CADA HECHO DE TU VIDA Cuenta

Un estudio sobre la capacidad inferencial de los estimadores calculados a partir de encuestas

Diciembre, 2018

INEC Buenas cifras, mejores vidas

Introducción

La Estadística es una herramienta sustancial al momento de planificar y establecer nuevos lineamientos dentro de la política pública. Por tal motivo, tener datos estadísticos confiables y oportunos se ha vuelto una necesidad para cualquier estado de derecho.

En Ecuador, el Instituto Nacional de Estadística y Censos (INEC), en su papel de ente rector de la Estadística Nacional, es el encargado de producir y avalar las operaciones estadísticas más relevantes de los sectores más sensibles e importantes del espectro nacional[1]. En este marco, el INEC ha coordinado y ejecutado censos de población y vivienda periódicos, encuestas sociodemográficas y económicas continuas de carácter coyuntural, medición de índices de precios, entre otros.

Uno de los métodos de recolección de datos más utilizado para satisfacer la necesidad de información es una encuesta por muestreo, es decir, una investigación aplicada a un subconjunto de la población de interés. Una encuesta por muestreo cuesta menos dinero, toma menos tiempo e incluso puede ser más precisa que una enumeración completa (Särndal, 2003, p. 3).

El presente documento desarrolla los conceptos básicos involucrados al momento de analizar estimadores provenientes de encuestas por muestreo. En la primera sección se introducen conceptos muestrales básicos. En la segunda sección se discute acerca de los errores muestrales y no muestrales. En la tercera sección se introduce el concepto de inferencia estadística. Por último, se presentan los diferentes parámetros a considerar al momento de desagregar un indicador.

Valores poblacionales y estadísticos

Las investigaciones estadísticas tienen como objetivo estimar valores poblacionales específicos. Un *valor poblacional* es una expresión numérica que resume el valor de una o más características para todos los N elementos de una población (Kish, 1965, p. 9). Algunos ejemplos de valores poblacionales de interés son:

- Total de hogares en un determinado instante de tiempo,
- Ganancia media de las empresas,
- Porcentaje de personas en edad de trabajar,
- Rendimiento $[\frac{kg}{ha}]$ de un determinado cultivo, entre otros.

Para Kish (1965), cualquier valor poblacional está determinado por los siguientes factores:

a. La población objetivo definida, por ejemplo:
   - Todas los hogares de una ciudad.
   - Empresas con actividad económica Manufactura.
   - Todas las personas de un determinado país.
   - Todas las haciendas en una determinada región.

b. La naturaleza de las variables a investigar y sus distribuciones. Algunos ejemplos son:
   - Los patrones de gasto de los hogares (dólares).
   - Las características de producción y ganancia media de las empresas de manufactura (variable categórica).
   - La participación en la fuerza laboral de las personas mayores a 15 años (variable dicotómica).

[1] Para más información acerca del Instituto Naciona de Estadística y Censos visite la siguiente página web.

- La producción de algún cereal en las haciendas (miles de kilogramos).

c. El método de observación, entre los que están:

- Entrevista cara a cara.
- Entrevista vía telefónica.
- Los registros de afiliamiento a la seguridad social.
- Medición objetiva de la producción usando equipo especializado.

d. La expresión matemática para derivar el valor poblacional desde los valores individuales de cada elemento.

Tanto el *valor poblacional* como el *valor verdadero* hacen referencia a las expresiones numéricas derivadas a partir de toda la población. La diferencia entre ellas surge de los errores de observación. El valor verdadero podría ser obtenido a partir de todos los elementos de la población, si las observaciones no fueran sujetas a errores (Kish, 1965, p. 9).

El *valor muestral*, o *estadístico*, es un *estimador* calculado a partir de los n elementos de la muestra. El estadístico es una *variable aleatoria*, que depende del diseño muestral[2] y de la combinación particular de elementos que fueron seleccionados. Por lo tanto, el estimador particular es uno entre todas las posibles estimaciones que podrían haber sido obtenidas por el mismo diseño muestral (Kish, 1965, p. 4).

Errores muestrales y no muestrales

Un estadístico es susceptible a diferentes fuentes de error. Särndal (2003) distingue cinco etapas en una encuesta, que van desde la planificación hasta la publicación de resultados, y sus fuentes de error asociadas, como se presenta a continuación.

1.- *Selección de la muestra*.

Esta etapa consiste en la ejecución del diseño muestral preconcevido. En esta, se calcula y selecciona un tamaño de muestra adecuado desde un marco existente o construido específicamente para la encuesta. Los errores asociados a esta etapa son:

- Errores de marco.- Dentro de las características que debe cumplir un marco muestral se encuentran:
  * Completitud: dentro del marco deben constar el total de elementos muestrales existentes en la población. De no ser así, existiría un error por omisión.
  * No duplicidad: cada registro que identifica un elemento muestral debe ser único, de otra manera se afectaría las probabilidades de selección de los mismos.
- Errores de muestreo: es el error causado por observar una muestra en vez del total de la población.

2.- *Recolección de datos*.

Esta etapa engloba el plan de trabajo y la ejecución de un modo específico de recolección (entrevista personal, entrevista telefónica, entrevista vía correo electrónico, entre otros). Los errores asociados a este etapa son:

- Errores de medida.- cuando el entrevistado da respuestas incorrectas; el entrevistador interpreta mal o influencia la respuesta del entrevistado; las preguntas del cuestionario no están claramente formuladas.

[2] El diseño muestral tiene dos aspectos: un *proceso de selección*, las reglas y operaciones por las cuales algunos miembros de la población son incluidos en la muestra; y un *proceso de estimación*, para calcular los estadísticos muestrales (Kish, 1965, p. 9-10).

- Errores debido a la no respuesta.- información no recolectada a nivel de individuo o variable. Por ejemplo, rechazo total o parcial.

3.- *Procesamiento de datos*.

En esta etapa se preparan los datos recolectados para su estimación y análisis. Los procesos involucrados son: codificación, digitación, edición, validación e imputación. Los errores asociados a este etapa son:

- Errores de codificación.
- Errores de transcripción.
- Errores introducidos o no corregidos en la edición.
- Errores en los valores imputados.

4.- *Estimación y análisis*.

En etapa implica el cálculo de los estadísticos de interés y sus respectivas medidas de precisión (estimador de la varianza, coeficiente de variación e intervalo de confianza). Otros análisis estadísticos pueden ser realizados, tales como comparación de subgrupos, análisis de correlación y regresión, etc. Los errores presentes en las etapas 1 a 3 afectan los estadísticos y, en el mejor de los casos, deberían ser considerados en el cálculo y estimación de sus medidas de precisión.

5.- *Diseminación de resultados y evaluación postencuesta*

Esta etapa engloba la publicación de los resultados de la encuesta, incluyendo una declaración general de las condiciones alrededor de la encuesta.

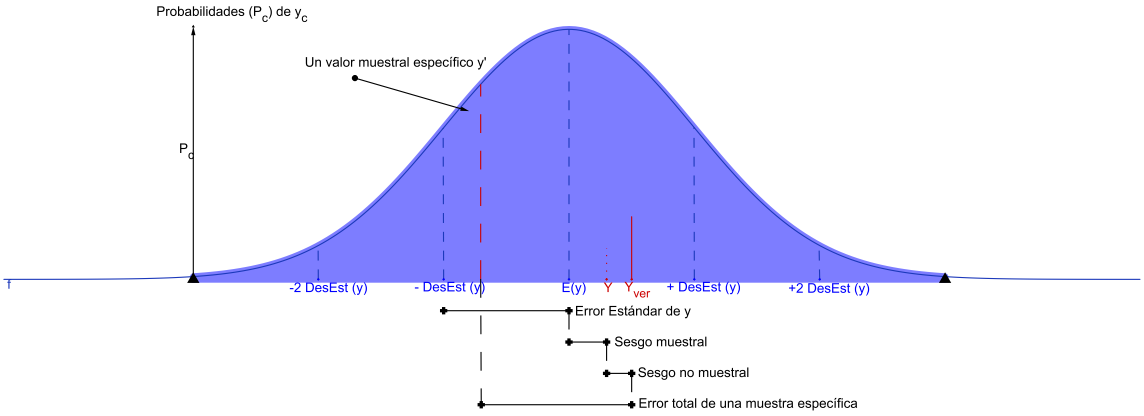
Los errores en los estadísticos están divididos tradicionalmente en dos grandes categorías: error muestral y error no muestral. El *error muestral* es, como se mencionó, el error causado por observar una muestra en vez del total de la población. El error muestral está sujeto a la variación intra-muestras. Los *errores no muestrales* incluyen todos los otros errores (Särndal, 2003, p. 16)[3].

Distribución muestral y error estándar

"Si pudiéramos extraer muestras repetidas del mismo tamaño de la misma población, y dividir las medias obtenidas sobre cierta variable, la distribución de estas medias se asemejaría a la curva normal en forma de campana"(Lininger, 1985, p. 110). Esta curva en forma de campana se conoce como *distribución muestral*. Cada estimador calculado a partir de una muestra tiene asociada una distribución muestral. La (fig.**1**) muestra una vista esquemática de la distribución muestral para un estadístico y cómo se integran y relacionan los siguientes conceptos:

- El *sesgo no muestral* $(Y_{VER} - Y)$ es la diferencia entre el valor verdadero Y_{VER} y el valor poblacional Y. Es generalmente desconocido y se debe a errores no muestrales.
- El *sesgo muestral* $(E(\widehat{y}) - Y)$ es la diferencia entre el valor esperado del estadístico $E(\widehat{y})$ y el valor poblacional Y. Se explica debido al error muestral.
- El *error total de una muestra específica* $(Y_{VER} - \widehat{y})$ puede descomponerse en error muestral y error no muestral, puesto que, el valor de un estadístico particular $\widehat{y}$ depende del conjunto de elementos seleccionados en la muestra y de los errores no muestrales presentes en las etapas de la encuesta.
- El *error estándar* $(ee(\widehat{y}))$ es el estimador puntual de la desviación estándar de la distribución muestral del estimador $\widehat{y}$. Como veremos ea continuación, la inferencia estadística se basa en los errores estándar.

Figura 1: Vista esquemática de una distribución muestral.



Fuente: Adaptado de (Kish, 1965, p. 12).

Särndal (2003) menciona que en general, se considera que un estimador ideal es aquel cuya distribución muestral está fuertemente concentrada alrededor del valor del parámetro desconocido. Esto garantiza una alta probabilidad de una estimación cercana. Sea $\widehat{y}$ un estimador de Y con varianza $V(\widehat{y})$ y sesgo $B(\widehat{y}) = E(\widehat{y}) - Y$. Una medida habitual de la exactitud de $\widehat{y}$ es el *Error Cuadrático Medio* (ECM):

$$\begin{aligned} ECM(\widehat{y}) = E[(\widehat{y} - Y)^2] &= V(\widehat{y}) + [B(\widehat{y})]^2 \\ &= [ee(\widehat{y})]^2 + [B(\widehat{y})]^2. \end{aligned} \tag{1}$$

El ECM depende tanto de la varianza como del sesgo. Si el ECM fuera la única preocupación, consideraríamos cómo el sesgo y la varianza cooperan en la producción de una ECM pequeño. Por ejemplo, un sesgo distinto de cero, o tal vez considerable, podría ser compensado por una pequeña varianza. Pero, además de un ECM pequeño, también se requiere que el sesgo sea pequeño en relación al error estándar. Esto es importante para que la inferencia estadística en encuestas sea válida[4].

Intervalos de confianza

Kish (1965) plantea que la teoría de Muestreo se ha concentrado en el cálculo de los errores estándar de los estimadores muestrales puesto que la inferencia estadística se basa en los errores estándar. Típicamente, la inferencia estadística toma la forma del intervalo

$$[\widehat{y} \; - \; t_p ee(\widehat{y}), \;\; \widehat{y} \; + \; t_p ee(\widehat{y})],$$

conocido como *intervalo de confianza*. De manera sencilla, la inferencia estadística se resume al hecho de que el intervalo de confianza contiene al valor poblacional Y con una cierta probabilidad asociada P. La longitud del intervalo de confianza depende del error estándar y, a través de t_p, del nivel de probabilidad P, usualmente aproximado por una distribución Normal o $t-$Student.

[3]Para profundizar en el tema, se invita al lector a revisar la Parte IV de Särndal (2003).
[4]Para profundizar en este tema, se invita al lector a revisar el Capítulo 5 de Särndal (2003).

Además, Kish (1965) menciona que los objetivos de la investigación se presentan comúnmente en términos de la precisión, que es función de la varianza estimada por la encuesta (error estándar). Si sesgos importantes, especialmente los errores ajenos al muestreo, están presentes y son distinguibles, la exactitud, que es función del error total (ECM), es una mejor medida de los objetivos de la encuesta que solamente la precisión.

Inferencia Estadística en encuestas

Parámetros de Revisión Estadística

Considerando todo lo expuesto anteriormente, el analista de información debe estar en la capacidad de generar herramientas de decisión al momento de discutir la representatividad de un indicador generado a partir de encuestas. Para tal fin, se sugiere que todo indicador debe ser analizado bajo el siguiente esquema:

- Análisis descriptivo de los indicadores a los niveles de desagregación requeridos.

  El análisis descriptivo de los indicadores debe ser el primer paso a dar para identificar las características, falencias y limitaciones de los indicadores y posibles errores en su cálculo. El análisis descriptivo también nos permite identificar valores perdidos o atípicos, además de darnos una aproximación de la distribución del indicador.

  El análisis descriptivo de un indicador incluye:

  - Tablas de frecuencia.
  - Medidas de tendencia central.
    * Mediana.
    * Media.
  - Medidas de dispersión.
    * Varianza.
    * Desviación estándar.

- Número de observaciones efectivas en la muestra.

  Al momento de desagregar un indicador, el procedimiento general suele ser dividir la muestra en subclases o subpoblaciones disjuntas cuyos elementos poseen características de nuestro interés. Por ejemplo, segmentar la muestra entre hombres y mujeres, hogares urbanos y rurales, y así.

  Sin embargo, si el número de observaciones efectivas dentro de cada subclase es considerablemente pequeño, el indicador en cuestión puede no estar representando el fenómeno de manera adecuada.

  Además, si el diseño muestral considera la selección de conglomerados en su primera etapa, es necesario analizar la distribución de las subclases o subpoblaciones en estos. (Kish, 1965, p. 17) menciona que en estos casos, el análisis depende de una normalidad aproximada para los valores muestrales dentro de cada conglomerado y su presencia en los mismos.

  Para este caso se recomienda utilizar los valores de la $t-$Student para aproximar el nivel de probabilidad de los intervalos de confianza, donde el grado de libertad se calcula restando el número de estratos del número de conglomerados de primera etapa [5].

  Se recomienda analizar el número de observaciones efectivas en todos los niveles de desagregación, o subclases, propuestos en las fichas metodológicas de los indicadores deseados.

- Coeficiente de variación e intervalo de confianza.

  El coeficiente de variación es una medida relativa de la precisión de los indicadores de una encuesta. En general se presenta en porcentaje y facilita la comparación entre indicadores de diferente naturaleza, o el mismo indicador pero en diferentes

[5] Para profundizar en esta temática se sugiere revisar la sección (8.6D) de Kish (1965, p. 17).

subpoblaciones o diferentes momentos en el tiempo.

El usuario debe observar que el coeficiente no sea relativamente grande. Esta medida debe contrastarse con el error que fue utilizado al momento de calcular la muestra, ya que para variables relacionadas a la variable de diseño deberían mantenerse en un rango cercano.

Por otra lado, el intervalo se lo puede leer de acuerdo al nivel de confianza fijado en cada investigación. Por ejemplo, con un nivel de confianza del 95% se puede decir que 95 de cada 100 muestras se encuentran dentro del límite inferior y superior del intervalo de confianza calculado.

Se recomienda calcular el coeficiente de variación y el intervalo de confianza en todos los niveles de desagregación propuestos en las fichas metodológicas.

Referencias

Kish, L. (1965). *Survey sampling*. J. Wiley.

Lininger, C. y Warwick, D. (1985). *La encuesta por muestreo: Teoría y Práctica*. Continental.

Särndal, C.E. y Swensson, B. y. W. J. (2003). *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer New York.

CADA HECHO DE TU VIDA *Cuenta*

@ecuadorencifras

@InecEcuador

t.me/ecuadorencifras

INEC/Ecuador

INECEcuador

INEC Ecuador