**Documento Stalyn**

**3.12 Varianza, desviación estándar y rangos**

**3.12.a Ultimate Cluster (UC) (subsección propuesta)**

**51.** Si bien el método se diseñó en un inicio para calcular las varianzas de los estimadores de totales, también puede emplearse junto con técnicas como la linealización de Taylor y el enfoque de ecuaciones de estimación para derivar varianzas de estimadores de otras cantidades poblacionales que puedan formularse como la solución de una ecuación de estimación. Gracias a esta característica, el método resulta flexible y aplicable en múltiples contextos de análisis de encuestas de hogares.

**3.12 Varianza, desviación estándar y rangos**

**3.12.a Ultimate Cluster (UC) (subsección propuesta)**

**52.** Por otra parte, un supuesto fundamental de este método establece que, dentro de cada estrato, las unidades primarias de muestreo (PSU) se eligen de forma independiente y con reemplazo. No obstante, en la práctica, la mayoría de las encuestas de hogares realizan la selección de PSU sin reemplazo, lo que suele generar diseños más eficientes. Por ello, las estimaciones de varianza que parten de la premisa de independencia constituyen aproximaciones de las verdaderas varianzas de muestreo. Cuando la fracción de muestreo es reducida (por ejemplo, menor al 5 %), dichas aproximaciones suelen ser apropiadas y con un nivel de precisión suficiente para su uso por parte de las oficinas nacionales de estadística o de analistas secundarios.

**3.12 Varianza, desviación estándar y rangos**

**3.12.a Ultimate Cluster (UC) (subsección propuesta)**

**53.** El método ***Ultimate Cluster*** destaca por su simplicidad, lo que lo hace muy atractivo. En la práctica, quienes realizan encuestas suelen preferirlo frente a métodos más sofisticados que consideran cada etapa del diseño de muestreo. Si bien estos procedimientos más detallados pueden ofrecer estimaciones de varianza algo más precisas, su aplicación es más compleja y demanda información más completa sobre el diseño muestral. En cambio, el método proporciona una aproximación razonable que suele ser suficiente en la mayoría de los casos prácticos, especialmente al estimar totales o medias. Un análisis sobre la precisión de esta aproximación y las posibles alternativas se presenta en Särndal, Swensson y Wretman (1992, p. 153).

**3.12.1 Métodos de replicación y estimación de varianzas (subsección propuesta)**

**54.** Para proteger la confidencialidad, los microdatos de encuestas disponibles públicamente suelen excluir información esencial del diseño (por ejemplo, identificadores de estratos o de unidades primarias de muestreo), lo que limita la capacidad de los usuarios para obtener estimaciones de varianza válidas. En estos casos, es recomendable que las oficinas nacionales de estadística (NSO) proporcionen pesos de replicación, lo que permite a los analistas calcular errores estándar apropiados. Sin variables de diseño ni pesos de replicación, los usuarios secundarios no pueden reproducir los errores estándar publicados ni considerar de manera adecuada el diseño complejo de la encuesta en la estimación de varianzas.

**3.12.1 Métodos de replicación y estimación de varianzas (subsección propuesta)**

**55.** Los métodos de replicación para estimar la varianza se fundamentan en generar subconjuntos de la muestra original, obtener estimaciones para cada uno de ellos y emplear la variabilidad observada entre dichas estimaciones para aproximar la varianza del estimador principal. Estos métodos resultan particularmente valiosos cuando la base de datos carece de información sobre estratos o identificadores de UPM, situación en la que no es posible aplicar el método de Última Conglomeración.

**3.12.1 Métodos de replicación y estimación de varianzas (subsección propuesta)**

**56.** El método bootstrap constituye una herramienta de replicación robusta y versátil para calcular varianzas tanto en encuestas como en otros ámbitos de análisis. Aunque fue introducido por Efron (1979) para datos que no provenían de encuestas, la adaptación más utilizada en encuestas de hogares es el **Bootstrap de Reescalamiento de Rao-Wu-Yue** (Rao, Wu y Yue, 1992), el cual se ajusta de manera óptima a diseños de muestreo estratificados y multietápicos, siendo una de las técnicas más empleadas para la estimación de varianzas en encuestas complejas.

**3.12.1 Métodos de replicación y estimación de varianzas (subsección propuesta)**

**57.** El método consiste en generar muchas réplicas de la muestra original, creando conjuntos de datos que simulan extracciones repetidas de la población. Para lograrlo, se añaden varias columnas de pesos de replicación en la base de datos original, cada una correspondiente a un remuestreo diferente. Esta tarea, que debe ser realizada por la Oficina Nacional de Estadística (ONE), sigue un proceso sencillo:

Primero, para cada estrato, se seleccionan aleatoriamente las UPM de la muestra original con reemplazo, de modo que algunas pueden repetirse y otras no aparecer. Cada UPM elegida se incorpora junto con todas sus observaciones. Cuando el tamaño de la muestra de primera etapa en el estrato *h* es mayor que dos (nh>2n\_h > 2), el número de UPM seleccionadas en cada réplica es nh−1n\_h-1, una menos que en la muestra inicial.

Este procedimiento se repite muchas veces, habitualmente cientos, para generar un gran número de réplicas. La cantidad de veces que una UPM *i* del estrato *h* aparece en la réplica *r* se representa como nhi(r)n\_{hi}^{(r)}, y puede variar entre 0 y nh−1n\_h-1.

A partir de cada réplica se calculan pesos bootstrap nuevos para todas las unidades, que reflejan cuántas veces fue seleccionada su UPM en esa réplica. Así se garantiza que las réplicas sigan representando a la población. El peso de la unidad *k* en la réplica *r* se obtiene mediante:

whik(r)=whik×nhnh−1×nhi(r)w\_{hik}^{(r)} = w\_{hik} \times \frac{n\_h}{n\_h-1} \times n\_{hi}^{(r)}

Por último, si los pesos originales ya incluyen ajustes por no respuesta o calibración, esos mismos ajustes deben aplicarse a cada conjunto de pesos bootstrap, para capturar la incertidumbre adicional que dichos ajustes generan.

**3.12.1 Métodos de replicación y estimación de varianzas (subsección propuesta)**

**58.** Cuando la Oficina Nacional de Estadística no incluye en la base de datos los identificadores de estratos o UPM, pero sí entrega los pesos de replicación Bootstrap, los investigadores pueden seguir calculando errores estándar de manera adecuada. En cada réplica *r*, el parámetro de interés θ\theta se estima como θ^(r)\hat{\theta}^{(r)}, empleando los pesos bootstrap whik(r)w\_{hik}^{(r)} en lugar de los pesos muestrales originales whikw\_{hik}. Posteriormente, la variabilidad entre todas las estimaciones obtenidas a partir de las réplicas se utiliza para aproximar la varianza del estimador original, de forma que la dispersión observada entre réplicas refleja el grado de incertidumbre del parámetro.

El estimador de la varianza mediante bootstrap se expresa como:

V^B(θ^)=1R∑r=1R(θ^(r)−θ~)2\hat{V}\_B(\hat{\theta}) = \frac{1}{R} \sum\_{r=1}^{R} \left(\hat{\theta}^{(r)} - \tilde{\theta}\right)^2

donde

θ~=1R∑r=1Rθ^(r)\tilde{\theta} = \frac{1}{R} \sum\_{r=1}^{R} \hat{\theta}^{(r)}

corresponde al promedio de todas las estimaciones replicadas, lo que asegura que la medida de variabilidad capture fielmente la incertidumbre del parámetro.

**3.12.1 Métodos de replicación y estimación de varianzas (subsección propuesta)**

**59.** El método Bootstrap ofrece múltiples beneficios. Aunque exige un procesamiento computacional considerable, ya que implica generar y analizar un gran número de réplicas, es muy eficaz para el análisis de diseños de encuesta complejos y permite estimar una amplia gama de parámetros, incluso aquellos de difícil cálculo con métodos tradicionales, como las medianas u otras estadísticas no lineales. Además, representa una solución práctica para la estimación de varianzas cuando otros métodos no están disponibles o resultan poco viables. Este enfoque resulta particularmente útil para los analistas que trabajan con bases de datos de encuestas que no incluyen **identificadores de estratos y de UPM,** pero que sí disponen de un conjunto de pesos de replicación.

**3.12.1 Métodos de replicación y estimación de varianzas (subsección propuesta)**

**60.** La sencillez del método Bootstrap facilita que los investigadores puedan estimar varianzas aun sin utilizar programas estadísticos especializados. Sin embargo, en la actualidad la mayoría de los paquetes estadísticos incluyen procedimientos para aplicar la replicación Bootstrap y calcular varianzas, lo que amplía su disponibilidad y permite obtener estimaciones robustas incluso para parámetros de alta complejidad. Esto lo convierte en una de las herramientas más versátiles para el análisis de encuestas. Pese a ello, su aplicación no es la más recomendable en encuestas repetidas con muestras que se superponen ni en situaciones donde las fracciones de muestreo son grandes y los tamaños de muestra son reducidos (Bruch, 2011).

**3.12.b Efecto del diseño (DEFF)**

**61.** De acuerdo con Kish (1965, p. 258), el efecto del diseño (DEFF) se define como la relación entre la varianza de un estimador calculado bajo un diseño de muestreo complejo y la varianza del mismo estimador cuando se emplea un muestreo aleatorio simple (MAS) con igual tamaño de muestra. Su estimación se formula como:

DEFF^=V^p(θ^)V^SRS(θ^)\widehat{\text{DEFF}} = \frac{\widehat{V}\_{p}(\hat{\theta})}{\widehat{V}\_{\text{SRS}}(\hat{\theta})}

donde V^p(θ^)\widehat{V}\_{p}(\hat{\theta}) corresponde a la varianza estimada de θ^\hat{\theta} bajo el diseño complejo p(s)p(s), mientras que V^SRS(θ^)\widehat{V}\_{\text{SRS}}(\hat{\theta}) representa la varianza estimada del mismo estimador bajo un MAS de igual tamaño. Tal como se discutió en el Capítulo 5, este indicador permite cuantificar cuánto aumenta la varianza como consecuencia de la conglomeración y otras características propias de los diseños complejos en comparación con un muestreo simple. Según Naciones Unidas (2008, p. 49), el DEFF puede entenderse de tres maneras: como el factor de incremento de la varianza frente al MAS, como una medida de la pérdida relativa de precisión o como una indicación del aumento en el tamaño de la muestra que sería necesario en un diseño complejo para alcanzar el mismo nivel de varianza que en un MAS.

**3.12.b Efecto del diseño (DEFF)**

**62.** Según Park et al. (2003), el efecto del diseño de una encuesta puede desglosarse en tres factores multiplicativos. El primero está relacionado con la ponderación desigual, ya que la presencia de pesos muestrales no uniformes suele incrementar ligeramente la varianza; por ello, el uso de pesos iguales resulta ventajoso y explica por qué los diseños auto-ponderados son preferidos en encuestas de hogares. El segundo factor corresponde a la estratificación, la cual, cuando se aplica correctamente, puede disminuir la varianza, aunque en la práctica su efecto reductor suele ser moderado. Finalmente, el tercer factor se asocia con el muestreo en varias etapas, que generalmente incrementa la varianza, dado que las unidades que pertenecen a un mismo conglomerado tienden a ser más homogéneas entre sí que en comparación con las de otros conglomerados.

**3.12.b Efecto del diseño (DEFF)**

**63.** Al analizar encuestas, el DEFF se convierte en un indicador fundamental para medir la calidad de las estimaciones y orientar el diseño de estudios futuros. Cuando su valor es elevado, evidencia que el diseño complejo introduce ineficiencias que incrementan la varianza y reducen la precisión de los resultados. Por el contrario, un valor cercano a uno revela que el diseño tiene un efecto mínimo sobre la varianza. Esta información permite a los investigadores identificar si es necesario ajustar la ponderación, optimizar la estratificación o modificar el tamaño del submuestreo para aumentar la eficiencia en levantamientos posteriores.

**3.12.b Efecto del diseño (DEFF)**

**64.** La interpretación de un **DEFF** alto debe hacerse con precaución, pues no siempre implica que el diseño muestral sea inadecuado. Es fundamental considerar el contexto de la encuesta; por ejemplo, un valor superior a tres podría parecer alarmante, pero en muchos casos se debe a limitaciones prácticas, como restricciones presupuestales, dificultades logísticas o la necesidad de garantizar la participación de los entrevistados. En ciertos levantamientos de hogares, puede ser indispensable seleccionar solo una fracción de los individuos elegibles en cada hogar en lugar de incluirlos a todos. Asimismo, situaciones propias del trabajo de campo, como problemas de cobertura o tasas de no respuesta, pueden aumentar la variabilidad de los pesos muestrales (véase el Capítulo 8) y, en consecuencia, elevar los valores del DEFF.

**3.8 Algunas reflexiones generales**

**65.** Para garantizar estimaciones oficiales precisas y confiables, el análisis de encuestas de hogares requiere un uso intensivo de herramientas computacionales especializadas. En esta sección se presentan algunos de los enfoques implementados en programas estadísticos que apoyan los distintos procesos involucrados en la producción de resultados. Entre los aspectos más relevantes para los analistas se incluyen la modelación de la no respuesta y la imputación de datos, el cálculo de errores estándar para cada indicador que se reportará en las tablas de salida y el estudio de las relaciones multivariadas entre las variables de la encuesta.

**4.1 Lectura de bases de datos y definición del diseño muestral**

**66.** Según Naciones Unidas (2005, sec. 7.8), es fundamental que la estructura de los diseños de muestreo complejos se tenga en cuenta en el proceso de inferencia al estimar estadísticas oficiales basadas en encuestas de hogares. El organismo advierte, a través de un ejemplo empírico, que ignorar este aspecto puede generar estimaciones sesgadas y errores de muestreo subestimados. En este contexto, se señalan a continuación las principales funcionalidades que ofrecen los programas estadísticos para el manejo de datos provenientes de este tipo de diseños. Para una revisión más exhaustiva, que incluye ejemplos de sintaxis y código computacional, se recomienda consultar Heeringa, West y Berglund (2017, Apéndice A).

**4.1 Lectura de bases de datos y definición del diseño muestral**

**67.** De manera general, herramientas estadísticas como R, Stata, SAS y SPSS cuentan con módulos y bibliotecas que optimizan la estimación de varianzas en muestras complejas, incorporando incluso métodos de replicación para varianzas basadas en el diseño. Mientras que R es un software de acceso libre, los otros programas requieren licencias de pago. Además de permitir el cálculo de estadísticas descriptivas, como medias, totales, proporciones, percentiles y razones, estas plataformas posibilitan el ajuste de modelos de regresión considerando la estructura del diseño de la encuesta. Los programas especializados en análisis de encuestas generan de forma automática el efecto del diseño, lo que apoya a los investigadores en la interpretación de la variabilidad de sus estimaciones.

**4.1 Lectura de bases de datos y definición del diseño muestral**

**68.** El uso de estos paquetes y herramientas computacionales implica que el usuario suministre información clave del diseño muestral, como los factores de expansión, la estratificación y los identificadores de conglomerados. Seguidamente, se ofrece un resumen general, aunque no exhaustivo, de las funciones y posibilidades que brindan los principales programas estadísticos.

**3.2 Algunas librerías de interés**

**69**. R es un software libre y de código abierto que ha ganado gran popularidad en el procesamiento de encuestas y la investigación social, convirtiéndose en una herramienta de elección para aplicar los desarrollos científicos y metodológicos más recientes en el análisis de datos de encuestas (R Core Team, 2024). Su carácter abierto permite que investigadores de todo el mundo aporten funciones y paquetes propios al **Comprehensive R Archive Network (CRAN),** poniéndolos a disposición de la comunidad académica y profesional. Entre sus recursos más destacados se encuentra el paquete **samplesize4surveys** (Gutiérrez, 2020), que facilita el cálculo de tamaños de muestra para individuos y hogares en encuestas repetidas, de panel y rotacionales. Asimismo, los paquetes **sampling** (Tillé y Matei, 2016) y **TeachingSampling** (Gutiérrez, 2015) ofrecen soporte para seleccionar muestras probabilísticas a partir de marcos de muestreo bajo diferentes diseños y algoritmos. Para el análisis de datos de encuestas de hogares, el paquete **survey** (Lumley, 2024) permite especificar el diseño muestral mediante la función *svydesign() y* obtener estimaciones correctas de errores estándar. El paquete **convey** (Pessoa et al., 2024) complementa este proceso al facilitar el cálculo de medidas de desigualdad. En el ámbito del modelado de regresiones, **svydiags** (Valliant, 2024) incluye herramientas de diagnóstico como análisis de residuos, valores de apalancamiento, factores de inflación de varianza y pruebas de colinealidad, mientras que **PracTools** (Valliant et al., 2025) proporciona utilidades para el cálculo del tamaño de muestra, el diseño de muestreo, la estimación de efectos de diseño y el análisis de componentes de varianza en esquemas multietápicos.

**3.7 Medidas descriptivos y reflexiones**

**73**. Al analizar datos de encuestas de hogares, uno de los productos más habituales son los **parámetros descriptivos,** cuyo propósito es sintetizar las principales características de la población. Estas estimaciones permiten ofrecer una representación clara y comprensible de la realidad poblacional a partir de la información obtenida en una muestra representativa.

**3.7 Medidas descriptivos y reflexiones**

**74.** Entre los resultados más comunes de este tipo de análisis se encuentran las frecuencias, proporciones, medias y totales. Las medias proporcionan información sobre el valor promedio de una variable, mientras que los totales reflejan su acumulado en toda la población. Las frecuencias cuentan cuántos hogares o individuos pertenecen a una categoría determinada —por ejemplo, el número de personas en situación de pobreza—, y las proporciones expresan la participación relativa de quienes presentan una característica específica, como el porcentaje de población pobre.

**3.7 Medidas descriptivos y reflexiones**

**75.** Actualmente, el análisis descriptivo va más allá de los parámetros básicos, incorporando métricas más complejas. Se estiman cuantiles de variables numéricas, como la mediana del ingreso de los hogares, para describir la distribución de los datos con mayor detalle. Además, se aplican indicadores especializados para evaluar fenómenos concretos, como los índices FGT para la medición de la pobreza, los indicadores de desigualdad (Gini, Theil, Atkinson) y los de polarización (Wolfson, DER), entre otros (Jacob, Damico y Pessoa, 2024).

**5.1 Estimaciones de totales**

**76.** En el análisis de encuestas de hogares, resulta esencial determinar el tamaño de las subpoblaciones, es decir, identificar cuántas personas u hogares pertenecen a categorías específicas y qué proporción representan dentro del total poblacional. Este tipo de estimaciones permite caracterizar el perfil demográfico y socioeconómico de la población, información que es clave para orientar la asignación de recursos, el diseño de políticas públicas y la formulación de programas sociales.

**5.1 Estimaciones de totales**

**77.** Así, es de gran utilidad conocer cuántas personas se encuentran por debajo de la línea de pobreza, cuántas no tienen empleo o cuántas han alcanzado determinado nivel educativo. Estos datos permiten atender las desigualdades, diseñar intervenciones más focalizadas y promover un desarrollo más equitativo. Analizar cómo se distribuyen los individuos entre las distintas categorías ofrece información indispensable para reducir brechas y avanzar en la construcción de un desarrollo inclusivo.

**5.1 Estimaciones de totales**

**78.** La estimación del tamaño de una población o subpoblación se realiza a partir de variables categóricas, las cuales segmentan a la población en grupos mutuamente excluyentes. Estas categorías pueden corresponder, por ejemplo, a quintiles de ingreso, estados de ocupación o niveles educativos alcanzados. El tamaño poblacional hace referencia al número total de individuos u hogares que, en la base de datos de la encuesta, pertenecen a una categoría determinada. Para obtener estas estimaciones, se combinan las respuestas de los encuestados con los pesos muestrales, que indican cuántas personas u hogares representa cada unidad de la muestra dentro de la población total.

El estimador del tamaño de la población se define como:

N^=∑h=1H∑i∈s1h∑k∈shiwhik\hat{N} = \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik}

donde shis\_{hi} corresponde a la muestra de hogares o individuos en la UPM ii del estrato hh; s1hs\_{1h} representa la muestra de UPM seleccionadas en el estrato hh; y whikw\_{hik} es el peso o factor de expansión de la unidad kk en la UPM ii del estrato hh.

**5.1 Estimaciones de totales**

**79.** La estimación del tamaño de una subpoblación sigue el mismo principio que el cálculo del tamaño poblacional total, pero se enfoca en un subconjunto definido por una característica específica. Así, para determinar cuántas personas pertenecen a una categoría en particular, se identifica dicho grupo en la base de datos de la encuesta y se suman sus pesos muestrales. Este procedimiento permite no solo conocer el tamaño total de la población, sino también cuantificar grupos de interés específicos.

Para llevarlo a cabo, se construye una variable binaria I(yhik=d)I(y\_{hik}=d), que toma el valor de uno si la unidad kk de la UPM ii en el estrato hh pertenece a la categoría dd —la cual no fue considerada en el ajuste de los pesos— de la variable discreta yy, y cero en caso contrario. El estimador muestral de este parámetro se expresa como:

N^d=∑h=1H∑i∈s1h∑k∈shiwhikI(yhik=d)\hat{N}\_d = \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} I(y\_{hik}=d)

En el caso de que dd sea una categoría incluida en el proceso de calibración de los pesos, el valor de N^d\hat{N}\_d coincidirá con el control externo utilizado y no debe interpretarse como una estimación.

**5.2 Estimación de proporciones**

**80.** Las proporciones permiten expresar el peso relativo que tienen determinados grupos dentro de la población. Por ejemplo, conocer el porcentaje de hogares que se encuentran por debajo de la línea de pobreza es esencial para evaluar las desigualdades socioeconómicas. Para obtener este indicador, se calcula el promedio ponderado de una variable dicotómica, lo que asegura que la estimación represente adecuadamente la distribución poblacional.

De acuerdo con Heeringa, West y Berglund (2017), al transformar las categorías de respuesta originales en variables indicadoras yy con valores de 1 y 0 (por ejemplo, 1 = Sí y 0 = No), la proporción estimada se obtiene mediante:

p^d=N^dN^=∑h=1H∑i∈s1h∑k∈shiwhikI(yhik=d)∑h=1H∑i∈s1h∑k∈shiwhik\hat{p}\_d = \frac{\hat{N}\_d}{\hat{N}} = \frac{\displaystyle\sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} I(y\_{hik}=d)} {\displaystyle\sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik}}

Dado que se trata de un estimador no lineal, su varianza puede aproximarse mediante la técnica de linealización de Taylor, utilizando como función de estimación zhik=I(yhik=d)−p^dz\_{hik}=I(y\_{hik}=d)-\hat{p}\_d. Actualmente, la mayoría de los programas estadísticos generan este tipo de proporciones junto con sus errores estándar, generalmente presentados en forma de porcentajes.

**5.2 Estimación de proporciones**

**81.** En situaciones donde las proporciones estimadas se aproximan a 0 o 1, se vuelve necesario aumentar el tamaño de la muestra para garantizar resultados estadísticamente sólidos. Además, es recomendable aplicar métodos que aseguren que los intervalos de confianza permanezcan en el rango [0,1], ya que los intervalos convencionales basados en la normal pueden desbordar estos límites y perder su utilidad interpretativa. Para resolver este inconveniente, se han desarrollado alternativas para la construcción de intervalos de confianza, como las propuestas por Rust y Hsu (2007) y Dean y Pagano (2015).

Una de estas alternativas recurre a la transformación logit de la proporción estimada y se expresa de la siguiente forma:

CI(p^d;1−α)=exp⁡[ln⁡(p^d1−p^d)±me(p^d)p^d(1−p^d)]1+exp⁡[ln⁡(p^d1−p^d)±me(p^d)p^d(1−p^d)]CI(\hat{p}\_d; 1-\alpha) = \frac{\exp \left[\ln\left(\frac{\hat{p}\_d}{1-\hat{p}\_d}\right) \pm \frac{me(\hat{p}\_d)}{\hat{p}\_d(1-\hat{p}\_d)}\right]}{1 + \exp \left[\ln\left(\frac{\hat{p}\_d}{1-\hat{p}\_d}\right) \pm \frac{me(\hat{p}\_d)}{\hat{p}\_d(1-\hat{p}\_d)}\right]}

donde me(p^d)me(\hat{p}\_d) representa el margen de error del estimador y se obtiene mediante

me(p^d)=t1−α/2,df×se(p^d),me(\hat{p}\_d) = t\_{1-\alpha/2, df} \times se(\hat{p}\_d),

siendo t1−α/2,dft\_{1-\alpha/2, df} el cuantil de la distribución t de Student con dfdf grados de libertad, dejando un área α/2\alpha/2 a su derecha. Los grados de libertad se calculan restando el número de estratos al número total de UPM (df=n−Hdf = n - H).

Existen además otros métodos que permiten calcular intervalos de confianza incluso cuando la proporción es exactamente cero o uno, asegurando resultados interpretables en casos extremos.

**3.10 Medias y totales**

**82**. Al trabajar con encuestas de hogares, el análisis de datos numéricos implica con frecuencia calcular estadísticas descriptivas como medias, totales y razones, ya que estas permiten sintetizar las principales características de la población y sirven de base para la toma de decisiones. Dichas estimaciones pueden calcularse para la población en su conjunto o para subgrupos específicos, de acuerdo con los propósitos de la investigación. Tal como destacan Heeringa, West y Berglund (2017), el cálculo de totales y medias poblacionales, junto con sus varianzas, ha sido esencial para el desarrollo de la teoría del muestreo probabilístico.

**3.10 Medias y totales**

**83.** La determinación de los totales poblacionales constituye uno de los pilares del análisis de encuestas. Tanto las medias como las proporciones y las razones derivan de los totales. Un total se define como la suma de una variable específica, por ejemplo, el ingreso o el gasto, a nivel de toda la población. Para estimar el ingreso total de todos los hogares de un país, se combinan los datos de la muestra aplicando los pesos muestrales que reflejan el diseño y aseguran representatividad. En el caso de variables numéricas simples, las estimaciones básicas son los totales y las medias, mientras que las razones permiten establecer comparaciones entre dos variables numéricas. Estos cálculos pueden hacerse para toda la población o de manera desagregada por dominios de estudio, dependiendo de las preguntas de investigación.

**3.10 Medias y totales**

**84.** Una vez establecido el diseño muestral, se realiza el cálculo de los parámetros de interés. En el caso de encuestas con diseños complejos que incorporan estratificación (h=1,2,...,Hh=1,2,...,H) y submuestreo en las UPM —ubicadas dentro de cada estrato hh e identificadas por ii—, el total poblacional se estima mediante la expresión:

Y^=∑h=1H∑i∈s1h∑k∈shiwhik yhik\hat{Y} = \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} \, y\_{hik}

Cuando se cuenta con respuesta completa, la estimación de la varianza de Y^\hat{Y} puede obtenerse aplicando el estimador de ***Ultimate Cluster*** descrito previamente en la Sección 9.2.

**3.10 Medias y totales**

**85.** El intervalo de confianza de nivel 1−α1-\alpha para el total poblacional YY se calcula mediante la expresión

Y^±t1−α/2,df×V^UC(Y^),\hat{Y} \pm t\_{1-\alpha/2, df} \times \sqrt{\hat{V}\_{UC}(\hat{Y})},

como se muestra en la ecuación (9-15). A medida que los grados de libertad aumentan, la distribución t de Student tiende a la normal, lo que explica que muchas Oficinas Nacionales de Estadística (ONE) utilicen esta aproximación para reportar intervalos de confianza. No obstante, es importante considerar que esta aproximación puede ser menos fiable cuando el tamaño de muestra es reducido, aunque suele ofrecer buenos resultados en encuestas de hogares extensas.

**3.10 Medias y totales**

**86.** Las medias o promedios poblacionales son igualmente esenciales para describir la tendencia central de una variable. Por ejemplo, el promedio de gasto de los hogares es un indicador representativo del comportamiento económico de la población. Su cálculo consiste en dividir el total de la variable entre el tamaño de la población, de modo que su exactitud depende de que ambos componentes se estimen correctamente. El estimador de la media poblacional se obtiene como la razón de dos totales:

Yˉ^=∑h=1H∑i∈s1h∑k∈shiwhikyhik∑h=1H∑i∈s1h∑k∈shiwhik=Y^N^.\widehat{\bar{Y}} = \frac{\displaystyle \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} y\_{hik}} {\displaystyle \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik}} = \frac{\hat{Y}}{\hat{N}}.

Dado que Yˉ^\widehat{\bar{Y}} es un estimador de naturaleza no lineal, su varianza exacta no puede expresarse en forma cerrada. Por esta razón, es necesario recurrir a técnicas de remuestreo, como el método bootstrap, o a aproximaciones basadas en series de Taylor. En este último caso, se utiliza su ecuación de estimación muestral

∑h=1H∑i∈s1h∑k∈shiwhik(yhik−Yˉ^)=0,\sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} (y\_{hik} - \widehat{\bar{Y}}) = 0,

lo que permite aplicar el estimador de varianza descrito en la Sección 9.2, considerando zhik=yhik−Yˉ^z\_{hik} = y\_{hik} - \widehat{\bar{Y}}. Estas estimaciones se calculan automáticamente en la mayoría de los paquetes estadísticos diseñados para manejar encuestas complejas.

**4.7 Estimando razones**

**87**. Las razones permiten expresar la relación entre dos variables. Un ejemplo común es la razón entre el gasto y el ingreso de los hogares, que ayuda a identificar patrones de consumo. La exactitud de este indicador depende de la correcta estimación de los totales que conforman el numerador y el denominador. Este tipo de medida resulta particularmente valiosa para construir indicadores que faciliten la comparación entre distintos grupos poblacionales o el monitoreo de cambios a lo largo del tiempo. Así, por ejemplo, el indicador 2.1.1 de los Objetivos de Desarrollo Sostenible (prevalencia de subalimentación) puede calcularse a partir de la razón entre el consumo de alimentos, medido en calorías o energía ingerida, y los requerimientos energéticos de la dieta, determinados según edad, sexo y nivel de actividad física.

El estimador puntual de la razón se obtiene dividiendo el estimador del total de la variable de interés entre el estimador del total de la variable de referencia:

R^=Y^X^=∑h=1H∑i∈s1h∑k∈shiwhikyhik∑h=1H∑i∈s1h∑k∈shiwhikxhik\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\displaystyle \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} y\_{hik}} {\displaystyle \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} x\_{hik}}

Para calcular la varianza de esta razón, se debe definir la función de estimación como zhik=yhik−R^xhikz\_{hik} = y\_{hik} - \hat{R}x\_{hik} —donde yy y xx son las variables del numerador y denominador— y emplear el estimador de varianza descrito en la Sección 9.2.

**4.5 Análisis de la relación entre dos variable continuas**

**88.** Para explorar cómo se relacionan dos variables numéricas en datos de encuestas, el análisis de correlación es una de las técnicas más empleadas. Así, por ejemplo, puede resultar de interés examinar si el ingreso de los hogares está asociado con su nivel de gasto y en qué medida. El coeficiente de correlación de Pearson, que varía entre –1 y 1, permite cuantificar tanto la fuerza como la dirección de esta relación lineal. Un valor positivo indica que ambas variables tienden a aumentar al mismo tiempo, un valor negativo señala que cuando una se incrementa la otra tiende a disminuir y valores próximos a cero sugieren una relación lineal débil o inexistente.

**4.5 Análisis de la relación entre dos variable continuas**

**89.** Para que la correlación obtenida sea representativa de la población y no solo de la muestra, en el análisis de encuestas se incorporan los pesos muestrales. Este procedimiento ajusta la estimación a los efectos del diseño complejo, contemplando la estratificación, la conglomeración y las probabilidades de selección desiguales. El cálculo del coeficiente implica evaluar la covarianza entre las dos variables y normalizarla por sus varianzas individuales, lo que elimina la influencia de las unidades de medida y facilita la interpretación.

**4.5 Análisis de la relación entre dos variable continuas**

**90**. Formalmente, el coeficiente de correlación de Pearson para dos variables numéricas de encuesta xx y yy se calcula mediante:

ρ^xy=∑h=1H∑i∈s1h∑k∈shiwhik(yhik−Yˉ^)(xhik−Xˉ^)∑h=1H∑i∈s1h∑k∈shiwhik(yhik−Yˉ^)2∑h=1H∑i∈s1h∑k∈shiwhik(xhik−Xˉ^)2\hat{\rho}\_{xy} = \frac{\displaystyle \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} (y\_{hik} - \widehat{\bar{Y}})(x\_{hik} - \widehat{\bar{X}})} {\sqrt{\displaystyle \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} (y\_{hik} - \widehat{\bar{Y}})^2} \sqrt{\displaystyle \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} (x\_{hik} - \widehat{\bar{X}})^2}}

Cuando las variables son categóricas u ordinales, se utilizan otras medidas de asociación. De igual manera, es posible ajustar modelos estadísticos para estudiar la relación entre una variable de respuesta y sus covariables, como se explica en la subsección 9.5.

**3.11 Medianas y percentiles**

**91.** Para estudiar la distribución de los datos más allá de la media, los percentiles y cuantiles son herramientas de gran utilidad, pues permiten segmentar la información en partes que muestran cómo se distribuyen los valores en la población. El percentil 10, por ejemplo, indica el punto por debajo del cual se ubica el 10 % de las observaciones, mientras que la mediana (percentil 50) divide la muestra en dos grupos de igual tamaño. Estas métricas no solo describen la tendencia central, sino también la dispersión y variabilidad de los datos, lo que resulta valioso para orientar políticas públicas, como gravar al 10 % de mayores ingresos o diseñar programas de subsidios para el 15 % más vulnerable. Su cálculo se basa en la función de distribución acumulada (CDF), que representa la proporción de individuos con valores menores o iguales a un umbral dado. Una vez estimada la CDF empleando los datos de la encuesta y sus pesos, se pueden derivar los cuantiles.

El estimador de la CDF en un diseño complejo se expresa como:

F^(t)=∑h=1H∑i∈s1h∑k∈shiwhikI(yhik≤t)∑h=1H∑i∈s1h∑k∈shiwhik(9−19)\hat{F}(t) = \frac{\displaystyle\sum\_{h=1}^{H}\sum\_{i\in s\_{1h}}\sum\_{k\in s\_{hi}}w\_{hik}I(y\_{hik}\leq t)} {\displaystyle\sum\_{h=1}^{H}\sum\_{i\in s\_{1h}}\sum\_{k\in s\_{hi}}w\_{hik}} \quad (9-19)

donde I(yhik≤t)I(y\_{hik}\leq t) toma valor 1 si yhiky\_{hik} es menor o igual a tt, y 0 en caso contrario. De acuerdo con Heeringa, West y Berglund (2017), para estimar cuantiles se ordenan los datos y(1),…,y(n)y\_{(1)},…,y\_{(n)} y se identifica jj tal que

F^(y(j))≤q≤F^(y(j+1)),\hat{F}(y\_{(j)})\leq q\leq \hat{F}(y\_{(j+1)}),

obteniéndose el estimador del cuantil qq-ésimo como

y^(q)=y(j)+q−F^(y(j))F^(y(j+1))−F^(y(j))(y(j+1)−y(j)).\hat{y}\_{(q)} = y\_{(j)} + \frac{q-\hat{F}(y\_{(j)})}{\hat{F}(y\_{(j+1)})-\hat{F}(y\_{(j)})}(y\_{(j+1)}-y\_{(j)}).

Al ser medidas no lineales, los cuantiles requieren métodos más sofisticados para estimar su varianza. Kovar, Rao y Wu (1988) recomiendan el uso de métodos de replicación para calcular errores estándar en este contexto.

**4.4 Estimación del coeficiente de Ginni en encuestas de hogares**

**92.** El estudio de la desigualdad económica es prioritario para gobiernos y organismos internacionales. Entre las métricas más utilizadas destaca el coeficiente de Gini, que compara la distribución de ingresos observada con una distribución perfectamente equitativa. En encuestas de hogares, su cálculo incorpora pesos muestrales ajustados al diseño de la encuesta, y con frecuencia se normalizan para simplificar el procesamiento. Sus valores oscilan entre 0 (igualdad perfecta) y 1 (máxima desigualdad). Este indicador es clave para evaluar cambios en la distribución del ingreso a lo largo del tiempo y para comparar niveles de desigualdad entre regiones o países.

**4.4 Estimación del coeficiente de Ginni en encuestas de hogares**

**93.** De acuerdo con Binder y Kovacevic (1995), el estimador del coeficiente de Gini se define como

G^=2∑h=1H∑i∈s1h∑k∈shiwhik∗F^hikyhik−1Yˉ^(9−22)\hat{G} = \frac{2\sum\_{h=1}^{H}\sum\_{i\in s\_{1h}}\sum\_{k\in s\_{hi}}w\_{hik}^{\*}\hat{F}\_{hik}y\_{hik}-1}{\widehat{\bar{Y}}} \quad (9-22)

donde whik∗w\_{hik}^{\*} es el peso normalizado:

whik∗=whik∑h=1H∑i∈s1h∑k∈shiwhik(9−23)w\_{hik}^{\*}=\frac{w\_{hik}} {\displaystyle\sum\_{h=1}^{H}\sum\_{i\in s\_{1h}}\sum\_{k\in s\_{hi}}w\_{hik}} \quad (9-23)

y F^hik\hat{F}\_{hik} representa la CDF estimada para el individuo kk en la UPM ii del estrato hh. Osier (2009) y Langel y Tillé (2013) ofrecen detalles técnicos relevantes para estimar la varianza de este indicador complejo.

**5.3 Tablas cruzadas**

**94.** En los levantamientos de encuestas de hogares, es habitual recopilar información sobre variables categóricas, que permiten clasificar a la población en grupos mutuamente excluyentes. Ejemplos comunes son el estado laboral (“ocupado”, “desempleado”, “inactivo”), el nivel educativo alcanzado (“primaria”, “secundaria”, “terciaria”) o el acceso a determinados servicios (“Sí”, “No”). Explorar si dos de estas variables están asociadas constituye un elemento central del análisis de encuestas, pues ofrece información valiosa para distintas aplicaciones: en política pública (relacionando educación y empleo para diseñar estrategias laborales), en evaluación de programas (detectando variaciones en el acceso a salud según el nivel de ingresos) o en investigación social (estudiando vínculos entre factores demográficos y servicios) para comprender dinámicas y tendencias sociales.

**5.3 Tablas cruzadas**

**95.** El estudio de la asociación entre dos variables categóricas implica verificar si la distribución de una depende de la otra. Para ello, se comparan las frecuencias de todas las combinaciones posibles de categorías. Por ejemplo, es posible contabilizar cuántos individuos corresponden simultáneamente a cada nivel educativo y estado laboral. Estas frecuencias pueden transformarse en proporciones que muestran la participación relativa de cada combinación en la población. Como punto de partida, suele emplearse una tabla de contingencia, que organiza en forma de cuadrícula los conteos o proporciones para cada cruce de categorías; en ella, un eje puede representar el estado laboral y el otro los niveles de educación.

**5.3 Tablas cruzadas**

**96.** Para formalizar este análisis, se introduce una notación específica. Sean xx y yy dos variables categóricas con RR y CC categorías, respectivamente. Para plantear pruebas de hipótesis sobre su independencia, se asume un modelo desuperpoblación, en el que los pares (xhik,yhik)(x\_{hik}, y\_{hik}) corresponden a observaciones de vectores aleatorios (X,Y)(X,Y) con distribución conjunta definida como:

Prc=Pr(X=r;Y=c),r=1,…,R; c=1,…,C(9−24)P\_{rc} = Pr(X=r ; Y=c), \quad r=1,\dots,R;\, c=1,\dots,C \quad (9-24)

y que cumple ∑r∑cPrc=1.\sum\_{r}\sum\_{c} P\_{rc}=1.

**5.3 Tablas cruzadas**

**97**. Si en lugar de una muestra se dispusiera de un censo que recogiera la información de xx y yy para toda la población, sería posible calcular el número de unidades en cada celda (r,c)(r,c) mediante:

Nrc=∑h=1H∑i∈U1h∑k∈UhiI(xhik=r;yhik=c)(9−25)N\_{rc} = \sum\_{h=1}^{H} \sum\_{i \in U\_{1h}} \sum\_{k \in U\_{hi}} I(x\_{hik}=r ; y\_{hik}=c) \quad (9-25)

y obtener las proporciones poblacionales prc=Nrc/N(++)p\_{rc} = N\_{rc}/N\_{(++)}, donde N(++)N\_{(++)} es el total de unidades en la población. Estas proporciones permiten aproximar las probabilidades PrcP\_{rc} bajo el modelo de superpoblación. En la práctica, como se trabaja con una muestra, estas proporciones se estiman a través de los estimadores ponderados explicados en secciones previas.

**5.3 Tablas cruzadas**

**98.** Las tablas cruzadas, o tablas de contingencia, constituyen una de las herramientas esenciales para el análisis de encuestas, ya que permiten organizar los datos de manera estructurada y mostrar cómo se distribuyen las frecuencias de dos o más variables categóricas. Gracias a este tipo de resumen, es posible identificar patrones y asociaciones que no serían evidentes a simple vista. Este procedimiento es ampliamente utilizado en la investigación y en el diseño de políticas públicas, pues facilita la exploración de relaciones entre variables. Por ejemplo, puede emplearse para analizar cómo varía la condición laboral en función del nivel educativo o cómo se distribuye el acceso a internet en zonas urbanas y rurales. Además, la literatura especializada suele referirse a estas como tablas de contingencia, y su interpretación puede reforzarse mediante gráficos de barras apiladas, los cuales permiten visualizar tendencias y diferencias de forma más intuitiva (véase la Sección 9.8 para un análisis más detallado).

**5.3 Tablas cruzadas**

**99.** Una tabla de contingencia puede definirse como una matriz de doble entrada en la que las filas, indexadas por r=1,…,Rr=1,\dots,R, y las columnas, indexadas por c=1,…,Cc=1,\dots,C, corresponden a las categorías de dos variables. Cada celda de la tabla contiene la frecuencia o proporción de casos que presentan simultáneamente la combinación (r,c)(r,c). De manera opcional, la tabla puede incluir totales marginales, que resumen los datos por fila o columna, y un total general que representa el tamaño poblacional total. Por ejemplo, las filas pueden representar niveles educativos y las columnas los estados de participación en el mercado laboral, permitiendo observar la distribución conjunta de ambas variables.

**5.3 Tablas cruzadas**

**100.** En el análisis de encuestas de hogares, las frecuencias incluidas en las tablas de contingencia no se obtienen mediante simples conteos de casos, sino a partir de frecuencias ponderadas, calculadas con los pesos muestrales de la encuesta. Cada celda de la tabla representa, de esta forma, el número estimado de personas en la población que pertenecen a la combinación específica de categorías.

En la mayoría de las encuestas de hogares, el resultado estándar es una tabla de doble entrada con las frecuencias ponderadas, que funcionan como estimaciones de las frecuencias poblacionales. Este tipo de resumen se presenta en la Figura x Tabla de doble entrada con frecuencias ponderadas estimadas, donde se ilustran las celdas, los totales marginales y el total general de la población:

**Figura x.** Tabla de doble entrada con frecuencias ponderadas estimadas

| **y** | **1** | **…** | **C** | **Total fila (N^(r+)\hat{N}\_{(r+)})** |
| --- | --- | --- | --- | --- |
| **x = 1** | N^11\hat{N}\_{11} | … | N^1C\hat{N}\_{1C} | N^(1+)\hat{N}\_{(1+)} |
| **…** | … | N^rc\hat{N}\_{rc} | … | … |
| **x = R** | N^R1\hat{N}\_{R1} | … | N^RC\hat{N}\_{RC} | N^(R+)\hat{N}\_{(R+)} |
| **Total col.** (N^(+c)\hat{N}\_{(+c)}) | N^(+1)\hat{N}\_{(+1)} | … | N^(+C)\hat{N}\_{(+C)} | N^(++)\hat{N}\_{(++)} |

La frecuencia estimada en cada celda (r,c)(r,c) se calcula mediante:

N^rc=∑h=1H∑i∈s1h∑k∈shiwhikI(xhik=r;yhik=c)(9−26)\hat{N}\_{rc} = \sum\_{h=1}^{H} \sum\_{i \in s\_{1h}} \sum\_{k \in s\_{hi}} w\_{hik} I(x\_{hik}=r; y\_{hik}=c) \quad (9-26)

y los totales marginales se obtienen como:

N^(r+)=∑cN^rc,N^(+c)=∑rN^rc,N^(++)=∑r∑cN^rc.\hat{N}\_{(r+)} = \sum\_{c} \hat{N}\_{rc}, \qquad \hat{N}\_{(+c)} = \sum\_{r} \hat{N}\_{rc}, \qquad \hat{N}\_{(++)} = \sum\_{r} \sum\_{c} \hat{N}\_{rc}.