Estimación de parámetros cuando la muestra es la población: reflexiones sobre los posibles caminos metodológicos

Cristian Flores Peñailillo[1*], Sandra Flores Alvarado[2], Sergio Alvarado Orellana[2]

[1]Estudiante de Magíster en Bioestadística, Escuela de Salud Pública
Facultad de Medicina, Universidad de Chile
[2]Programa de Bioestadística, Escuela de Salud Pública
Facultad de Medicina, Universidad de Chile

Resumen

La inferencia estadística se refiere a la obtención de conclusiones a partir de información que está sujeta a aleatoriedad. Esta información se obtiene a partir de muestras con las que se espera conocer lo que ocurre en la población. Ahora, si se cuenta con acceso a los datos poblacionales como en registros poblacionales ¿en qué queda el proceso inferencial? Cuando ocurre esto a la muestra que origina estos grandes sets de datos se le denomina población aparente. Tener acceso a este conjunto de datos no es sinónimo de que los datos sean reales ya que con estas grandes bases el investigador no tiene acceso al instrumento con que se originan los datos, por lo que no se sabe qué tan imperfecto es. Por otro lado, población puede ser considerada como una población verdadera o como una realización proveniente de una super-población. Muchos estudios realizan inferencia a partir de datos poblacionales sin la debida reflexión acerca de qué exactamente están infiriendo y, más aún, sin clarificar la super-población que quieren conocer. Ante esto una posible solución se deriva de considerar el paradigma bayesiano en estos análisis, en el que se actualice la información previa.

Palabras clave: estudios de muestreo, población, registros, inferencia estadística.

—¿Qué camino debo seguir?
—Según adónde quieras llegar— observó el Gato.
—Me es absolutamente igual un sitio que otro...— dijo Alicia.
—Entonces también da lo mismo un camino que otro— añadió el Gato.
—Es que con tal de llegar a alguna parte...— agregó Alicia a modo de explicación.

Alicia y el Gato de Cheshire

Lewis Carroll, *Alicia en el País de las Maravillas*.

1. Introducción

La inferencia estadística se refiere a la obtención de conclusiones a partir de información que está sujeta a aleatoriedad. Para esto, se deben considerar dos términos de forma conceptual y, a su vez, con un asidero razonable en el mundo real: la población y la muestra. La población es la colección completa de elementos o individuos en estudio cuyo tamaño de denota por N, y la muestra un subconjunto tomado de la población, cuyo tamaño de denota por n.

Considerando esto, el principio básico de la inferencia estadística es que las conclusiones acerca de una población de interés pueden ser obtenidas usando la información contenida en una muestra tomada de esa población [1]. Entre la muestra y las conclusiones hay un camino más o menos intrincado —si se quiere adjetivar de alguna forma— antes de dar el salto inferencial.

Ahora bien, son distintos pero ¿ambos términos se mantienen separados al llevarlos a la realidad de un problema en particular? Si consideramos estudios transversales, tomar una muestra en un momento dado permite hacer inferencias hacia la población que dio origen a la muestra en ese momento. Si se piensa en estudios longitudinales, seguir una muestra adecuada en el tiempo para estudiar variables permite conocer su comportamiento en la población. Sin embargo, ambos tipos de conclusiones descansan en dos supuestos fundamentales respecto a las muestras seleccionadas: a) que representan adecuadamente a la población de la que proceden, y b) que su selección fue probabilística. Si no se consideran estos supuestos, ¿qué se puede decir de las conclusiones a las que se arriben, por medio de la inferencia, en conjuntos de datos o muestras que no cumplan con la representatividad y selección probabilística?

Por otra parte, si se quieren descartar los supuestos para el trabajo con muestras y se pretende *usar la población para arribar a conclusiones basadas en la inferencia estadística*, ¿qué conclusiones puede dar la inferencia, si ya se cuenta con la información de toda la población? ¿Qué se puede hacer además de describir la población? Pero, más importante respecto a lo que se

*criflores@ug.uchile.cl

puede o no se puede hacer con los datos a mano, ¿qué es pertinente hacer? Es decir, empleando métodos inferenciales a datos poblacionales, ¿hacia qué, quién, cuándo o hacia dónde, se está infiriendo? Para este dilema se han planteado, no sin críticos, la existencia de súper-poblaciones.

El objetivo de este ensayo es contribuir a la discusión estadística respecto al empleo de métodos inferenciales sin la adecuada reflexión del porqué de su uso, más allá de que sea solicitado como forma de validar resultados. En ningún caso es una demonización al trabajo con datos poblacionales o una apología acérrima a la inferencia frecuentista. Por el contrario, busca ser una ayuda para orientar trabajos que empleen estas herramientas con la adecuada reflexión y evitar que los investigadores actúen como meros autómatas.

2. Población y muestra

Una población corresponde a la colección completa de unidades de observación que se quiere estudiar [2], es decir, acerca de quienes se quiere conocer algo y cuyo número de elementos se denota como N. Para identificar correctamente a la población se suele definir la población objetivo y la población muestreada. La definición de población objetivo es la dada al inicio de este párrafo, en tanto que la población muestreada [2] es aquella población de la cual una muestra será tomada.

Para entender el matiz entre ambos conceptos un ejemplo de la vida real puede ser iluminador. Suponga que una agencia quiere estudiar las preferencias académicas de los profesionales que cursan primer año de postgrado, este grupo es su población objetivo. Sin embargo, la población muestreada de la agencia puede definirse como los profesionales que han solicitado antecedentes para cursar programas en tres principales universidades.

Una vez que se tiene claridad en cuanto a la población a estudiar, si ésta no se puede identificar de manera exhaustiva o, como en el ejemplo anterior, no se puede llegar por temas logísticos a todos los miembros que conforman la población, se toma una muestra. La muestra corresponde a un subconjunto de una determinada población y cuyo número de elementos se denota con n. Idealmente se espera que la muestra tomada represente adecuadamente el grado de diversidad de la población de la cual procede, por lo que la selección de buenas muestras se hace imprescindible si se quiere conocer a la población. La Figura 1 resume gráficamente estas ideas.

3. Inferencia

Como se mencionó, ya sea poque es conceptualmente imposible acceder a una población (imagine que su población en estudio es, por ejemplo, las hojas del *Platanus orientalis*) o por temas logísticos, se utilizan muestras para conocer lo que sucede en la población. En el campo estadístico el proceso por el cual se llega a conclusiones generales (poblacionales) a partir de casos particulares (muestras) se conoce como inferencia estadística.

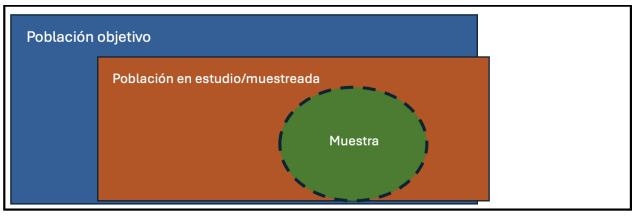

Figura 1: Representación de población, población en estudio y muestra. Adaptado de [2].

Formalmente se define como la obtención de conclusiones a partir de información que se encuentra sujeta a aleatoriedad, es decir, las conclusiones a las que se llegue acerca de la población de interés pueden ser hechas a partir de la información contenida en la muestra [1].

Si bien los conceptos de población, muestra e inferencia intuitivamente están relacionados, la realización de inferencias confiables utilizando la información de la muestra en gran medida depende de la relación establecida entre ambos conjuntos. Si llamamos a la población P y a su tamaño (número de elementos) como N, y a la muestra m y su tamaño como n, tenemos las siguientes relaciones:

$$m \subseteq P,$$
$$n \ll N.$$

Es decir, la muestra está contenida en la población y el número de elementos es considerablemente menor. Considerando esto, ¿se encuentra el camino libre para realizar inferencias? Casi. Retomando el ejemplo de la agencia, ¿sería lo mismo realizar inferencias tomando sólo a los postulantes a programas de la Facultad de Medicina que si se considerara al resto de facultades, para un mismo tamaño de muestra? Intuitivamente se puede decir que la inferencia no sería la misma.

En términos estadísticos, para realizar inferencias confiables se espera que las muestras sean representativas de la diversidad existente en la población de la cual proceden y que sean probabilísticas, es decir, que los miembros de la población presenten una probabilidad de ser incluidos en la muestra.

El que las muestras cumplan con estas características disminuye el sesgo al que puedan estar afectas las estimaciones realizadas. Se entiende como sesgo a un error o desviación sistemática en el muestreo, medición o en los procesos de estimación, lo que resulta en estadísticos considerablemente mayores o menores que la característica poblacional que se quiere estimar.

Al alero de estos elementos opera la inferencia frecuentista, en donde la probabilidad de un evento se interpreta considerando un gran número de repeticiones del experimento, de manera que se pueda llegar a conclusiones acerca del estado de una población. Debe hacerse notar que operan otros conceptos como distribución de probabilidad, teoría de muestreo, muestras grandes, convergencia, y otros conceptos de los cuales se asume que el lector tiene nociones.

4. Cuando la población es la muestra

Como se mencionó, el proceso de inferencia estadística —siguiendo un orden lógico— va desde la información observada en la muestra a conclusiones respecto a la población no observada. Suponiendo que la muestra es ideal, *i.e.*, cumple todos los supuestos, se puede tener una medida de probabilidad respecto de las conclusiones alcanzadas. Ahora, en cambio, si se cuenta con acceso a los datos poblacionales como en registros poblacionales ¿en qué queda el proceso inferencial? La inferencia frecuentista es clara, inferir respecto a los datos poblacionales es improcedente. La Figura 2 representa la disyuntiva.

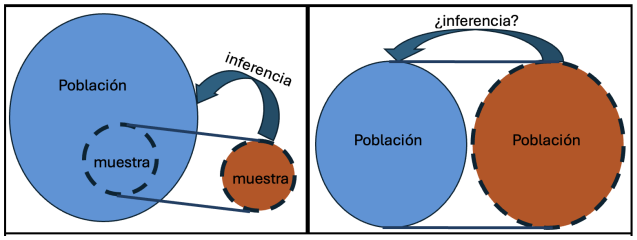

Figura 2: Representación del dilema de uso de datos poblacionales. Elaboración propia.

No pocos son los autores entregados a despejar dudas y dar respuestas acerca de este dilema. Thygesesn y Ersbøll [3] comentan este tema refiriéndose a los estudios basados en bases de datos poblacionales. Al respecto, su discusión está centrada más bien en los sesgos y la información como tal que pueda existir en las bases de datos poblaciones y como éstos afectan, o más bien condicionan, las investigaciones que se puedan realizar a partir de ellos. En otras palabras, respecto a las fallas del instrumento con que se realiza la recolección de datos. En la Tabla 1 se mencionan las fortalezas y debilidades.

Respecto a la inferencia, la controversia entre el uso de muestras y poblaciones queda referida a que, aunque el registro considere a toda la población, esto puede considerarse como una muestra de una gran población potencial teórica, en un tiempo y lugar, y por lo tanto, como la realización de un proceso estocástico. A esta población se le conoce como súper-población [3, 4]. La Figura 3 representa esta idea.

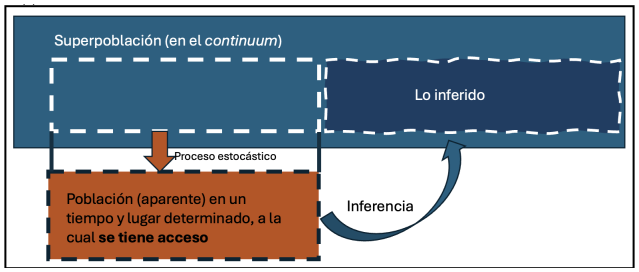

Figura 3: Representación de los conceptos de población (aparente) y super-población. Elaboración propia.

5. La súper-población

El constructo teórico de la súper-población no está exento de polémicas dado que puede ser contrario a la inferencia frecuentista clásica. Hartley y Sielklen [4], tomando de ejemplo las encuestas, plantean el punto de vista de súper-poblaciones para muestreos de población finita. Bajo este paradigma la población finita de interés de tamaño N a la cual se tiene acceso proviene de una población infinita y considera el procedimiento estocástico que genera la muestra de n unidades en el siguiente procedimiento de dos pasos: (1) tomar una muestra grande de tamaño N desde la súper-población infinita, y (2) tomar una muestra de $n < N$ de la gran muestra de tamaño N.

Tabla 1: Fortalezas y debilidades de datos basados en registros de la población [3].

Fortalezas	Debilidades
• Los datos ya existen, por lo que el investigador no debe recolectarlos. • Los datos contienen información de toda la población objetivo. • Los datos son recolectados independientemente de la pregunta de investigación. • Tiempo transcurrido entre las mediciones. • Ajuste de confusores está disponible para toda la población.	• Los datos son recolectados por personas distintas al investigador. • Falta de información de confusores. • Dificultad de manejar datos perdidos. • Calidad de datos baja o desconocida. • Falta de truncación. • Diferencias sin importancia se vuelven significativas.

En este proceso, el paso (1) es imaginario y se asume que es resultado del muestreo de elementos independientes e idénticamente distribuidos. En base a lo anterior, el uso de métodos de inferencia frecuentista no es descabellado, sino más bien necesario.

Frente a esto, el diálogo entre Alicia y el Gato de Cheshire en el bosque del País de las Maravillas no es surreal sino que del todo plausible y sensato ¿Qué camino debo seguir? [5].

6. Poblaciones aparentes: ¿Inferir o no inferir? Ésa es la cuestión

Entonces corresponde analizar la situación de qué hacer cuando el conjunto de datos disponibles es toda la información existente y por ende, no hay más datos que recolectar. A este conjunto de todas las unidades se le denomina población aparente [6]. Al respecto, los problemas conceptuales son referidos, en primer lugar, a que los datos no son obtenidos a partir de muestras generadas por un muestreo probabilístico y, en segundo, a que la población aparente es el resultado de un mecanismo de generación de datos que produce una sola muestra con un solo set de datos, es decir, no se puede esperar otro.

Ante esto, y en respuesta a la pregunta de Alicia, se perfilan dos caminos: (i) que los datos sean tratados como fijos, lo que significa que la población aparente es una población verdadera y en consecuencia la estadística inferencial es irrelevante; o (ii) que los datos sean tratados como la realización de una variable aleatoria, en una muestra a partir de algún de proceso que le da origen y que podría, en principio, producir un número muy grande de otras realizaciones. Estas realizaciones, en consecuencia, constituyen una súper-población.

Hasta aquí todos los caminos presentan escollos. Refiriéndonos a la inferencia frecuentista, ¿es realista asumir que el diseño de muestreo fue probabilístico cuando en la práctica suele romperse o no seguirse al pie de la letra, o que la interpretación de la incertidumbre acerca de estimaciones de parámetros como distribución de un resultado sobre un número hipotético y muy grande de ensayos idénticos e independientes sean consistentes con cómo fueron generados los datos? Con esto, ¿qué tan buenas pueden ser las herramientas inferenciales al considerar los supuestos imaginarios del largo plazo?

Ahora, si la población aparente es considerada una población verdadera, los datos no podrían haber sido diferentes, por lo tanto, el proceso de generación de datos es determinístico y el uso de inferencia estadística es irrelevante. Si, por otro lado, se la considera como resultado de un proceso estocástico, debe definirse la población que da origen a esta población aparente, con lo cual pasaría a ser una realización de la súper-población, con lo que al aplicar métodos de inferencia frecuentista lo que se obtendrá de tal proceso es una distribución de muestreo de las poblaciones aparentes. Sin embargo, este razonamiento no considera que la población aparente no es generada por medio de un muestreo aleatorio y además que esta población aparente viene a ser la única realización posible, por lo que no presenta error de muestreo.

Clarificado el *quid* del debate se han levantado voces de uno y otro lado. Silva Ayçaguer [7], al referirse al uso de pruebas de significación cuando los datos son poblacionales, es claro al señalar que la inferencia en esos casos carece de sentido. Aún más, considerando a las súper-poblaciones, este enfoque en raras ocasiones es declarado por los investigadores, es decir, nunca queda definida la súper-población hacia la cual va dirigida la inferencia. Por otro lado, se plantea la dificultad teórica el súper-universo como un constructo forzoso para darle sentido a la inferencia cuando se usan datos poblacionales.

En la misma línea, aunque menos crítico del concepto de súper-poblaciones, Schneider [8] recalca el hecho de que si bien es posible realizar inferencias a una población infinita imaginaria, no se debe abusar del concepto. Es decir, algunas súper-poblaciones son plausibles de existir, mientras que otras no. Para estas últimas, el uso de métodos inferenciales no corresponde y responde más bien al sinsentido de querer interpretar los datos como procedente de una población imaginaria de modo que se pueda hacer inferencia. Vale hacerse la pregunta ¿qué significado puede ser extraído de tal proceso imaginario?

Por último, ante esta disyuntiva frecuentista, el paradigma bayesiano ofrece una aproximación distinta [6, 9]. Considerando la interpretación frecuentista de la probabilidad como el valor límite de frecuencia relativa de un evento como número independiente de ensayos que crece hasta el infinito, puede considerarse la concepción subjetiva de la probabilidad, es decir, una expresión de la incertidumbre basada en la experiencia. Esto se conoce como inferencia Bayesiana.

Este enfoque de la inferencia inicia con una densidad de probabilidad *a priori* que expresa la creencia que se tiene sobre un evento. Luego, esta creencia previa es revisada considerando nueva información, *i.e.*, los datos obtenidos de una muestra. La consideración en simultáneo de ambas informaciones mediante el empleo del teorema de Bayes da como resultado la obtención de una distribución de probabilidad *a posteriori*, que expresa una actualización de la información *a priori* a la luz de la nueva información del evento. Bajo este enfoque parece tener más sentido el uso de datos poblacionales, los cuales pueden ser analizados bajo la información previa que se posea. Este enfoque, sin embargo, no es de uso trivial puesto que depende en gran medida de que tan confiable es la información *a priori*

con la que se cuenta [6, 7].

Para finalizar, en respuesta a la inquietud de Alicia, trabajar con este tipo de datos requiere de reflexión abundante, profusa y compartida con pares respecto a lo que dilucidar a partir de los datos y en cuanto a la estrategia metodológica que se emplee para que sus conclusiones tengan sentido. Esto ya que si no se tiene claridad de lo que se quiere conocer y de cómo se conocerá lo desconocido, la respuesta del Gato Cheshire es iluminadora: "da lo mismo un camino que otro", puesto que no se sabrá qué es lo que se está conociendo.

Si se quieren explicitar estas preguntas podrían formularse durante todo el proceso investigativo al usar datos poblacionales: ¿qué información contienen?, ¿con qué instrumentos se recolectó la información?, ¿constituyen una población o una muestra?, si es una población y se quiere inferir, ¿qué se quiere inferir?, ¿tiene sentido imaginar una súper-población que dio origen a los datos?, ¿se están utilizando todos los métodos disponibles para abordar el problema?, ¿es aplicable la inferencia frecuentista?, ¿será más adecuado el enfoque Bayesiano?, ¿qué dice el (bio)estadístico de los datos y del problema? Y, por supuesto, la más importante: ¿qué se quiere conocer?

Referencias

[1] Marschner IC. (2014). *Inference Principles for Biostatisticians*. New York: Chapman and Hall/CRC.

[2] Lohr SL. (2021). *Sampling: Design and Analysis*, 3rd Ed. Boca Raton: CRC Press.

[3] Thygesen LC, Ersbøll AK. (2014). When the entire population is the sample: strengths and limitations in register-based epidemiology. *European Journal of Epidemiology*, *29*, 551–558. 10.1007/s10654-013-9873-0

[4] Hartley HO, Sielken RL. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics*, *31*(2), 411-422. 10.2307/2529429

[5] Carroll L. (1922). *Alicia en el País de las Maravillas*. Madrid: Editorial Rivadeneyra.

[6] Berk RA, Western B, Weiss RE. (1995). Statistical Inference for Apparent Populations. *Sociological Methodology*, *25*, 421-458.

[7] Silva Ayçaguer LC. (1997). *Cultura estadística e investigación científica en el campo de la salud: una mirada crítica*. Madrid: Díaz de Santos.

[8] Schneider JW. (2016), The imaginarium of statistical inference when data are the population: Comments to Williams and Bornmann. *Journal of Informetrics*, *10*(4), 1243–1248. 10.1016/j.joi.2016.09.011

[9] Rubin DB. (1995). Bayes, Neyman, and Calibration. *Sociological Methodology*, *25*, 473-479. 10.2307/271076