Ignorabilidad: Un supuesto clave en la dinámica de la inferencia estadística en Ciencias de la Salud

EDUARDO ALARCÓN-BUSTAMANTE*
FACULTAD DE MATEMÁTICAS, DEPARTAMENTO DE ESTADÍSTICA, PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
NÚCLEO MILENIO SOBRE MOVILIDAD INTERGENERACIONAL: DEL MODELAMIENTO A LA POLÍTICA (MOVI) [NCS2021072]
LABORATORIO INTERDISCIPLINARIO DE ESTADÍSTICA SOCIAL, LIES, SANTIAGO, CHILE

Me imagino que más de una vez escuchaste que la *positividad nacional*, detectada por test, del COVID-19 era de un porcentaje específico. Es más, muchas veces te asustaste por una positividad alta o te relajaste por una positividad baja. Por ejemplo, el 25 de enero de 2022, y basado en el resultado de 80.163 tests, el Ministerio de Salud reportó que

> *La positividad para las últimas 24 horas a nivel país es de 15,82 % y en la Región Metropolitana es de 13,59 %.*

Entonces, ¿por qué se decía nacional si sólo una parte de Chile se tomó el test? ¿Existe algún procedimiento que nos permita determinar que a partir de lo observado en los resultados de los test se pueda hablar de una positividad nacional?

El teorema de probabilidades totales [1] nos puede ayudar a responder esta pregunta. Nuestro interés recae en aprender sobre la positividad nacional. Sea $\mathcal{M}$ el espacio muestral definido por los chilenos y chilenas. En este espacio muestral se define para cada $m \in \mathcal{M}$ las siguientes variables aleatorias. Denotemos por $T(m)$ una variable aleatoria binaria, tal que $T(m) = 1$ si el individuo m se tomó el test y 0 si no. Sea $R(m) = 1$ si el individuo m es positivo por COVID-19, de acuerdo al resultado del test. Queremos saber la proporción de individuos positivos a nivel nacional. En otras palabras, nos interesa $\mathrm{P}(R = 1)$, la cual se puede descomponer como

$$\begin{aligned}\mathrm{P}(R = 1) =& \mathrm{P}(R = 1|T = 1)\mathrm{P}(T = 1) \\ &+ \mathrm{P}(R = 1|T = 0)\mathrm{P}(T = 0)\end{aligned}$$

Es decir, la positividad a nivel nacional, $\mathrm{P}(R = 1)$, es el promedio ponderado entre la proporción de positivos que se tomó el test[1], $\mathrm{P}(R = 1|T = 1)$, y la de quienes no se lo tomaron, $\mathrm{P}(R = 1|T = 0)$.

Esta última cantidad es imposible de ser observada, pues nunca sabremos la proporción de positivos entre quienes no se tomaron el test. Entonces ¿cómo podemos hablar de positividad nacional utilizando sólo la información de quienes se tomaron el test? Técnicamente, ¿qué permite afirmar que $\mathrm{P}(R = 1) = \mathrm{P}(R = 1|T = 1)$? La única forma de que ocurra esto es haciendo lo siguiente:

$$\mathrm{P}(R = 1|T = 1) = \mathrm{P}(R = 1|T = 0).$$

Es decir, debemos asumir que la proporción de infectados por COVID-19 entre quienes se tomaron el test es idéntica a la proporción de contagiados entre los que no lo hicieron. Este supuesto es denominado *ignorabilidad*[2]. Formalmente, estamos diciendo que la positividad es constante entre quienes se toman o no el test, i.e., $R \perp\!\!\!\perp T$. Por lo tanto, estamos asumiendo que no importa si la persona m se tomó el test y, en consecuencia, la positividad nacional se puede determinar sólo con aquellos que sí se lo tomaron.

A pesar de lo fuerte y cuestionable del supuesto de ignorabilidad, es una herramienta útil para aprender sobre un parámetro a partir de sólo una parte de la población. Como nunca observaremos la proporción de contagiados entre quienes no se tomaron el test, entonces nunca sabremos si este supuesto se cumple o no. Es decir, el supuesto de ignorabilidad es irrefutable [5]. El supuesto de ignorabilidad, a grandes rasgos, afirma que lo que no podemos conocer tiene el mismo comportamiento que lo que sí conocemos.

En este artículo pretendo mostrar, brevemente con otro ejemplo, que la ignorabilidad es clave para hacer inferencias. El ejemplo que utilizo es en base a los resultados obtenidos desde la Encuesta Nacional de Salud (ENS) y los factores de expansión que se utilizan en ella. En particular, muestro dónde está presente la ignorabilidad y cómo podemos detectarla.

Ignorabilidad en la ENS 2016-2017

> *La muestra ENS 2016-2017 es representativa de la población nacional, regional, urbana y rural, además de probabilística, se puede definir como estratificada geográficamente, de conglomerados y multietápica, con*

*esalarcon@uc.cl

[1] Para efectos de este ejemplo, consideraremos que hablar de una persona con resultado positivo es equivalente a decir que esta persona está contagiada.

[2] Para detalles técnicos sobre este supuesto, ver [2, 3]. Para una discusión sobre el uso de la ignorabilidad en otros ámbitos, ver [4].

[3] epi.minsal.cl/wp-content/uploads/2018/05/DISE%C3%91O-MUESTRAL-ENS-2016-2017.pdf

distribución no proporcional por estrato de muestreo.[3]

Lo anterior es declarado en el informe metodológico de la ENS 2016-2017. Pero, ¿qué significa una muestra representativa? En términos coloquiales esto significa que la muestra es capaz de representar el comportamiento poblacional. Para poder hacer esta afirmación debemos considerar el concepto de ignorabilidad. Es más, el diseño muestral tras las muestras representativas es denominado *diseño de muestreo ignorable* (para detalles técnicos puede ver [6, 7]). En este tipo de diseños se asume que al conocer características de los individuos de la población de interés, tales como sexo biológico, nivel socioeconómico, región, manzana, entre otras, entonces ser seleccionado o no en la muestra es irrelevante para hacer inferencias sobre la variable de estudio [4].

Supongamos que queremos conocer la proporción de mujeres que consume tabaco en el país. Podemos definir sobre $\mathcal{M}$ las siguientes variables aleatorias para todo $m \in \mathcal{M}$: sea $C(m) = 1$ si el individuo m consume tabaco y 0 si no; $S(m) = 1$ si la persona m fue seleccionada en la muestra y 0 si no; $G(m) = 1$ si la persona m es mujer y 0 si no. Entonces, nuestro interés es la proporción de consumidores de tabaco entre las mujeres, $\mathrm{P}(C = 1|G = 1)$. Por el teorema de probabilidades totales, sabemos que esta proporción la podemos descomponer como el promedio ponderado de las consumidoras de tabaco que fueron seleccionadas en la muestra y las que no, i.e.,

$$\begin{aligned}\mathrm{P}(C = 1|G = 1) =& \mathrm{P}(C = 1|G = 1, S = 1)\mathrm{P}(S = 1|G = 1) \\ &+ \mathrm{P}(C = 1|G = 1, S = 0)\mathrm{P}(S = 0|G = 1)\end{aligned}$$

Acá, $\mathrm{P}(S = 1|G = 1)$ es la proporción de seleccionados en la muestra, entre las mujeres del país y $\mathrm{P}(C = 1|G = 1, S = 1)$ es la proporción de consumidoras de tabaco entre las mujeres seleccionadas en la muestra. Análogamente, $\mathrm{P}(C = 1|G = 1, S = 0)$ es la proporción de consumidoras de tabaco entre las no seleccionadas en la muestra. Esta última cantidad es imposible de ser estimada. Sin embargo, el supuesto de ignorabilidad nos permite hacer inferencias sobre $\mathrm{P}(C = 1|G = 1)$, considerando solo la información observada. El diseño ignorable nos dice que $C \perp\!\!\!\perp S|\{G = 1\}$. Es decir, por el hecho de ser mujer, ser seleccionado o no en la muestra es irrelevante para aprender sobre la proporción de mujeres fumadoras en Chile. La traducción de esta independencia condicional, es que la proporción de consumidoras de tabaco entre las seleccionadas es idéntica a la proporción de consumidoras de tabaco entre las no seleccionadas, i.e.,

$$\mathrm{P}(C = 1|G = 1, S = 1) = \mathrm{P}(C = 1|G = 1, S = 0).$$

Así, para estimar la proporción de consumidoras de tabaco a nivel nacional, basta con conocer la proporción de consumidoras de tabaco entre las seleccionadas en la muestra, ya que la ignorabilidad permite que

$$\mathrm{P}(C = 1|G = 1) = \mathrm{P}(C = 1|G = 1, S = 1)$$

Así, decir que una muestra es representativa (que la muestra representa a la población) es análogo a decir que creemos en que el comportamiento de los no seleccionados es idéntico al de los seleccionados en la muestra.

Ignorabilidad en los factores de expansión

Bien sabemos que para extrapolar los resultados de la encuesta hacia la población, se utilizan los denominados factores de expansión. Pero ¿cuál es la filosofía que hay detrás de estos factores que nos permiten hacer la extrapolación? Para exponer este punto, tomaré como referencia el documento de trabajo "Fundamentos de la nueva metodología de calibración de los factores de expansión de la Encuesta Nacional de Empleo"[4].

Los factores de expansión provienen de un estimador conocido como Estimador de Horvitz-Thompson, el cual es útil para estimar el total poblacional de una característica de interés. Este estimador tiene como base el Principio de Representatividad, el cual, como se manifiesta en el documento, significa que *cada elemento incluido en una muestra se representa a sí mismo y a un grupo de elementos que no pertenecen a la muestra seleccionada, cuyas características son cercanas a las del elemento incluido en la muestra* [8].

Si miramos con detención, el principio de representatividad no es nada más ni nada menos que el supuesto de ignorabilidad aplicado a la estimación del factor de expansión, pues estamos asumiendo que el comportamiento de la variable respuesta es homogéneo entre individuos con las mismas características.

Entonces, si estamos dispuestos a creer en que personas con las mismas características tendrán la misma respuesta, entonces no tendremos problemas en creer las conclusiones de la ENS (o cualquier encuesta).

Discusión

A través del artículo he mostrado con algunos ejemplos que la ignorabilidad es un supuesto clave para hacer inferencias sobre un parámetro de interés. Aunque se presentan sólo dos ejemplos, no es difícil imaginar que la gran mayoría de los estudios, no sólo en Ciencias de las Salud, utilizan este supuesto para extrapolar los resultados a la población. Otro ejemplo, que no desarrollé, es el uso de la ignorabilidad en estudios de caso y control, ya que un mismo individuo no puede estar en ambos grupos y por lo tanto se supone que: *independiente del grupo en el que está, su respuesta hubiese sido la misma ya que ella solo dependerá de otras características (como la edad, sexo, entre otras)*.

[4] https://www.ine.gob.cl/docs/default-source/documentos-de-trabajo/documento-de-trabajo-fundamentos-de-la-nueva-calibraci%C3%B3n-de-los-factores-de-expansi%C3%B3n-en-la-ene.pdf?sfvrsn=3de3a0e1_4

La ignorabilidad nos ayuda a extraer conclusiones que son fáciles de entender, razonar y absorber. Sin embargo, no es el único camino. Para ver otra forma de analizar encuestas puede leer [9].

La ignorabilidad es un supuesto fuerte y por lo tanto las conclusiones pueden llegar a ser muy débiles. Este efecto se denomina *Ley de credibilidad decreciente* [10] y determina que entre más fuertes son los supuestos, más débiles son las conclusiones.

Es indispensable entender que los datos no hablan por sí solos, si no que son los supuestos los que nos ayudan a entender esta dinámica. Es más, como es manifestado en [10], la lógica de la inferencia se puede resumir así

$$\text{Datos} + \text{Supuestos} = \text{Conclusiones}$$

Es decir, para un conjunto fijo de datos si cambio los supuestos, mis conclusiones pueden cambiar rotundamente.

Financiamiento: Eduardo Alarcón-Bustamante ha sido parcialmente financiado por el Proyecto FONDECYT de Postdoctorado 3220422 y por ANID −− Programa Iniciativa Científica Milenio −− Código NCS2021072.

Referencias

[1] Kolmogorov AN. Foundations of the theory of probability. New York: Chelsea Pub. Co., 1950.

[2] Rosenmbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*; 1983; 70(1):41-55.

[3] Imbens G. The role of the propensity score in estimating dose-response functions. *Biometrika*; 2000; 87(3):706-710.

[4] Alarcón-Bustamante E. Ignorar o no ignorar, esa es la cuestión. *Cuadernos de Beauchef*; 2022; 6(1):15-33.

[5] Manski C. Identification for prediction and decision. New York: Harvard University Press, 2007.

[6] Scott A. Some comments on the problem of randomization in surveys. *Sankhyã*; 1977; 39:1-9.

[7] Sugden R, Smith T. Ignorable and informative designs in survey sampling inference. *Biometrika*; 1984; 71(3):495-506.

[8] Gutiérrez A. Estrategias de muestreo, diseño de encuestas y estimación de parámetros. Bogotá: Ediciones de la U, 2016.

[9] San Martín E, Alarcón-Bustamante E. Dissecting Chilean surveys: the case of missing outcomes. *Chilean journal of statistics*; 2022; 13(1):17-46.

[10] Manski C. Public Policy in an uncertainty world: analysis and decisions. New York: Harvard University Press, 2013.