Encuesta para la Medición del Impacto COVID-19 en la Educación (ECOVID-ED) 2020

Diseño muestral

INEGI

Instituto Nacional de Estadística y Geografía

Encuesta para la Medición del Impacto COVID-19 en la Educación

ECOVID-ED

Diseño muestral

INEGI

Índice Página

1. Objetivo de la encuesta

Medir el impacto que ha provocado la contingencia por coronavirus COVID-19 en la educación de niñas, niños, adolescentes y jóvenes de 3 a 29 años en México, a fin de conocer las consecuencias que ha traído consigo el cierre temporal de las instituciones educativas en el país para evitar los contagios, durante el ciclo pasado (2019-2020) y el actual ciclo escolar (2020-2021).

2. Población objetivo

La población objeto de estudio son las personas de 3 a 29 años, que cuenten con un teléfono y que residan dentro del territorio nacional en la fecha del levantamiento.

3. Cobertura geográfica

La encuesta está diseñada para dar resultados a nivel nacional.

4. Marco de la encuesta

El marco de muestreo de la ECOVID-ED se conformó a partir del Plan Nacional de Numeración, publicado por el Instituto Federal de Telecomunicaciones (IFT) y actualizado a noviembre de 2020 y de los cuales se generaron de manera aleatoria una muestra de números telefónicos en cada una de las 32 entidades federativas. El diseño estadístico es probabilístico por lo que los resultados se pueden generalizar a nivel nacional, unietápico porque la selección de los números se realizó en una etapa, se dice que es estratificado debido a que cada entidad federativa es un estrato de diseño.

4.1 Estratificación

El diseño de la muestra es estratificado debido a que cada una de las 32 entidades federativas conforman los estratos de diseño, al interior de cada entidad se ha se seleccionado una muestra aleatoria de números telefónicos en forma independiente.

5. Tamaño de la muestra

El tamaño de la muestra se calculó para una proporción mínima de 1%, la cual requiere el tamaño de muestra mayor, lo que garantiza que las estimaciones de proporciones mayores al 1% queden cubiertas con este tamaño. La expresión empleada para el cálculo es la siguiente:

$$n = \frac{z^2 q}{r^2 p(1 - tnr)}$$

Donde:

- n = tamaño de la muestra.
- p = estimación de la proporción de interés.
- q = 1-p.
- r = error relativo máximo aceptable.
- z = valor asentado en las tablas estadísticas de la distribución normal estándar que garantiza obtener las estimaciones con una confianza prefijada.
- tnr = tasa de no respuesta máxima esperada.

Considerando una confianza de 90%, un error relativo esperado de 2.9%, una proporción de 1%, y la tasa de no respuesta esperada de 85%, se determinó una muestra a nivel nacional de 37 475 números telefónicos, el cual se ajustó a 37 754.

6. Asignación de la muestra

La muestra se distribuyó en las 32 entidades, en 6 entidades la muestra fue menor de 1 000 números telefónicos, mientras que en las otras 26 entidades la muestra de números telefónicos fue mayor de 1000 en cada una de ellas.

7. Selección de la muestra

La selección de la muestra de números telefónicos para la ECOVID-ED, se realizó de manera independiente al interior de cada entidad federativa entre las ciudades autorrepresentadas y, el resto urbano y rural el procedimiento se realizó de manera aleatoria.

La probabilidad de seleccionar el i-ésimo número telefónico en la e-ésima entidad, está dada por la siguiente expresión:

$$p_{ei} = \frac{k_e}{K_e}$$

Su factor de expansión[1] está dado por:

$$F_{ei} = \frac{K_e}{k_e}$$

Donde:

k_e = números telefónicos a seleccionar en la e-ésima entidad.

K_e = total de números telefónicos activos que conforman el marco muestral en la e-ésima entidad.

F_{ei} = factor de expansión del i-ésimo número telefónico de la e-ésima entidad.

8. Ajuste a los factores de expansión

Los factores de expansión elaborados conforme al procedimiento antes descrito se ajustaron en base a los siguientes conceptos:

8.1 Ajuste por no respuesta

El ajuste por no respuesta atribuida al informante se efectuó a nivel de entidad, mediante la siguiente expresión:

$$F^*_{ei} = F_{ei} \frac{\sum_e F_{ei} Q_e}{\sum_e F_{ei} Q^*_e}$$

1 El Factor de Expansión se define como el inverso de la probabilidad de selección. En la Norma Técnica del Proceso de Producción de Información Estadística y Geográfica para el Instituto Nacional de Estadística y Geografía, el Factor de Expansión se nombra Ponderador.

Donde:

F^{*}_{ei} = factor de expansión corregido por no respuesta del i-ésimo número telefónico, de la e-ésima entidad.

F_{ei} = factor de expansión del i-ésimo número telefónico, de la e-ésima entidad.

Q_{e} = total de números telefónicos seleccionados en la e-ésima entidad.

Q^{*}_{e} = total de números telefónicos seleccionados con respuesta en la e-ésima entidad.

8.2 Ajuste por proyección de población

Los factores corregidos por no respuesta se ajustan por proyecciones de población generadas por CONAPO al mes de abril de 2020, el procedimiento es el siguiente:

$$F^{**}_{ei} = F^{*}_{ei} \frac{X_{e}}{\hat{Y}_{e}}$$

Donde:

X_{e} = proyección de la población de CONAPO para la e-ésima entidad.

$\hat{Y}_{e}$ = estimación de la población obtenida de la ECOVID-ED en la e-ésima entidad.

F^{*}_{ei} = factor de expansión ajustado por no respuesta ,de la i-ésima vivienda con teléfono, de la e-ésima entidad.

F^{**}_{ei} = factor de expansión ajustado por proyección de población, de la i-ésima vivienda con teléfono, de la e-ésima entidad.

8.3 Ajuste por la cobertura de población con teléfono

Los factores de expansión ajustados por la proyección de población se corrigen por la cobertura de población que cuenta con teléfono por entidad federativa. Con información obtenida de ENDUTIH 2018 la cobertura a nivel nacional es aproximadamente de 94% para las viviendas con teléfono, el procedimiento es el siguiente:

$$F^{*}_{ei} = F_{ei} \frac{\hat{Z}_{e}}{\hat{Y}_{e}}$$

Donde:

$\hat{Z}_{e}$ = estimación de la población que reside en viviendas con teléfono obtenida de la ENDUTIH 2018 en la e-ésima entidad.

$\hat{Y}_{e}$ = estimación de la población que reside en las viviendas con teléfono obtenida de la ECOVID-ED en la e-ésima entidad.

F^{*}_{ei} = factor de expansión ajustado a la población que reside en viviendas con teléfono reportada por ENDUTIH 2018 para la i-ésima vivienda con teléfono, de la e-ésima entidad.

F_{ei} = factor de expansión ajustado por proyección de población de la i-ésima vivienda con teléfono, de la e-ésima entidad.

8.4 Ajuste por calibración

Debido a que, la muestra telefónica de la ECOVID-ED presenta algunas diferencias respecto a las estructuras de la ENOE cuarto trimestre de 2020, se procedió a ajustar sus factores de expansión mediante técnicas de calibración, a efecto de reducirlas.

La calibración consiste en minimizar una función que impone condiciones para generar nuevos factores de expansión (w_k), partiendo de los factores originales (d_k), tomando como referencia la información muestral de p variables de interés la muestra $(X_j \quad j=1,...,p)$, así como fuentes auxiliares de información (cuarto trimestre de 2020 de ENOE) traducidas en totales $t_j \quad j=1,...,p$. El procedimiento es el siguiente:

$$\min \sum_{k=1}^{n} d_k \, F\left(\frac{w_k}{d_k}\right)$$

Sujeta a:

$$\sum_{k=1}^{n} w_k \, X_{kj} = t_j; \quad j=1,...,P$$

Para este ajuste se consideró la función distancia $F(z)=z\,ln\,(z)-z+1$ conocida como Raking Ratio propuesta por Deville y Sarndal[2].

En una primera calibración, se ajustaron la población total, las estructuras demográficas de la muestra por población de 3 a 29 años de edad y sexo de la población, tomando como referencia las estimaciones del cuarto trimestre de 2020 de la ENOE.

Posteriormente, en una segunda calibración, los factores de expansión de la ECOVID-ED se ajustaron de manera que se mantuvieran las estructuras siguientes de la ENOE:

- Población de 3 a 29 años de edad.
- Nivel de escolaridad.

Las variables de escolaridad utilizadas para calibrar los factores de expansión la ECOVID-ED fueron las siguientes:

- Ninguno
- La unión de Preescolar y Primaria
- La unión de Secundaria y Carrera técnica con secundaria terminada
- La unión de Preparatoria o Bachillerato y Carrera técnica con preparatoria terminada (profesional técnico)
- La unión de Licenciatura o profesional y Maestría
- Doctorado

Cabe mencionar que la calibración tiene el efecto de minimizar las posibles desviaciones de las estimaciones de las variables más relevantes de la encuesta, que surgen, en este caso, al considerar una muestra aleatoria. Por lo anterior, a

[2] Deville, Jean–Claude, Sarndal, Carl–Erik (1992). "Calibration Estimators in Survey Sampling". Journal of the American Statistical Association, Vol 87. Núm. 418, pp 376–382

medida que las variables sean desglosadas en otras sub categorías conceptuales, las estimaciones se deberán tomar con reserva por la posible presencia de sesgos, debido al volumen reducido de muestra.

9. Estimadores

Para la ECOVID-ED se pueden calcular diversos tipos de estimadores. A continuación, se presentan los siguientes:

- Estimador del total

  El estimador del total de la característica X y Y están dados por:

$$\hat{X} = \sum_{e=1}^{L} \sum_{i=1}^{m_e} F_{ei} X_{ei}$$

$$\hat{Y} = \sum_{e=1}^{L} \sum_{i=1}^{m_e} F_{ei} Y_{ei}$$

  Donde:

  F_{ei} = factor de expansión final de la i-ésima vivienda con teléfono, de la e-ésima entidad.

  X_{ei} = valor observado de la característica de interés X en la i-ésima vivienda con teléfono, en la e-ésima entidad.

  - Para la estimación de tasas se utiliza el estimador de razón:

$$\hat{R} = \frac{\hat{X}}{\hat{Y}}$$

    Donde, $\hat{Y}$ se define en forma análoga a $\hat{X}$.

- Estimador de una proporción

$$\hat{p}_e = \frac{\sum_{i=1}^{m_e} F_{ei} X_{ei}}{\sum_{e=1}^{m_e} F_{ei}}$$

$$\hat{q}_e = 1 - \hat{p}_e$$

$$\hat{p} = \frac{\sum_{e=1}^{L} \sum_{i=1}^{m_e} F_{ei} X_{ei}}{\sum_{e=1}^{L} \sum_{ie=1}^{m_e} F_{ei}}$$

10. Estimación de errores de muestreo

Para la evaluación de los errores de muestreo de las principales estimaciones nacionales se emplearon las expresiones de estimación de varianza para los estimadores de razón, estimadores de totales y para proporciones empleando las siguientes expresiones:

- Totales

$$\hat{V}\left(\hat{X}\right)=\sum_{e=1}^{L}\hat{M}_e\sum_{i=1}^{m_e}\frac{F_{ei}\left(X_{ei}-\overline{X}_e\right)^2}{m_e-1}$$

$$\hat{V}\left(\hat{Y}\right)=\sum_{e=1}^{L}\hat{M}_e\sum_{i=1}^{m_e}\frac{F_{ei}\left(Y_{ei}-\overline{Y}_e\right)^2}{m_e-1}$$

Donde:

$$\overline{X}_e=\frac{\sum_{i=1}^{m_e}F_{ei}X_{ei}}{\sum_{i=1}^{m_e}F_{ei}}$$

$$\overline{Y}_e=\frac{\sum_{i=1}^{m_e}F_{ei}Y_{ei}}{\sum_{i=1}^{m_e}F_{ei}}$$

$$\hat{M}_e=\sum_{i=1}^{m_e}F_{ei}X_{ei}$$

$$m_e=\sum_{i=1}^{m_e}1$$

- Proporciones

$$\hat{V}\left(\hat{p}\right)=\sum_{e=1}^{L}\frac{\hat{p}_e\,\hat{q}_e}{m_e-1}$$

- Razones

$$\hat{V}\left(\hat{R}\right)=\frac{\hat{V}\left(\hat{X}\right)+\hat{R}^2\hat{V}\left(\hat{Y}\right)-2\hat{R}\hat{COV}\left(\hat{X},\hat{Y}\right)}{\hat{Y}^2}$$

$$COV\left(\hat{X},\hat{Y}\right)=\sum_{e=1}^{L}\hat{M}_e\frac{\sum_{i=1}^{m_e}\left(x_{ei}-\overline{X}_e\right)\left(y_{ei}-\overline{Y}_e\right)}{m_e-1}$$

Donde:

x_{ei} = la variable de estudio X de la i-ésima viviendas, en la e-ésima entidad.

$\overline{X}_e$ = promedio de la variable de estudio X en la e-ésima entidad.

m_e = número de personas en la e-ésima entidad.
L = número de estratos a nivel nacional.
$\hat{Y}^2$ = el cuadrado del total ponderado de la característica Y.

Estas definiciones de X son análogas para la variable de estudio Y.

Las estimaciones de la desviación estándar (DE) y coeficiente de variación (CV) se calculan mediante las siguientes expresiones:

$$DE = \sqrt{\hat{V}\left(\hat{\theta}\right)} \qquad CV = \frac{\sqrt{\hat{V}\left(\hat{\theta}\right)}}{\hat{\theta}}$$

Donde:

$\hat{\theta}$ = estimador del parámetro poblacional θ.
$\hat{V}\left(\hat{\theta}\right)$ = estimador de la varianza bajo muestreo aleatorio simple estratificado.

Finalmente, el intervalo de confianza $I_{1-\alpha}$ al 100(1-α)%, se calcula de la siguiente forma:

$$I_{1-\alpha} = \left(\hat{\theta} - z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\theta}\right)}, \hat{\theta} + z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\theta}\right)}\right)$$

Donde α es el nivel de significancia.

11. Homologación de la Semaforización para los Umbrales de Indicadores de precisión estadística[3]

Para facilitar la interpretación de las precisiones estadísticas de la información pública en tabulados, el Comité de Aseguramiento de la Calidad, en la cuarta sesión celebrada el 1 de noviembre de 2018, aprobaron los siguientes umbrales y especificaciones para la publicación en los tabulados los CV, así como su semaforización de estos.

Umbrales aprobados para la cobertura de la CV

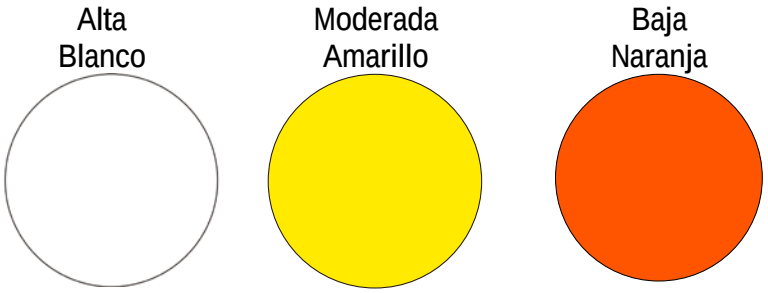
Interpretación	Semaforización	Viviendas/Hogares/Otras unidades diferentes a las económicas DGES/DGEGSPJ
Alta	Blanco	[0%, 15%)
Moderada	Amarillo	[15%, 30%)
Baja	Naranja oscuro	>=30%

Umbrales aprobados para el reporte de la precisión de acuerdo con el coeficiente de variación en los tabulados de resultados de los proyectos con muestreo probabilístico (acuerdo CAC-007/01/2018).

[3] La fuente de esta información está basada en el documento del Comité de Aseguramiento de la Calidad depositado en el siguiente sitio http://intranet.inegi.org.mx/calidad/wp-content/uploads/2017/02/Homologacion_de_umbrales.pdf

A partir del segundo trimestre de 2018, se publican los siguientes indicadores de precisión estadística en la presentación de resultados en tabulados de todas las encuestas con muestreo probabilístico del INEGI: error estándar, intervalo de confianza y coeficiente de variación (CV). Adicionalmente, se estandariza la coloración en los tabulados para indicar el nivel de precisión de las estimaciones con base en el CV. A continuación, se presenta el código RGB de los colores utilizados en la semaforización:

Parámetros RGB para la semaforización del coeficiente de variación.



Alta	Moderada	Baja
Rojo: 255 Verde: 255 Azul: 255	Rojo: 255 Verde: 234 Azul: 0	Rojo: 255 Verde: 84 Azul: 0

El siguiente texto explicativo aparece en cada uno de los tabulados publicados de encuestas por muestreo probabilístico.

Las estimaciones que aparecen en este cuadro están coloreadas de acuerdo con su nivel de precisión, en *Alta, Moderada* y *Baja*, tomando como referencia el coeficiente de variación CV (%). Una precisión *Baja* requiere un uso cauteloso de la estimación en el que se analicen las causas de la alta variabilidad y se consideren otros indicadores de precisión y confiabilidad, como el intervalo de confianza.

Nivel de precisión de las estimaciones:

Alta, CV en el rango de (0,15)

Moderada, CV en el rango de [15, 30)

Baja, CV de 30% en adelante