Datos Faltantes:

¿Imputarlos o expresar la incerteza inducida por ellos?

Laboratorio Interdisciplinario de Estadística Social LIES UC

Núcleo Milenio de Movilidad Intergeneracional, MOVI

Mayo, 2022

¿Qué hallarás en este documento?

- Encontrarás una introducción a la herramienta básica que utilizamos para generalizar o extrapolar los resultados de la Encuesta CADEM. Esta herramienta se llama **Ley de Probabilidades Totales**.

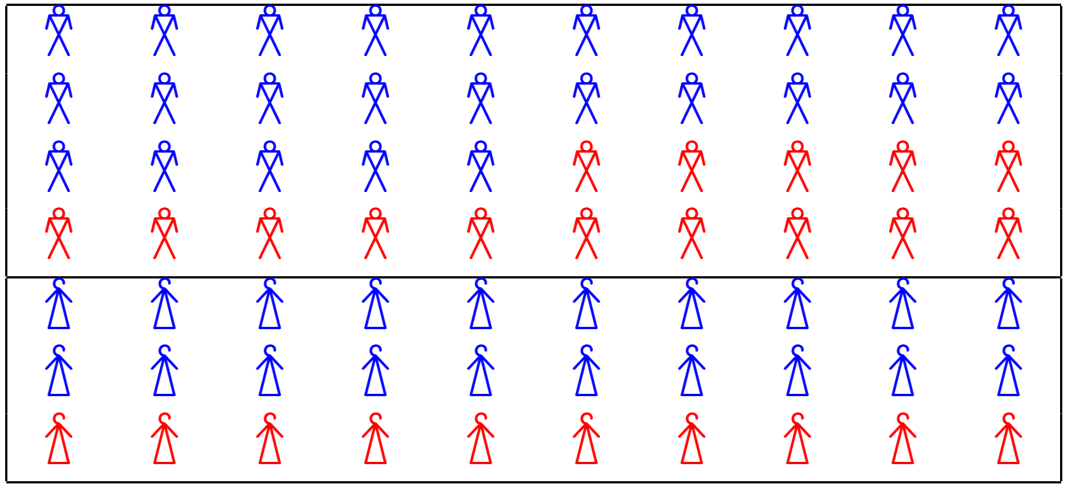
- Encontrarás cómo extrapolamos los resultados de la encuesta CADEM **explicitando la incerteza inducida por la tasa de no respuesta**

- Para mayor información, te invitamos a *leer el siguiente artículo*:

  E. San Martín & E. Alarcón-Bustamante (2022), *Dissecting Chilean Surveys: The case of missing outcomes Case*. Chilean Journal of Statistics, 13 (1) 17–45.

Ejemplo de una encuesta

- Vamos a suponer que de 70 personas encuestadas, 45 declaran que votarán **apruebo** en el plebiscito de salida, mientras que 25 declaran que votarán **rechazo**.
- El **espacio muestral** corresponde al conjunto de etiquetas o identificadores de cada uno de los encuestados.
- Los 70 encuestados son clasificados según su sexo biológico: hombre y mujer .
- Esta información puede representarse de la siguiente manera:



Probabilidades marginales

- La proproción (o probabilidad) de **apruebo** es igual a

$$P(\textbf{apruebo}) = \frac{45}{70}$$

–esto es, 64.29 %.

- La proporción (o probabilidad) de **rechazo** es igual a

$$P(\textbf{rechazo}) = \frac{25}{70}$$

–esto es, 35.71 %.

- Las probabilidades anteriores se calculan teniendo en cuenta **toda la población de encuestados**: por eso se llaman **probabilidades marginales**.

Probabilidades condicionales

- Las proporciones de **apruebo** y **rechazo** pueden reportarse también bajo una determinada **condición**, por ejemplo el sexo biológico.
- Así, la proporción de **apruebo** **entre los hombres** o **bajo la condición que el encuestado es hombre**, lo que escribimos como $P(\textbf{apruebo} \mid \text{🚹})$, está dada por

$$P(\textbf{apruebo} \mid \text{🚹}) = \frac{25}{40}.$$

- La proporción de **apruebo** **entre las mujeres** o **bajo la condición que la encuestada es mujer**, lo que escribimos como $P(\textbf{apruebo} \mid \text{🚺})$, está dada por

$$P(\textbf{apruebo} \mid \text{🚺}) = \frac{20}{30}.$$

- Cada una de estas probabilidades se calcula con respecto a la población definida por la **condición**, por lo que se llaman **probabilidades condicionales**.

Ley de Probabilidades Totales

- Es posible relacionar la probabilidad marginal $P(\textbf{apruebo})$ con las probabilidades condicionales $P(\textbf{apruebo} \mid \text{🧍})$ y $P(\textbf{apruebo} \mid \text{🧍‍♀️})$.
- Para ello solo es necesario considerar la proporción de hombres y de mujeres entre los encuestados:

$$P(\text{🧍}) = \frac{40}{70}, \qquad P(\text{🧍‍♀️}) = \frac{30}{70}.$$

Ley de Probabilidades Totales

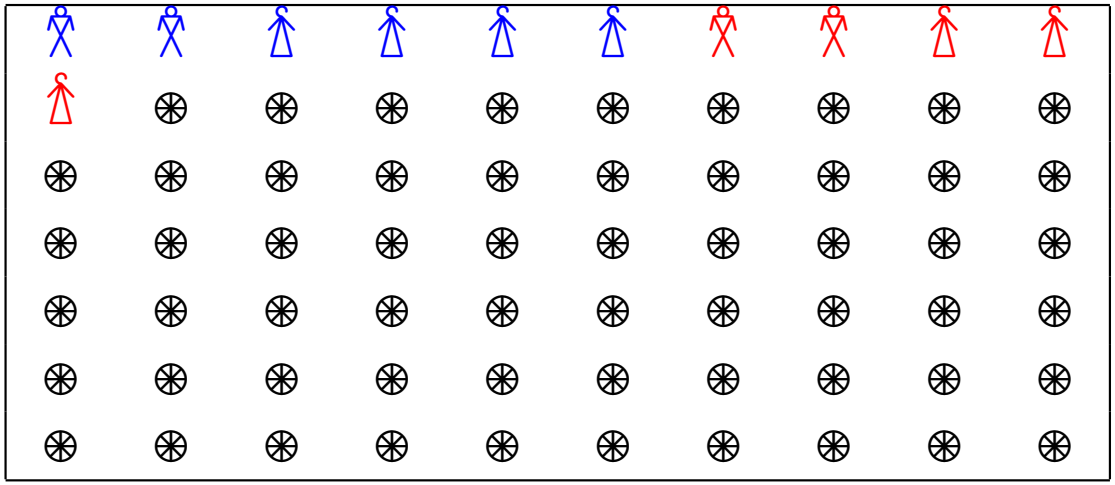
- Teniendo en cuenta entonces que de los 45 apruebo, 25 son hombres y 20 son mujeres, podemos hacer la siguiente descomposición:

$$
\begin{aligned}
P(\textbf{apruebo}) &= \frac{45}{70} \\
&= \frac{25+20}{70} \\
&= \frac{25}{40} \times \frac{40}{70} + \frac{20}{30} \times \frac{30}{70} \\
&= P(\textbf{apruebo} \mid \text{🚹}) \, P(\text{🚹}) \; + \; P(\textbf{apruebo} \mid \text{🚺}) \, P(\text{🚺})
\end{aligned}
$$

- La descomposición anterior siempre se puede hacer: basta descomponer el espacio muestral en subpoblaciones o grupos, y luego caracterizar las probabilidades condicionales de una variable de interés para cada grupo. Esto se llama **Ley de Probabilidades Totales**.

¿Qué ocurre cuando hay datos faltantes?

- Supongaamos que se encuestaron telefónicamente 70 personas (hombres y mujeres), de las cuales solo 11 de ellas accedieron a responder.
- De las que respondieron, 6 personas declararon que votarán **apruebo** en el plebiscito de salida, y 5 personas declararon que votarán **rechazo**.
- Las restantes personas **no quisieron responder la encuesta**: en el diagrama están representadas por ⊛, enfatizando el hecho que no siquiera se pudo recoger información adicional acerca de ellos.



¿Qué ocurre cuando hay datos faltantes?

- En este caso, el **espacio muestral** corresponde al conjunto de números telefónicos (celulares) de los encuestados.
- Los **datos faltantes** inducen una partición sobre el espacio muestral en dos grupos: aquellos que respondieron la encuesta, lo que denotaremos por "$Z = 1$", y los que no respondieron la encuesta, lo que denotaremos por "$Z = 0$".
- La información disponible nos permite calcular la proproción de encuestados que respondieron la encuesta y de los que no la respondieron:

$$P(Z = 1) = \frac{11}{70}, \qquad P(Z = 0) = \frac{59}{70}.$$

- $P(Z = 1)$ es denominada **tasa de respuesta** y $P(Z = 0)$ **tasa de no respuesta**.

¿Qué ocurre cuando hay datos faltantes?

- Para constatar el impacto de los datos faltantes sobre la variable de interés (en este caso, la proporción de los que **aprobarán** en el plebiscito de salida), usamos la Ley de Probabilidades Totales:

$$P(\textbf{aprueba}) = P(\textbf{aprueba} \mid Z = 1)\, P(Z = 1) \; + \; P(\textbf{aprueba} \mid Z = 0)\, P(Z = 0),$$

donde

  - $P(\textbf{aprueba} \mid Z = 1)$ corresponde a la proporción de los que aprobarán **entre los encuestados que accedieron a responder la encuesta**; esta proporción es igual a

$$P(\textbf{aprueba} \mid Z = 1) = \frac{6}{11}.$$

  - $P(\textbf{aprueba} \mid Z = 0)$ corresponde a la proporción de los que aprobarán **entre los encuestados que no accedieron a responder la encuesta**: **esta proporción ES IMPOSIBLE DE SER CALCULADA, pues se desconocen las preferencias de aquellos que no respondieron la encuesta**.

¿Qué ocurre cuando hay datos faltantes?

- Luego,

$$P(\textbf{aprueba}) = \frac{6}{11} \times \frac{11}{70} + \underbrace{P(\textbf{aprueba} \mid Z = 0)}_{\textbf{???}} \times \frac{59}{70},$$

de donde se concluye que es **imposible extrapolar los resultados que se obtienen a partir de los que respondieron la encuesta, a la población de interés que está representada por el espacio muestral**.

¿Cómo trata CADEM los datos faltantes?

- El Diseño Metodológico de CADEM señala lo siguiente:

  Estimar la magnitud del rechazo es fundamental debido a la relación directa que puede tener con los sesgos de autoselección en las encuestas de opinión pública. El cálculo de la tasa de rechazo se utiliza asimismo como una medida de validación de los resultados. **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**[1].

[1] Tomado de https://cadem.cl/wp-content/uploads/2021/07/Disen%CC%83o-Metodolo%CC%81gico-2018-VF.pdf

¿Cómo CADEM trata los datos faltantes?

- Usando la descomposición de $P(\textbf{apruebo})$ anteriormente señalada, CADEM asume que

$$P(\textbf{aprueba} \mid Z = 1) = P(\textbf{aprueba} \mid Z = 0).$$

- Este supuesto **NO es empíricamente testeable** pues depende del comportamiento no observado de quienes no responden la encuesta.

- Esto último advierte contra una afirmación sin contenido hecha por CADEM:

  [...] *pero cuando existe evidencia que ambos grupos no son equivalentes, el rechazo puede introducir serias distorsiones en los resultados.*

  Evidencia empírica jamás existirá pues lo que no se observa, no se observa.

- Finalmente, bajo dicho supuesto, **no hay necesidad de explicitar la población de interés** sobre la cual generalizar los resultados que se obtienen a partir de la encuesta.

La lógica de la investigación empírica

- La inferencia estadística requiere combinar datos con supuestos. De hecho,

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

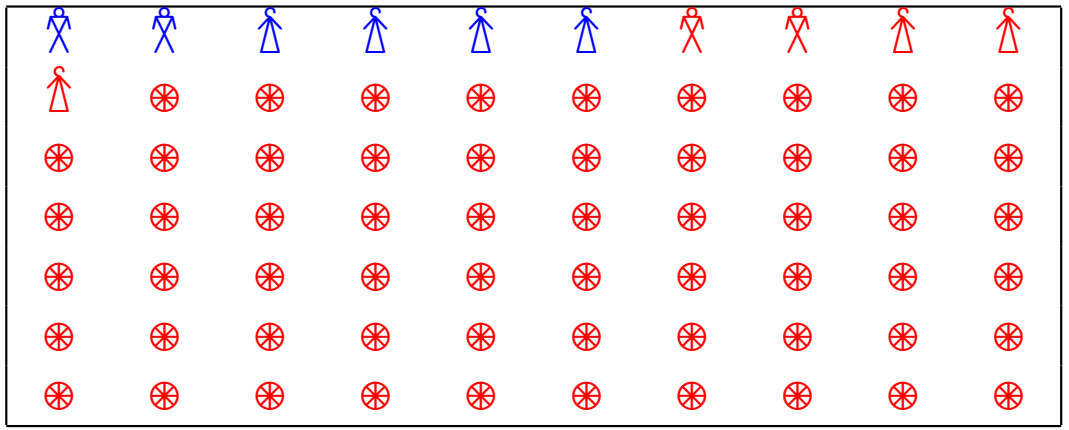
  Entonces, para un conjunto fijo de datos, cuando los supuestos cambian, las conclusiones pueden cambiar dramáticamente.
- Ya vimos que CADEM hace un supuesto que es imposible de testear ¿Qué pasa si hacemos un supuesto empírico y más creíble que el impuesto por CADEM?

Explicitar la incerteza inducida por los datos faltantes

- Volvamos a considerar la descomposición de $P(\textbf{aprueba})$:

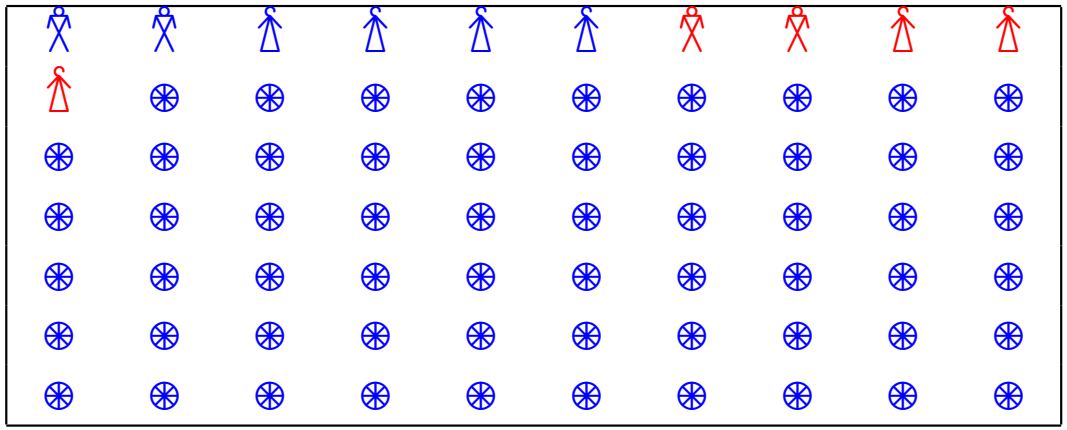
$$P(\textbf{aprueba}) = P(\textbf{aprueba} \mid Z = 1)\, P(Z = 1) \;+\; P(\textbf{aprueba} \mid Z = 0)\, P(Z = 0).$$

- La proporción desconocida $P(\textbf{aprueba} \mid Z = 0)$ siempre está entre dos valores extremos: puede ser al menos igual a 0, lo que equivale a afirmar que, entre todos aquellos que no respondieron la encuesta, ninguno hubiese **aprobado**; esto es,



Explicitar la incerteza inducida por los datos faltantes

- Dicha proporción puede ser a lo más igual a 1, lo que equivale a afirmar que, entre todos aquellos que no respondieron la encuesta, todos hubiesen **aprobado**; esto es,



Explicitar la incerteza inducida por los datos faltantes

- Cuando $P(\textbf{aprueba} \mid Z = 0)$ toma su mínimo valor (*entre todos los que no respondieron, nadie aprueba el plebiscito de salida*), entonces el mínimo valor que toma $P(\textbf{aprueba})$ es igual a

$$P(\textbf{aprueba} \mid Z = 1)\, P(Z = 1).$$

- Cuando $P(\textbf{aprueba} \mid Z = 0)$ toma su máximo valor (*entre todos los que no respondieron, todos aprueban el plebiscito de salida*), entonces el máximo valor que toma $P(\textbf{aprueba})$ es igual a

$$P(\textbf{aprueba} \mid Z = 1)\, P(Z = 1) + P(Z = 0).$$

- Por lo tanto, $P(\textbf{aprueba})$ pertenece al siguiente intervalo:

$$[\, P(\textbf{aprueba} \mid Z = 1)\, P(Z = 1),\ \ P(\textbf{aprueba} \mid Z = 1)\, P(Z = 1) + P(Z = 0)\,].$$

Explicitar la incerteza inducida por los datos faltantes

- Este intervalo representa **todos los posibles valores** que puede tomar $P($**aprueba**$)$ como fruto de la extensión de los resultados recolectados en la encuesta entre los que responden.

- Además, el mismo intervalo contiene una **medida de la incerteza inducida por la tasa de no respuesta**; esta medida corresponde al **largo del intervalo**, que es igual a

$$P(Z = 0),$$

  es decir, la tasa de los que no responden.

- Mayor es dicha tasa, mayor es la incerteza debida a la no respuesta.

Explicitar la incerteza inducida por los datos faltantes

- Usando la información del ejemplo, tenemos que la cota inferior del intervalo de $P(\textbf{aprueba})$ es igual a

$$\frac{6}{11} \times \frac{11}{70} \approx 0{,}086.$$

- La cota superior del intervalo de $P(\textbf{aprueba})$ es igual a

$$\frac{6}{11} \times \frac{11}{70} + \frac{59}{70} \approx 0{,}929.$$

- Por lo tanto, cuando se extrapolan los resultados obtenidos entre los que responden la encuesta a la población representada por el espacio muestral, lo único que podemos decir es:
  1. Si entre los que no respondieron, **nadie** aprueba el plebiscito de salida: un 8.6 % de la población de interés aprueba el plebiscito de salida.
  2. Si entre los que no respondieron, **todos** aprueban el plebiscito de salida: un 92.9 % de la población de interés aprueba el plebiscito de salida.