

## Ignorar o no ignorar, esa es la cuestión

Eduardo Alarcón-Bustamante<sup>1</sup>

*“The credibility of inferences decreases with  
The strength of the assumptions maintained”.*  
The law of Decreasing Credibility (Manski, 2013)

### Introducción

Pensemos en la siguiente situación: un profesor revisa las pruebas de su curso y, al momento de ingresar las notas al libro, se percató de que falta la nota de un estudiante. Angustiado, verifica si este rindió la prueba y se da cuenta que sí lo hizo. El gran problema del profesor es cómo inferir acerca de la nota de este estudiante. Una posibilidad es tratar de predecir la nota e *imputarla*. “Inteligentemente”, el profesor busca compañeros del estudiante que sean *gemelos* entre ellos, digamos: mismo sexo biológico, mismo sector de residencia y misma edad. Calcula el promedio de notas de los gemelos, obteniendo así la nota de la prueba extraviada (supongamos 3,5). Conclusión: *el estudiante obtuvo una calificación que no es suficiente ni siquiera para aprobar la prueba*. El supuesto que hace el profesor es: todos aquellos estudiantes con las mismas características observadas tienen, en promedio, la misma nota.

---

<sup>1</sup> Núcleo Milenio de Movilidad Intergeneracional, MOVI. Laboratorio Interdisciplinario de Estadística Social LIES, Facultad de Matemáticas, Pontificia Universidad Católica de Chile. Escuela de Psicología, Pontificia Universidad Católica de Chile. Este trabajo ha sido parcialmente financiado por el Proyecto FONDECYT de Postdoctorado 3220422.

Al momento de ingresar las notas al sistema el profesor se percata de que el estudiante tiene todas sus notas sobre 6,5 y piensa que el 3,5 no refleja su realidad. Finalmente, decide calcular el promedio de sus notas anteriores e imputarla en la nota faltante (supongamos 6,8). Conclusión: *el estudiante tuvo la mejor nota del curso en la prueba*. El supuesto es: el estudiante tiene un rendimiento homogéneo en todas las pruebas.

Notemos que, frente a la misma situación y a los mismos datos (calificaciones del estudiante y sus compañeros), el resultado, a causa del supuesto hecho, cambia rotundamente. ¿Por qué ocurre esto? Los datos no hablan por sí solos: son los supuestos que uno impone los que ayudan a obtener las conclusiones (resultados). Esto es lo que Manski (2013) define como “lógica de la inferencia empírica”: “La inferencia requiere supuestos que relacionen los datos con la población de interés”.

A pesar de que el profesor tiene una solución a su problema, la pregunta es ¿cuál era la nota real? ¿1,0; 2,1; 3,5; 6,3; 6,8; 7,0? Nunca lo sabremos, ya que el profesor extravió la prueba. Aunque 3,5 y 6,8 son dos posibilidades (pues lo único que sabemos es que las notas en Chile están entre 1,0 y 7,0), tenemos algunas situaciones en las que, dependiendo nuevamente de los supuestos, esta calificación sería totalmente distinta:

- La persona no estudió y entregó la prueba en blanco: hubiese obtenido 1,0.
- La persona no estudió, pero adivinó una respuesta: obtuvo 1,4.
- La persona, a pesar de haber estudiado mucho, no entregó la prueba a la perfección: obtuvo 6,8.
- La persona estudió mucho y entregó todo perfecto: hubiese obtenido 7,0.

El profesor ignoró completamente el potencial comportamiento del estudiante e imputó su calificación. Lo más delicado, no consideró el impacto que esta decisión puede tener. Imputar la nota fue seductor pero puede ser muy peligroso (Dempster y Rubin, 1983). Seductor, porque hace creer al profesor que solucionó el problema de su prueba extraviada; peligroso, porque esta imputación puede ser completamente distinta a la realidad.

Tanto en investigación en ciencias sociales como en otras ciencias nos encontramos con el problema de hacer inferencias de poblaciones de interés que no observamos completamente. Por ejemplo, en esta época de pandemia por covid-19 se ha reportado diariamente en Chile la tasa de positividad “en términos de país”; sin embargo, esta tasa de positividad es respecto de la población que sí accedió a realizarse el test. Entonces, ¿por qué extrapolamos los resultados como si todo Chile se lo hubiese realizado? Estos resultados son obtenidos ignorando el potencial comportamiento de la positividad entre quienes no se tomaron el test. Un potencial comportamiento es el siguiente: *las personas que piensan no estar infectadas tienden a realizarse menos el test y se puede pensar que la proporción de infectados que no se tomó el test es menor a la proporción de infectados que sí lo hizo.*

Para algunos profesores universitarios otra situación de interés es la evaluación docente. En este contexto, sabemos que solo un porcentaje de estudiantes contesta la encuesta docente; sin embargo, al momento de hacer los reportes se recalca, por ejemplo: “el 45% ciento de los estudiantes del curso no recomienda al profesor”. La realidad es que “entre aquellos estudiantes que respondieron la encuesta docente, el 45% no recomienda al profesor”. La extrapolación, por lo general, ignora las potenciales respuestas de aquellos estudiantes que no respondieron la encuesta. Podemos, por ejemplo, pensar en el siguiente comportamiento: “las mujeres que no respondieron la encuesta tienden a rechazar más al profesor que aquellos hombres que sí la respondieron”.

En todas las situaciones antes descritas, la parte de la población de la que se ignora el potencial comportamiento es considerada como dato no observado (“datos faltantes”, en la jerga estadística). La gran pregunta es ¿cómo hacer inferencias con estos datos faltantes? Una forma fácil, y ya vemos que muy utilizada, es ignorarlos, eliminarlos y trabajar solo con lo completamente observado (independiente del contexto en el que estemos trabajando). También existen algunos procedimientos estadísticos que intentan atacar este problema; sin embargo, ninguno de ellos está sustentado en los datos recolectados, sino más bien en supuestos acerca de los datos faltantes (para detalles técnicos de algunos de estos procedimientos, ver Little and Rubin, 2019), los cuales, en palabras de Manski (2007), son irrefutables: ante todo, los datos no son observados.

Se puede demostrar matemáticamente que ignorar los datos no observados es análogo a asumir que el comportamiento de lo que no observamos es exactamente el mismo que aquel que sí observamos. Esto, a su vez, es lo mismo que decir que el comportamiento de la población es completamente conocido, pero solo hay algunos datos que no observo. Sin embargo, como asumimos que se conoce completamente el comportamiento de la población, observar o no un(os) dato(s) es irrelevante para hacer inferencias sobre ella. Por otra parte, ya que asumimos que conocemos su comportamiento, es contradictorio pensar en hacer inferencias sobre una población con una parte de ella, si ya aceptamos que conocemos su comportamiento completo.

El asunto es ¿qué tan dispuestos estamos a creer en el supuesto que ignora los datos no observados? Este ensayo pretende dar a conocer algunas situaciones de la investigación empírica en las que los supuestos utilizados para tomar decisiones ignoran el potencial comportamiento de los datos no observados. Específicamente, se analizan dos problemas que ofrece la ausencia de datos faltantes: predicción del rendimiento de estudiantes universitarios a través de puntajes en el test de selección y análisis de encuestas.

## **Predicción del rendimiento de estudiantes universitarios a través de puntajes en el test de selección**

Para hacer selección de personal o selección a la universidad, frecuentemente se utiliza en este proceso un test estandarizado. Sabemos (o queremos creer) que este tipo de test selecciona a aquellos postulantes que tienen una mayor probabilidad de tener éxito, por ejemplo, en un programa de pregrado en la universidad. En estos test se espera que mayores puntajes se transformen en mejor rendimiento y, por lo tanto, la decisión de seleccionar o no es tomada con un puntaje de corte y seleccionando aquellos postulantes con un puntaje sobre este corte (Alarcón-Bustamante, San Martín & González, 2020).

La calidad del test es generalmente evaluada mediante su “capacidad predictiva”, en el sentido de que nos gustaría saber qué tanto se puede predecir el rendimiento del postulante con los puntajes del test. De acuerdo con Donoso (1998), una de las razones de la evolución del sistema de admisión en Chile ha sido la necesidad de aumentar la capacidad predictiva de las pruebas aplicadas. Históricamente, la evaluación de la capacidad predictiva se ha hecho con procedimientos estadísticos que relacionan la variable de interés (promedio ponderado acumulado al primer año de universidad) y el puntaje obtenido en el test. Sin embargo, como solo se selecciona a una proporción de los postulantes, el rendimiento de los no seleccionados no puede ser observado. En consecuencia, los estudios de capacidad predictiva se hacen solo con aquellos postulantes seleccionados en los que se debería observar el rendimiento. Más aún, la situación puede ser un poco más “grave” pues, en ocasiones, tampoco el puntaje en el test es observado (por ejemplo, en una prueba opcional). En este caso, una práctica habitual es también trabajar con aquellos postulantes seleccionados a quienes se observa el puntaje. En este contexto, por ejemplo, Geiser & Studley (2002), en su estudio de capacidad predictiva de los puntajes del SAT acerca del rendimiento en estudiantes de la Universidad de California, declaran que:

The only exclusions from the sample were students with missing SAT scores or high-school GPAs; students who did not complete their freshman year and/or did not have a freshman GPA recorded in the UC Corporate Student Database...

Si bien este modo de operar es válido, otros procedimientos estadísticos intentan lidiar con este problema: coeficientes de correlación corregidos por restricción de rangos (e.g., Mendoza & Mumford, 1987) o estimadores de coeficientes de regresión que consideran el sesgo de selección (Heckman, 1976; 1979). Aunque estos procedimientos se ven promisorios y sirven para aprender sobre la capacidad predictiva del test sobre el rendimiento de los estudiantes, es explícitamente asumido que los datos no observados provienen de la misma distribución de lo observado, lo cual es solo una opción dentro de infinitas posibilidades. De hecho, hay autores que prefieren no utilizar estos procedimientos. Por ejemplo, Koretz & Lengi (2018), declaran que una de las razones de no usar este procedimiento en su análisis es que

...the conventional correction for restriction of range can be seriously misleading if the selection function differs from the simple selection assumed in the derivation of the correction.

De acuerdo con Manzi et al. (2008), al analizar de este modo se puede caer en una subestimación de la capacidad predictiva del test de selección. Sin embargo, como no se observan los potenciales rendimientos, esta subestimación que se menciona puede ser incluso una sobreestimación: nunca lo sabremos.

Como hemos mencionado, hacer el análisis de la capacidad predictiva solo con aquellos seleccionados es equivalente a *ignorar* los potenciales rendimientos de quienes no fueron seleccionados. Al respecto, el Ministerio de Educación, el año 2014<sup>2</sup>, recomienda:

---

<sup>2</sup> Recuperado el 11 de mayo de 2022 desde [https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2017/06/A3N23\\_EvaluacionPSU.pdf](https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2017/06/A3N23_EvaluacionPSU.pdf)

Sofisticar los análisis para evaluar la capacidad predictiva de la prueba en la población general. Los análisis hasta ahora evalúan el desempeño de la prueba para predecir los desempeños de alumnos que fueron seleccionados en las universidades estudiadas. No sabemos si estos resultados son aplicables para predecir los desempeños potenciales de los estudiantes que quedaron fuera.

Pensemos en estos potenciales rendimientos como *datos faltantes* en un problema de selección, pues (siempre) observamos los puntajes obtenidos en el test de los postulantes, pero no así el rendimiento. La pregunta es, entonces, ¿cómo podemos considerar los potenciales rendimientos si no los observamos? La respuesta, aunque sea un poco larga, sería la siguiente: el supuesto que permite trabajar solo con aquellos postulantes a los cuales se les observa el rendimiento, es el siguiente: *no importa si fue o no seleccionado, el puntaje obtenido en el test me entrega toda la información que necesito respecto del potencial rendimiento y, por lo tanto, ignoraré lo que no observo.*

Si usted está dispuesto a creer en este supuesto, está bien: no sabemos cómo son los potenciales desempeños y el supuesto anteriormente mencionado es “un” posible escenario, entre muchos. A modo de ejemplo, uno puede pensar en los siguientes:

- Un postulante no seleccionado con puntajes bajos tendrá menor rendimiento que un seleccionado con puntajes altos.
- Si asumimos que el proceso de selección fue correcto, entonces un postulante no seleccionado con un puntaje cualquiera tendrá un menor rendimiento que aquel postulante seleccionado con el mismo puntaje.
- Si asumimos que el proceso de selección fue incorrecto, entonces un postulante no seleccionado con un puntaje cualquiera tendrá un mayor rendimiento que aquel postulante seleccionado con el mismo puntaje.

Así como en estos escenarios, que se comparan entre individuos seleccionados y no en una misma carrera, podemos pensar en otros

que involucren la misma carrera, pero en distintas universidades y con distintos niveles de exigencia. En este sentido, pensemos en lo que alguna vez hemos escuchado: si hubiese entrado en una universidad más fácil (más difícil), me hubiese ido mejor (peor).

Todos los escenarios posibles expuestos no ignoran los potenciales desempeños de los no seleccionados y pueden ser incluidos dentro del modelo que se utilice para aprender sobre la capacidad predictiva de un test de selección. Por ejemplo, en Alarcón-Bustamante et al. (2020), una característica esperada del test de selección es utilizada para considerar los potenciales rendimientos de los estudiantes no seleccionados, a saber, “a mayor puntaje en el test, mayor será el rendimiento”. Bajo este supuesto, se puede demostrar que el modelo de regresión tendrá una pendiente no negativa como función del puntaje. Así, los autores entregan una banda de todos los valores plausibles de la capacidad predictiva del test que son concordantes con este supuesto. Entonces, si es posible considerarlos en el modelo a utilizar, ¿es justo ignorar estos potenciales rendimientos?

## **Análisis de encuestas**

Todos estamos conscientes de que una de las principales razones para hacer inferencias de la población a partir de una muestra es el alto costo monetario que tiene extraer muestras grandes o estudiar toda la población.

Antes de pensar en el análisis de una encuesta, pensemos de dónde provienen los datos que se analizarán: 1) no todas las personas son seleccionadas para contestar la encuesta; 2) si son seleccionadas, no están obligadas a contestarla; 3) las personas que acceden a contestarla no están obligadas a contestar todas las preguntas. Estos tres puntos nos llevan a tener tres fuentes de datos faltantes: a) las posibles respuestas de los no seleccionados; b) las posibles respuestas de quienes, a pesar de ser seleccionados en la muestra, no accedieron a contestar la encuesta, no eran realmente elegibles o simplemente no se sabe si eran elegibles,



y c) las posibles respuestas de quienes, a pesar de ser seleccionados y acceder a contestar la encuesta, no respondieron alguna pregunta.

Si existen estas tres fuentes con datos que no observamos, ¿por qué, al momento de reportar resultados y extraer conclusiones, se reportan solo respecto de las respuestas observadas? Más aún, se reportan como si fueran las de toda la población de interés. La respuesta a esto es simplemente porque se asume que quienes no contestaron, de haber contestado, lo hubiesen hecho igual que quienes sí lo hicieron<sup>3</sup>.

A continuación haré la disección de cada uno de los puntos que llevan a las diferentes fuentes de datos faltantes, mostrando el supuesto que se usa para poder trabajar solo con los datos observados.

## **1) Tratamiento para las posibles respuestas de los no seleccionados**

Este punto se puede considerar como parte de un diseño del muestreo. En palabras técnicas se denomina “diseño de muestreo ignorable”: *cuando se conocen características de interés (e.g., estratos, conglomerados, sexo biológico, nivel socioeconómico, entre otros), ser seleccionado o no en la muestra no es relevante para hacer inferencia sobre la variable que se está estudiando*<sup>4</sup>. Así, considerar que ser seleccionado o no en una muestra (cuando se conocen otras características) no es relevante para hacer inferencias, permite trabajar solo con las unidades seleccionadas y es el primer paso del muestreo: seleccionar quienes contestarán la encuesta. Ello no significa que los seleccionados estén dispuestos a contestarla; así pasamos al punto 2.

---

<sup>3</sup> Para conocer una forma alternativa de reportar resultados de una encuesta que considere distintas fuentes de no respuesta, ver San Martín y Alarcón-Bustamante (2022).

<sup>4</sup> Para detalles técnicos, ver Scott (1977) y Sugden y Smith (1984).

- 2) Tratamiento para las posibles respuestas de quienes, a pesar de ser seleccionados en la muestra, no accedieron a contestar la encuesta, no eran realmente elegibles o simplemente no se sabe si eran elegibles

En este punto pondré dos ejemplos concretos en la palestra: encuestas CASEN y CADEM<sup>5</sup>.

## Encuesta CASEN

La encuesta CASEN es ampliamente utilizada en Chile para estimar ingresos y pobreza a nivel nacional. Para ello se toma una muestra de viviendas y se realiza una entrevista a una persona idónea dentro de la vivienda. Los resultados son extrapolados y las conclusiones son extraídas. En su última versión, denominada “CASEN en pandemia 2020”, el proceso para obtener las respuestas consistió en dos etapas: una vez definidas las potenciales viviendas para la entrevista<sup>6</sup>, la primera etapa fue de precontacto, cuyo objetivo fue “solicitar a los habitantes de las viviendas seleccionadas datos de contacto telefónico de sus residentes, así como informar sobre la aplicación de la encuesta”. Esta se aplicó en la segunda etapa en aquellas viviendas que dieron un número de contacto en la etapa anterior y “fue aplicada en la mayoría de las viviendas a través de entrevistas telefónicas. Solo en un acotado número de casos, ubicados en zonas aisladas o sin cobertura telefónica, la encuesta se realizó únicamente de manera presencial”.

Al final del proceso, las conclusiones obtenidas se basaron en las respuestas entregadas en 62.540 viviendas. Como es de esperar, no

---

<sup>5</sup> Para un análisis crítico de las encuestas CASEN y CADEM, puede ver las columnas científicas, escritas por Ernesto San Martín y Eduardo Alarcón-Bustamante. “¿Podemos confiar en la desigualdad que nos muestra la Casen?” y “Por qué no podemos confiar en la Encuesta CADEM”, publicadas en terceradosis.cl. Además, puede ver la columna “¿Qué hace más o menos confiable una encuesta?”, del diario *Visión UC*.

<sup>6</sup> Puede ver la estrategia muestral en: [http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2020/Diseno\\_Muestral\\_Casen\\_en\\_Pandemia\\_2020.pdf](http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2020/Diseno_Muestral_Casen_en_Pandemia_2020.pdf)

llegaron a esta cantidad inmediatamente, sino que hubo viviendas que no fue posible contactar, que no eran elegibles (empresas, viviendas de veraneo, dormitorios, hospitales, entre otras), que rechazaron contestar la entrevista, etc.

Sin ir más lejos, en el documento “Desempeño en el trabajo de campo”<sup>7</sup> se puede distinguir las siguientes definiciones y números de viviendas que clasifican en cada una de ellas para ambas etapas:

**Tabla 1.**

*Definiciones de trabajo de campo de CASEN y número de viviendas en cada definición en ambas etapas.*

Grupo	Definición	Etapas de precontacto	Etapas de aplicación (solo en aquellas que se logró entrevista en precontacto)
Entrevistada	Entrevista completa	86.189	62.540
Rechazo	Incluye situaciones de rechazo al inicio de la encuesta como entrevistas interrumpidas. En este grupo también quedan clasificadas las entrevistas completadas de manera parcial, las que quedan con código de entrevistas interrumpidas.	3.138	7.804
No contacto	Corresponde a viviendas u hogares seleccionados en la muestra donde no fue posible interactuar con un informante idóneo. Incluye los casos en que se realizó contacto con un integrante del hogar inhabilitado para actuar como informante idóneo.	8.403	15.591

<sup>7</sup> Recuperado el 18 de mayo de 2022 desde [http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2020/Nota\\_tecnica7\\_Desempeno\\_del\\_Trabajo\\_Campo\\_Casen\\_en\\_Pandemia\\_2020.pdf](http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2020/Nota_tecnica7_Desempeno_del_Trabajo_Campo_Casen_en_Pandemia_2020.pdf)

Otras no entrevistadas	Corresponde a viviendas u hogares seleccionados en la muestra en los que fue posible interactuar presencialmente con un informante idóneo, pero no fue posible aplicar la encuesta por razones distintas al rechazo.	118	189
Elegibilidad desconocida <sup>8</sup>	Para estos casos se desconoce si la vivienda está ocupada durante el periodo de medición. Esta situación puede ocurrir debido a problemas de localización, dificultades de acceso o desinformación sobre el uso de la edificación. Este grupo de códigos solo fue utilizado en el precontacto o la aplicación presencial de la encuesta.	1.324	<i>“Dada la cercanía temporal entre el proceso de precontacto y la fase de aplicación no hay códigos de elegibilidad desconocida para el proceso de entrevista telefónica. En este sentido, <b>las unidades que no logran ser contactadas al número de teléfono recolectado durante el precontacto se consideran dentro del grupo de las unidades elegibles no contactadas</b>”.</i>
No elegibles	Unidades seleccionadas en la muestra, que al momento de ser contactadas para la realización de la encuesta tenían un uso distinto al de vivienda particular ocupada.	3.858	65

Como se puede apreciar, las estimaciones de pobreza e ingreso son realizadas con entrevistas realizadas en 62.540 viviendas, las cuales son solo una pequeña proporción de las que se intentó contactar (99.172); entonces, ¿por qué se extrapola tan fácilmente la información a toda la población? La respuesta es la misma: se ignoran las potenciales respuestas de quienes, por algún u otro motivo, no estaban disponibles para contestar la encuesta. Así, todos los análisis y procedimientos estadísticos son realizados con la información disponible: *estar o no*

<sup>8</sup> Respecto de viviendas con elegibilidad desconocida, el documento de desempeño en el trabajo de campo menciona: *“el cálculo de las tasas de rechazo, contacto y respuesta que se presentan en este documento, corresponde a una medida conservadora respecto de otras formas de cálculo definidas por AAPOR para estos indicadores, en tanto asumen que las unidades de elegibilidad desconocida son viviendas elegibles y se incluyen en el denominador de estas”.*

*en la muestra es irrelevante para hacer inferencias sobre la población objetivo.*

## Encuesta CADEM

La encuesta CADEM fue diseñada para “entregar semanalmente información confiable, oportuna y contingente sobre el debate político, económico y social de Chile”<sup>9</sup>. Su población objetivo son personas de 18 años o más que habiten en el territorio nacional, y consiste en preguntar semanalmente, mediante llamados telefónicos, la opinión sobre diversos temas a aproximadamente 700 personas. El proceso se lleva a cabo hasta conseguir este número deseado de personas que contesten la encuesta. En este proceso se puede distinguir diferentes fuentes de potenciales opiniones que no son observadas, pues el proceso es como sigue:

- CADEM cuenta con una base de datos de más de 18 millones de números de teléfono según su diseño metodológico. En este mismo documento declaran que la población de 18 años o más asciende a un poco más de 13 millones. Entonces, al llamar por teléfono no se sabe si la persona que (potencialmente) contestará el teléfono cumple con el requisito (es elegible).
- Si no contesta la llamada, entonces nunca se sabrá si la persona era elegible o no. Si la contesta y es elegible (es seleccionado para contestar la encuesta), no está obligada a contestar la encuesta.
- El proceso se detiene en el momento en que 700 personas elegibles, aproximadamente, acceden a contestar la encuesta.

---

<sup>9</sup>Diseño metodológico CADEM, recuperado de <https://cadem.cl/wp-content/uploads/2022/03/Metodologia2020.pdf>

Así, tenemos dos fuentes de potenciales opiniones: la de los seleccionados que no accedieron a contestar la encuesta y la de los elegibles que no fueron contactados. CADEM es transparente en definir cómo ignora las potenciales respuestas de quienes no contestaron la encuesta:

*Bajo el supuesto de que quienes rechazan contestar son iguales a quienes contestan, la magnitud de la tasa de rechazo no ofrece mayores inconvenientes, pero cuando existe evidencia que ambos grupos no son equivalentes, el rechazo puede introducir serias distorsiones en los resultados...*

¿Es justo pensar de quienes no contestan que, de haberlo hecho, su respuesta habría sido la misma? Esta es “una” opción entre infinitas posibilidades. En otras palabras, si entre los que contestaron el 35% está de acuerdo con una afirmación, ¿será factible que entre quienes no contestaron también el 35% esté de acuerdo? Nunca lo sabremos, pues entre los que no contestaron hay muchos posibles resultados respecto de estar de acuerdo con la afirmación, a saber, entre 0% y 100%.

Por otro lado, y no menos importante, CADEM, al transparentar este supuesto, hace una afirmación que no es empíricamente medible: *cuando existe evidencia de que ambos grupos no son equivalentes, el rechazo puede introducir serias distorsiones en los resultados*. Ante todo, los datos no son observados y, por lo tanto, nunca sabremos si ambos grupos son o no equivalentes.

### **3) Tratamiento de las posibles respuestas de quienes, a pesar de ser seleccionados y acceder a contestar la encuesta, no respondieron alguna pregunta**

La tercera fuente de incertidumbre está relacionada con aquellas personas que no contestaron alguna pregunta en la encuesta. Para mostrar este punto, seguiré con el análisis de ambas encuestas anteriormente descritas

## Encuesta CASEN

A pesar de que en muchas de las preguntas realizadas en esta encuesta faltan datos, el tratamiento de ellos es respecto de dos variables de interés: ingreso de los trabajadores e ingreso de los pensionados<sup>10</sup>. De hecho, se declara que:

La metodología definida para la imputación de datos faltantes a las preguntas de ingreso se basa en la técnica de imputación por medias, asignando a cada dato faltante (receptor) el valor declarado en promedio por los casos más similares posibles (donantes)<sup>11</sup>...

Esta metodología asume que los ingresos son homogéneos entre los grupos definidos por los donantes, pues el dato faltante es imputado con el ingreso promedio observado de aquellos con características similares: ¿cuánto es el ingreso que no se declaró? Si bien una posibilidad es que el ingreso no reportado sea justamente el promedio, nunca sabremos dicho ingreso, pues no fue declarado. De hecho, puede ir desde \$0 hasta una cantidad finita desconocida.

Para ambos ingresos, de trabajadores y pensionados, podemos hacer otros supuestos que son menos fuertes que la homogeneidad entre sus gemelos y no ignoran el potencial ingreso no reportado. Por ejemplo, para los trabajadores asalariados, podemos pensar que *aquellos trabajadores con menos años de estudio que no reportaron el ingreso tendrán un sueldo inferior a aquellos que sí lo reportaron y tienen más años de estudio*.

## Encuesta CADEM

La encuesta CADEM no se hace problemas en este sentido, pues entre sus opciones está “No sabe o no contesta”. Sin embargo,

---

<sup>10</sup> El resto de datos faltantes se deja como tal y se trabaja solo con lo observado.

<sup>11</sup> Reporte técnico *Medición de los ingresos y la pobreza en Chile. Encuesta CASEN en pandemia 2020*. Recuperado desde [http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2020/Medicion\\_de\\_la\\_pobreza\\_en\\_Chile\\_2020.pdf](http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2020/Medicion_de_la_pobreza_en_Chile_2020.pdf)

es importante mencionar que también se puede hacer este tipo de supuestos que no ignoran las potenciales respuestas de los que no respondieron. Por ejemplo, ante las preguntas relacionadas con aprobación presidencial, se podría suponer que *las mujeres que no respondieron la pregunta de la aprobación presidencial tienen una mayor probabilidad de aprobar la gestión presidencial que aquellos hombres que sí la respondieron*. Este último supuesto no es medible (al igual que el supuesto que ignora los datos faltantes); sin embargo, es más razonable y considera otro potencial comportamiento de los datos no observados.

## Discusión

A lo largo del documento se han explicitado los supuestos tradicionalmente hechos para realizar inferencias sobre la población de interés. Estos supuestos recaen en ignorar los potenciales valores que pueden tomar los datos faltantes en diferentes situaciones y se concentran en solo un potencial escenario, a saber, *lo no observado se comporta igual que lo observado*. Sabemos que hay más escenarios posibles, pero todo el análisis se reduce exclusivamente a este. Sin embargo, es preciso recalcar que los datos no hablan por sí solos, sino que son los supuestos realizados los que ayudan para la extracción de conclusiones. De hecho, la lógica de la inferencia estadística es “datos + supuestos = conclusiones” (Manski, 2013). Así, para un conjunto de datos fijos, cambiar supuestos implica necesariamente que cambien las conclusiones.

Notemos que el supuesto de ignorabilidad es para todos los casos en los que faltan datos, independiente si estamos hablando de opiniones, salarios o rendimiento universitario. Entonces, ¿cómo es posible que la estadística sea una herramienta tan flexible que, sin importar el contexto, ignorar los datos faltantes siempre funciona? El problema no es la estadística, el problema recae en que quienes leemos reportes seamos capaces de entender que las conclusiones obtenidas son un fiel reflejo del supuesto. Sin embargo, los potenciales escenarios



propuestos en este trabajo están basados en el contexto en el que estamos y no se reducen a ignorar los datos faltantes.

Este documento pretende hacer un llamado urgente a los lectores de reportes técnicos involucrados en la toma de decisiones e implementación de políticas públicas: es estrictamente necesario entender que los supuestos utilizados son los que permiten extraer las conclusiones y no los datos desnudos. Si estamos dispuestos a creer en supuestos muy fuertes, debemos considerar que las conclusiones pueden ser extremadamente débiles, tal es el caso de pensar en que aproximadamente 700 personas representan la opinión de todo un país; esta débil conclusión es obtenida gracias a un fuerte supuesto. Lo anterior es lo que se denomina "ley decreciente de la debilidad de las conclusiones": entre más fuerte el supuesto, más débil es la conclusión (Manski, 2013).

## Referencias

- Alarcón-Bustamante, E., San Martín, E. & González, J. (2020). Predictive validity under partial observability. In M. Wiberg, M. Molenaar, J. González, U. Böckenholt & J. S. Kim (Eds.), *Quantitative Psychology*. IMPS2019. Springer Proceedings in Mathematics & Statistics, vol. 322. Springer, Cham.
- Dempster, A. P. & Rubin, D. B. (1983). Introduction. In I. O. W. G. Madow, *Incomplete Data in Sample Surveys (vol. 2): Theory and Bibliography* (pp. 3-10). New York: Academic Press.
- Donoso, S. (1998). La reforma educacional y el sistema de selección de alumnos a las universidades: impactos y cambios demandados. *Estudios Pedagógicos*, (24), 7-30.
- Geiser, S. & Studley, R. (2002). UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California. *Educational Assessment*, 8(1), 1-26.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 46, 931-961.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153-161.
- Koretz, D. & Langi, M. (2018). Predicting freshman grade-point average from test scores: Effects of variation within and between high schools. *Educational Measurement: Issues and Practice*, 37(2), 9-19.
- Little, R. & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. New York: Wiley.
- Manski, C. (2007). *Identification for prediction and decision*. New York: Harvard University Press.
- Manski, C. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*. New York: Harvard University Press.

- Manzi, J., Bravo, D., Pino, G. del, Donoso, G., Martínez, M. y Pizarro, R. (2008). *Estudio de la validez predictiva de los factores de selección a las universidades del consejo de rectores, admisiones 2003 al 2006* (Tech. Rep.). Comité Técnico Asesor, Honorable Consejo de Rectores de las Universidades Chilenas.
- Mendoza, J. & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, 12(3), 282-293.
- San Martín, E. & Alarcón-Bustamante, E. (2022). Dissecting Chilean surveys: the case of missing outcomes. *Chilean Journal of Statistics*, 13(1), 17-46.
- Scott, A. J. (1977). Some comments on the problem of randomisation in surveys. *Sankhyā O*, 39, 1-9.
- Sugden, R. & Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3), 495-506.