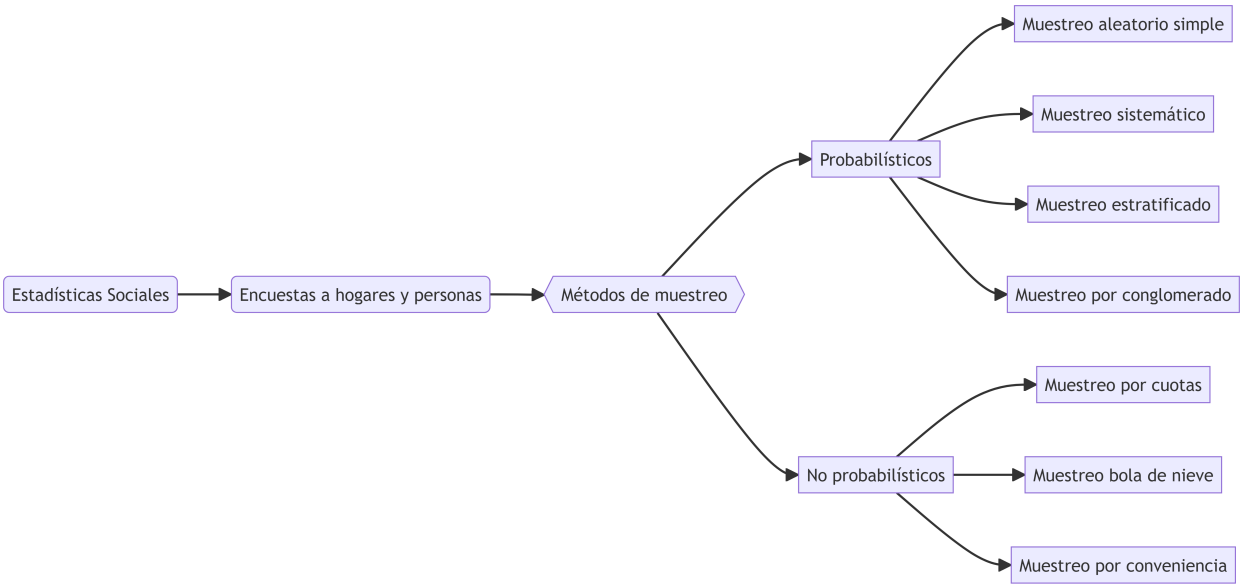
```
Rows: 1,045,200
Columns: 23
$ año                      <dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 202…
$ eleccion_tipo            <chr> "GENERAL", "GENERAL", "GENERAL", "GENERAL", "…
$ recuento_tipo            <chr> "PROVISORIO", "PROVISORIO", "PROVISORIO", "PR…
$ padron_tipo              <chr> "NORMAL", "NORMAL", "NORMAL", "NORMAL", "NORM…
$ distrito_id              <chr> "01", "01", "01", "01", "01", "01", "01", "01…
$ distrito_nombre          <chr> "Ciudad Autónoma de Buenos Aires", "Ciudad Au…
$ seccionprovincial_id     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ seccionprovincial_nombre <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ seccion_id               <chr> "001", "001", "001", "001", "001", "001", "00…
$ seccion_nombre           <chr> "Comuna 01", "Comuna 01", "Comuna 01", "Comun…
$ circuito_id              <chr> "00001", "00001", "00001", "00001", "00001", …
$ circuito_nombre          <chr> "00001", "00001", "00001", "00001", "00001", …
$ mesa_id                  <chr> "00001", "00001", "00001", "00001", "00001", …
$ mesa_tipo                <chr> "NATIVOS", "NATIVOS", "NATIVOS", "NATIVOS", "…
$ mesa_electores           <dbl> 345, 345, 345, 345, 345, 345, 345, 345, 345, …
$ cargo_id                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ cargo_nombre             <chr> "PRESIDENTE Y VICE", "PRESIDENTE Y VICE", "PR…
$ agrupacion_id            <dbl> 134, 132, 135, 136, 133, 0, 0, 0, 0, 0, 134, …
$ agrupacion_nombre        <chr> "UNION POR LA PATRIA", "JUNTOS POR EL CAMBIO"…
$ lista_numero             <dbl> NA, NA, NA, NA, NA, 0, 0, 0, 0, 0, NA, NA, NA…
$ lista_nombre             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ votos_tipo               <chr> "POSITIVO", "POSITIVO", "POSITIVO", "POSITIVO…
$ votos_cantidad           <dbl> 96, 65, 44, 13, 4, 8, 3, 0, 0, 0, 102, 86, 38…
```

Ejercicio 2: Probability vs. Nonprobability Sampling

Palabras clave: muestreo representativo, muestreo por cuotas, poblaciones difíciles de encuestar, inferencia dependiente del modelo, encuestas por Internet, big data, registros administrativos.

Este trabajo presenta una selección de los principales desarrollos en la investigación sobre encuestas desde su introducción a finales del siglo XIX, centrándose en las encuestas de hogares y personas, con énfasis en los métodos de muestreo utilizados. Aunque se han producido avances significativos en otras áreas, como los modos de recolección de datos y el diseño de cuestionarios, estos temas quedan fuera del alcance del análisis.

Figura 1. Mapa conceptual



Kiær's Representative Method of Statistical Surveys

El método representativo de Kiær, desarrollado en 1897, fue pionero en el diseño de encuestas estadísticas a gran escala en Noruega, utilizando un muestreo por áreas en dos etapas para garantizar la representatividad. Este enfoque buscaba reflejar un microcosmos de la población mediante criterios basados en el censo de 1891, seleccionando distritos administrativos y personas dentro de ellos. Sin embargo, su trabajo enfrentó fuertes críticas en reuniones del International Statistical Institute (ISI) y en Noruega, donde su diseño fue cuestionado por no ser verdaderamente representativo y por errores en la estimación de ciertas variables clave. La controversia, especialmente en torno a un proyecto de ley sobre seguridad social, llevó al abandono del método representativo en Noruega.

Tras décadas de debate, el muestreo representativo volvió a ganar aceptación en el ISI en 1924, cuando se reconoció la validez de las "investigaciones parciales". Un informe de 1926 concluyó que las muestras eran aceptables si eran suficientemente representativas, ya fuera mediante selección aleatoria o intencional. Este cambio marcó un avance hacia los métodos de muestreo que conocemos hoy, integrando principios de representatividad estadística en la investigación.

Neyman's Seminal Paper

En 1934, Neyman presentó un artículo en el que comparaba los métodos de selección aleatoria y selectiva, lo que marcó un hito en el muestreo estadístico. En su trabajo, discutió cómo se pueden realizar inferencias a partir de muestras probabilísticas de poblaciones finitas y definió el intervalo de confianza en este contexto. También criticó las limitaciones del muestreo selectivo, señalando que, en muchos casos, no representaba adecuadamente a la población, como ocurrió con el estudio realizado por Gini y Galvani sobre el Censo General Italiano. Su trabajo promovió la adopción generalizada del muestreo probabilístico, especialmente por parte de las oficinas estadísticas nacionales, y sentó las bases para el desarrollo de muchos métodos y teorías de muestreo en las décadas siguientes.

Sin embargo, para aplicar el diseño de muestreo ideal de Neyman, es necesario cumplir con ciertas condiciones, como disponer de un marco de muestreo completo, probabilidades de selección conocidas y respuestas precisas de los encuestados. A pesar de estas condiciones ideales, en la práctica surgen problemas como la no cobertura (cuando algunos elementos de la población no se incluyen en el marco de muestreo) y la no respuesta (cuando los individuos seleccionados no responden). Estos problemas pueden introducir errores en las estimaciones obtenidas a partir de las muestras.

A pesar de la robustez del muestreo probabilístico, este tiene desventajas, como los altos costos y los plazos más largos para la recolección de datos. Para hacer frente a estos desafíos, se han desarrollado métodos menos rigurosos, conocidos como "pseudo-probabilísticos", que intentan aplicar un enfoque de muestreo probabilístico pero dependen de suposiciones de modelado. Métodos como el muestreo por cuotas, ampliamente utilizado en la investigación de mercados, son ejemplos de estas técnicas, que aunque no son muestreo probabilístico, buscan aproximarse a sus resultados.

Quota Sampling

En 1936, el sondeo de la revista Literary Digest para predecir las elecciones presidenciales de Estados Unidos resultó en un claro sesgo debido a la muestra utilizada. Al seleccionar a los participantes de directorios telefónicos, propietarios de automóviles y votantes registrados, se excluyó a sectores más pobres de la población, lo que favoreció a los votantes de clases altas. Aunque el tamaño de la muestra fue grande (10 millones de personas), la falta de ajuste de ponderaciones resultó en una predicción incorrecta. Esto muestra que un tamaño de muestra grande no garantiza estimaciones precisas, como también lo señalaron los ajustes posteriores realizados por Lohr y Brick, que aún no resolvieron completamente el problema.

Para enfrentar sesgos similares, los investigadores de mercado y los encuestadores desarrollaron métodos de muestreo por cuotas. Este método controla las características demográficas de los entrevistados, como sexo, edad o estado laboral, con el fin de garantizar que la muestra sea representativa de esos grupos. Aunque el muestreo por cuotas tiene la ventaja de ser más barato y rápido que el muestreo probabilístico, también se basa en la suposición de que los no respondedores dentro de cada grupo son sustituidos aleatoriamente, lo que puede generar sesgos si no se maneja adecuadamente. Aunque algunos estudios han encontrado que los resultados de las muestras por

cuotas son similares a los de las probabilísticas, esto no siempre ocurre, lo que resalta la limitación de este enfoque.

Pseudo-Probability Sample Designs for "Hard-to-Survey Populations"

En los últimos años, ha habido un aumento significativo en el uso de métodos de encuestas sociales para estudiar las características de poblaciones difíciles de encuestar, que son pequeños segmentos de la población general sin un marco de muestreo separado. Este tipo de población incluye grupos sensibles, como niños de 1 año para encuestas de vacunación o personas cuya pertenencia a ciertos grupos es confidencial. Para estos casos, se han desarrollado varios diseños de muestreo. Un ejemplo común es el **método de muestreo del Programa Ampliado de Inmunización (EPI) de la OMS**, que emplea un diseño de muestreo de dos etapas en comunidades seleccionadas al azar, sin necesidad de listar hogares. Este método se ha utilizado ampliamente en países en desarrollo para medir la inmunización infantil.

Otros métodos incluyen el **muestreo basado en lugares**, que se utiliza para muestrear poblaciones difíciles de encuestar que frecuentan ciertos lugares, como las poblaciones nómadas o aquellas cuyo estatus es sensible. Este método requiere la construcción de un marco de lugares y la selección de muestras de ubicación y períodos de tiempo para la recolección de datos. Finalmente, el **muestreo dirigido por los propios encuestados (Respondent Driven Sampling - RDS)** es una técnica que se basa en las redes sociales de los miembros de la población, utilizada para muestrear poblaciones ocultas como usuarios de drogas inyectables o trabajadores sexuales. Aunque el RDS puede generar una muestra de probabilidad en circunstancias ideales, en la práctica, es difícil garantizar que se cumplan todas las condiciones necesarias para obtener una muestra representativa.

Internet Surveys

El selección de muestras a través de internet es un enfoque relativamente reciente para realizar investigaciones sociales, que ha ganado gran popularidad debido a la posibilidad de obtener respuestas de grandes muestras a bajo costo y con alta velocidad. Sin embargo, los métodos de muestreo no probabilísticos que se utilizan en este tipo de encuestas generan preocupaciones sobre posibles sesgos en las estimaciones de los resultados. Las personas sin acceso a internet, o con acceso limitado, quedan excluidas de estos estudios, lo que hace que los participantes no sean una muestra representativa de la población general.

Un tipo de muestreo por internet es el muestreo en río, en el que se colocan invitaciones para participar en encuestas en varios sitios web, generalmente ofreciendo alguna forma de compensación. Este proceso de selección introduce **sesgos que cuestionan la representatividad de la muestra**, además de plantear dudas sobre la sinceridad y reflexión de las respuestas. Otro enfoque es el de los paneles de internet opt-in, donde personas son seleccionadas para participar en encuestas a lo largo del tiempo a cambio de un pago. Aunque estas encuestas no probabilísticas han sido mejoradas con métodos de ponderación y ajustes complejos, persisten las dudas sobre la representatividad de las respuestas y sobre si los datos externos utilizados para calibrar las muestras pueden reflejar con precisión a la población general.

Model-Dependent Inference

La inferencia dependiente del modelo ha ganado relevancia en el ámbito de las encuestas sociales, especialmente para abordar imperfecciones en el muestreo, como la no cobertura y la no respuesta. En particular, los **enfoques basados en modelos** se han vuelto necesarios para realizar análisis en subgrupos cuando los tamaños muestrales no son suficientes para obtener estimadores precisos a partir de métodos de muestreo basados en el diseño. Sin embargo, la adopción de estos métodos ha generado debates, ya que los estadísticos que prefieren enfoques basados en el diseño consideran que los métodos dependientes del modelo pueden ser menos confiables, sobre todo si el modelo está mal especificado. A pesar de estas preocupaciones, los avances en los métodos de ajuste, como la estimación de pequeñas áreas, han permitido que estos enfoques se utilicen de manera más frecuente, especialmente cuando se requiere estimar parámetros en áreas geográficas específicas.

El enfoque dependiente del modelo se utiliza especialmente cuando se buscan estimaciones para subgrupos pequeños o áreas administrativas definidas geográficamente, donde las muestras son demasiado pequeñas para obtener resultados precisos mediante métodos de diseño tradicionales. Estos enfoques de predicción permiten hacer estimaciones cuando los datos de los marcos de muestreo son insuficientes o no están disponibles para toda la población objetivo. Aunque estas estimaciones basadas en modelos pueden reducir la varianza en comparación con los estimadores basados en el diseño, no están exentas de sesgos si los modelos no se ajustan correctamente. A pesar de ello, la creciente aceptación de métodos como la estimación de pequeñas áreas ha sido clave para su implementación exitosa en diversas áreas, especialmente en el contexto de encuestas sociales de gran escala.

Analytic Uses of Survey Data

En la década de 1970, con la expansión de la potencia de cómputo y el software, los datos de encuestas recolectados mediante diseños complejos comenzaron a utilizarse principalmente en análisis secundarios para estudiar relaciones entre variables, buscando conexiones causales. Inicialmente, la regresión múltiple fue la técnica principal, con un enfoque en la magnitud de los coeficientes de regresión. Algunos analistas argumentaban que el interés no era para la población finita específica, sino para estimar parámetros de una superpoblación más general, donde el diseño de la muestra se volvía irrelevante, a menos que los pesos y el agrupamiento fueran variables predictoras importantes.

Con el tiempo, el uso de regresión se amplió para incluir otros modelos y técnicas multivariadas como el análisis de datos categóricos, el modelado multinivel y los análisis longitudinales. Estas técnicas han sido aplicadas a datos de encuestas complejas, permitiendo una mayor flexibilidad en el análisis de diversas relaciones entre variables, aunque su aplicación con datos complejos requiere enfoques especializados.

Administrative Records and Big Data

En los últimos tiempos, **ha aumentado el interés en el uso de registros administrativos como una fuente alternativa de datos para la investigación**. Esta alternativa ofrece ventajas notables en términos de costos y tamaño de muestra, pero también plantea importantes desafíos,

como las cuestiones de privacidad y confidencialidad. A pesar de su atractivo, los registros administrativos tienen limitaciones, como la cobertura de los datos, la consistencia en la medición de las variables, y la validez de los datos a lo largo del tiempo. Además, muchas veces estos registros no contienen toda la información necesaria para los análisis, lo que puede requerir la vinculación de varios conjuntos de registros, un proceso que conlleva sus propios problemas de calidad y confidencialidad.

Además de los registros administrativos mantenidos por el gobierno, existen otras fuentes de datos para la investigación social, como los registros de organizaciones privadas. Sin embargo, estos también presentan problemas de calidad y acceso. Otra fuente de datos relevante es el "big data", como el que se obtiene de la ubicación de los teléfonos móviles, que puede proporcionar información sobre los tiempos de viaje de los usuarios. A pesar de su atractivo, **los grandes conjuntos de datos pueden ser engañosos**, como se evidenció en el caso de Google Flu Trends, que inicialmente fue prometedor para predecir brotes de gripe, pero terminó fallando. Este tipo de ejemplos advierte sobre **la necesidad de un análisis crítico al utilizar grandes volúmenes de datos**.

Conclusiones

El debate entre muestreo probabilístico y muestreo no probabilístico fue crucial en los inicios de la investigación sobre encuestas. No fue hasta el trabajo de Neyman (1934) que el muestreo probabilístico se consolidó como el estándar para encuestas a gran escala. Este método garantiza que los resultados estén sujetos únicamente a errores de muestreo medibles, a diferencia del muestreo no probabilístico, donde siempre existe la duda sobre la representatividad de la muestra respecto a las variables de interés.

Aunque el muestreo probabilístico ofrece mayor rigor, implica mayores costos y tiempos, lo que ha llevado al desarrollo de métodos alternativos más accesibles para ciertas poblaciones. Hoy en día, se utilizan modelos que integran datos de distintas fuentes, como encuestas probabilísticas y no probabilísticas, registros administrativos y big data. Esto ha impulsado investigaciones sobre cómo combinar estas fuentes para realizar inferencias válidas y confiables.

Bibliografía

Kalton, G., 2023. Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day. Statistics in Transition new series, 24(3), pp. 1-22. https://doi.org/10.59170/stattrans-2023-029