Técnicas para integración de muestras provenientes de diversas fuentes.

M. Rueda, B. Cobo Rodríguez, R. Ferri García, J. Rueda

UNIVERSIDAD DE GRANADA

Abril 2025

Index

1. Introducción
2. Métodos de estimación para corregir el sesgo de muestras no probabilísticas
3. Estimación con datos de muestreos no probabilísticos
4. Algunos ejemplos de encuestas
5. Integración de encuestas probabilísticas y no probabilísticas
6. Software

Introducción

Era digital: uso extendido de encuestas web.

- Rastreador de estudios de Covid-19 (**Matías y Levitt (2023)**) 90 % utilizó muestreos no probabilísticos
- Encuestas relacionadas con el covid en España (**Sánchez-Cantalejo et al 2023**) 73 %

Ventajas: flexibilidad, rapidez, bajo coste,...

Desafíos: para garantizar la efectividad y representatividad de los datos obtenidos.

Introducción

Objetivos

- Revisar las principales técnicas de estimación a partir de datos no probabilísticos.
- Breve introducción a los estimadores que integran datos de encuestas probabilísticas y no probabilísticas.
- Algunos ejemplos de encuestas reales en donde se han aplicado estas técnicas.
- Una descripción general del software disponible para aplicar estas técnicas.

Tipos de muestreo

Muestreo probabilístico

Cuando puede calcularse de antemano cuál es la probabilidad de obtener cada una de las unidades.
La selección de cada elemento es un experimento aleatorio.
Todos los elementos son elegibles

Muestreo no probabilístico

- Muestreo intencional u opinático
  En él es la persona que selecciona la muestra la que procura que ésta sea *representativa* (ejemplo: muestreo por cuotas)
- Muestreo sin norma
  En él se toma la muestra por razones de comodidad o capricho (ejemplo: muestreo en el lugar)

Muestreo probabilístico

- Es el único sobre el que se puede desarrollar una teoría científica que permita predecir la validez de las estimaciones muestrales.
- El muestreo probabilístico es el *gold standar* para la inferencia estadística en las encuestas.
- Declive de los métodos tradicionales de administración de cuestionarios en términos de tasa de respuesta y calidad.
- Los nuevos métodos de administración (web, app, redes sociales,...) ofrecen grandes tamaños de muestra a costes bajos y rápidos.
- Auge en los últimos años de muestreos no probabilísticos (muestreo en bola de nieve, paneles online, encuestas online a voluntarios,...).

Muestreos no probabilísticos

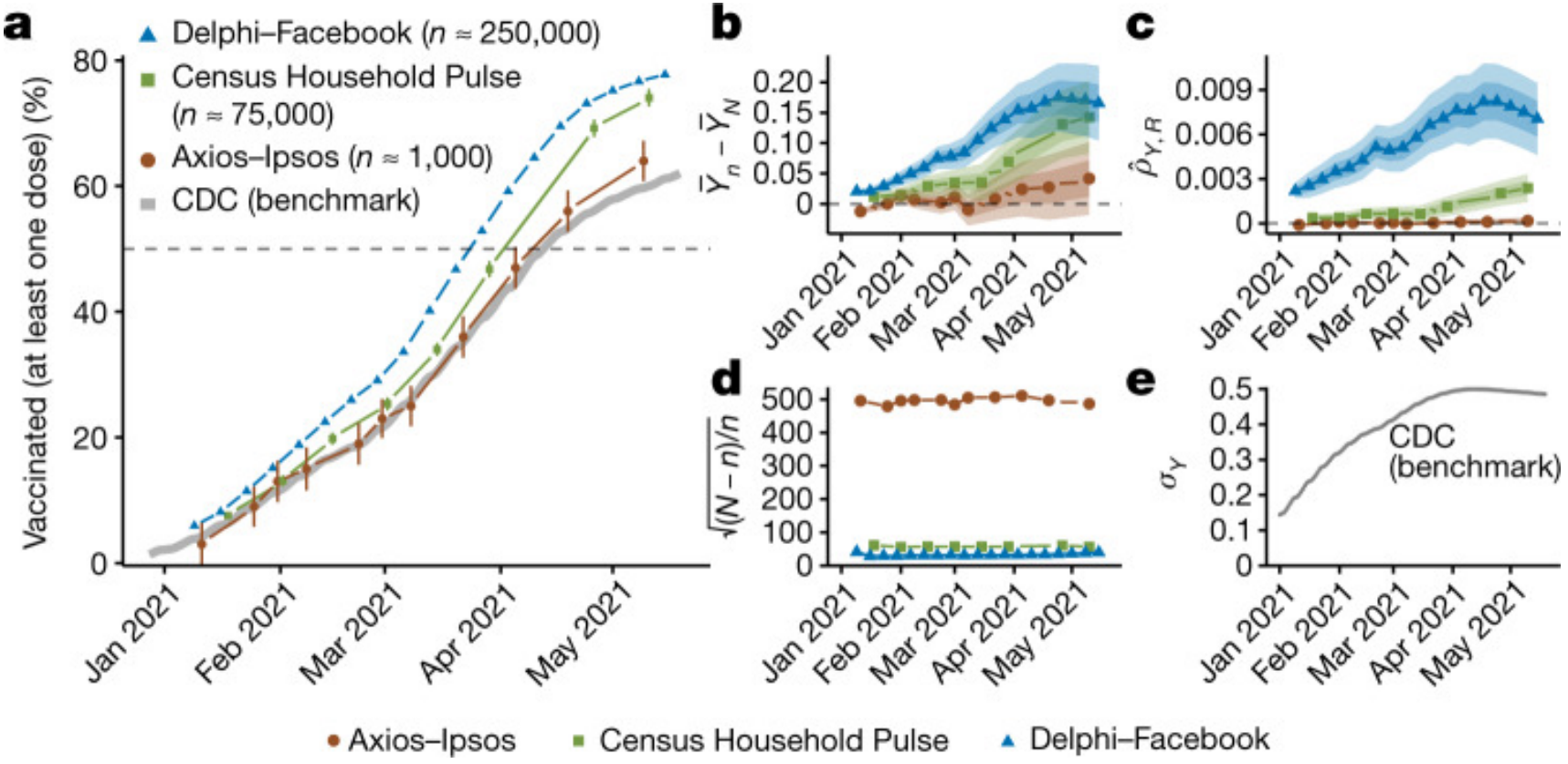
- La selección de las unidades no es aleatoria, no se pueden determinar las probabilidades de participar de cada unidad
- No se pueden aplicar los principios de la inferencia estadística.
- No ofrece ninguna garantía de representatividad de la muestra, con lo que los sesgos potenciales son muchísimos,
- Si las características de la muestra utilizada difieren de las de la población de referencia hay una amenaza en la validez interna y externa de la investigación,
- Resultan útiles en investigaciones exploratorias, donde el interés reside en determinar si existe o no un problema concreto.

Muestreo no probabilístico

Bradley, et al. (2021) Nature 600(7890):695-700. Unrepresentative big surveys significantly overestimated US vaccine uptake

- Paradoja del Big Data: El aumento del tamaño muestral reduce los IC pero aumenta el efecto del sesgo.
- Estimacion de la vacunación de la primera dosis del COVID-19 en adultos estadounidenses del 9 de enero al 19 de mayo de 2021 a partir de dos encuestas: Delphi–Facebook (250.000 respuestas por semana) y Census Household Pulse (75.000 cada dos semanas)
- Delphi-Facebook sobrestimó la vacunación en 17 % Census Household Pulse en 14 % (error estándar 5 %)
- Axios-Ipsos con 1.000 respuestas por semana proporcionó estimaciones mucho más precisas.

Bradley, et al. (2021) Nature 600(7890):695-700.



Definiciones básicas

- $U = \{1, \ldots, i, \ldots, N\}$
- s_r muestra probabilística n_r, π_i, $d_i = 1/\pi_i$
- s_v muestra no probabilística n_v, autoseleccionada de $U_f \in U$.
- y variable de interés
- $\mathbf{x}$ vector de covariables

Muestra probabilística

$$\hat{Y}_R = \sum_{i \in s_r} d_i y_i$$

Insesgado

$$\widehat{V}_p(\hat{Y}_R) = \sum_{i,j \in s_r} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

Muestra no probabilística

$$\hat{Y}_v = N \sum_{i \in s_v} \frac{y_i}{n_v}$$

No es insesgado

No se puede determinar la varianza muestral

Muestra probabilística

$$\hat{Y}_R = \sum_{i \in s_r} d_i y_i$$

Insesgado

$$\widehat{V}_p(\hat{Y}_R) = \sum_{i,j \in s_r} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

Muestra no probabilística

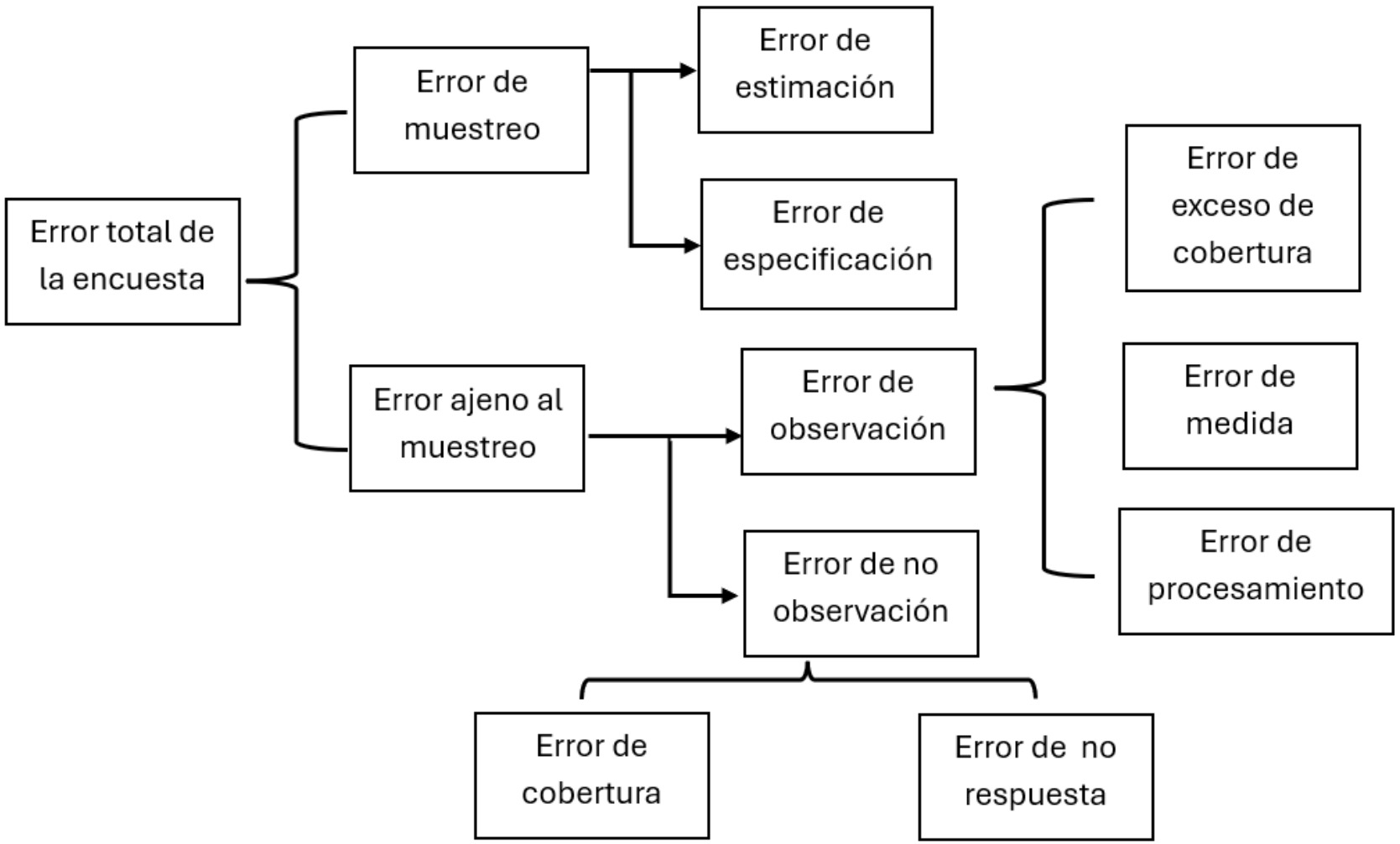
$$\hat{Y}_v = N \sum_{i \in s_v} \frac{y_i}{n_v}$$

No es insesgado

No se puede determinar la varianza muestral

Sesgos

Definición: *Error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unas respuestas frente a otras*



Sesgo de cobertura

Una parte de la población sujeto de estudio no puede ser contactada, y no estará representada en la muestra
Razones: existen hogares donde no existe acceso a internet, los sujetos muestrales no se hayan incluido en el marco muestral

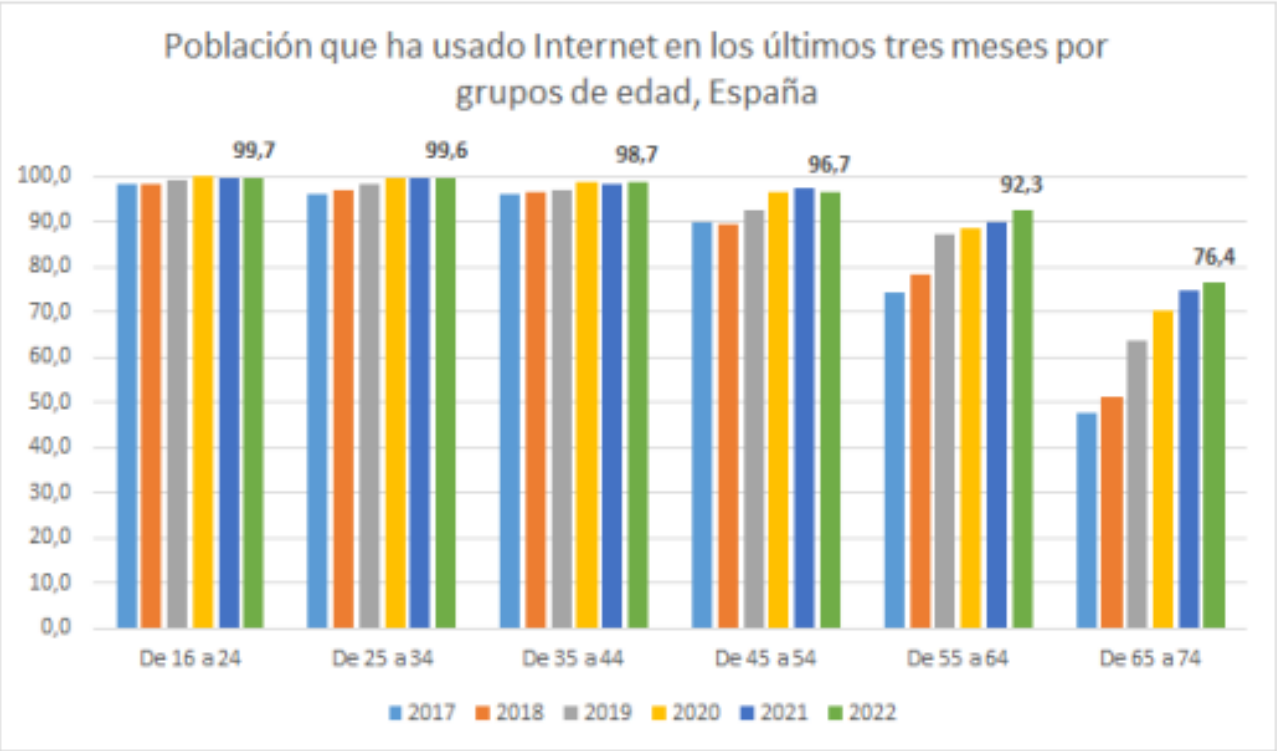


Figura 5.3: Adaptado de 'Población que ha usado Internet en los últimos tres meses por grupos de edad' por Instituto Nacional de Estadística, 2022.

Sesgo de selección

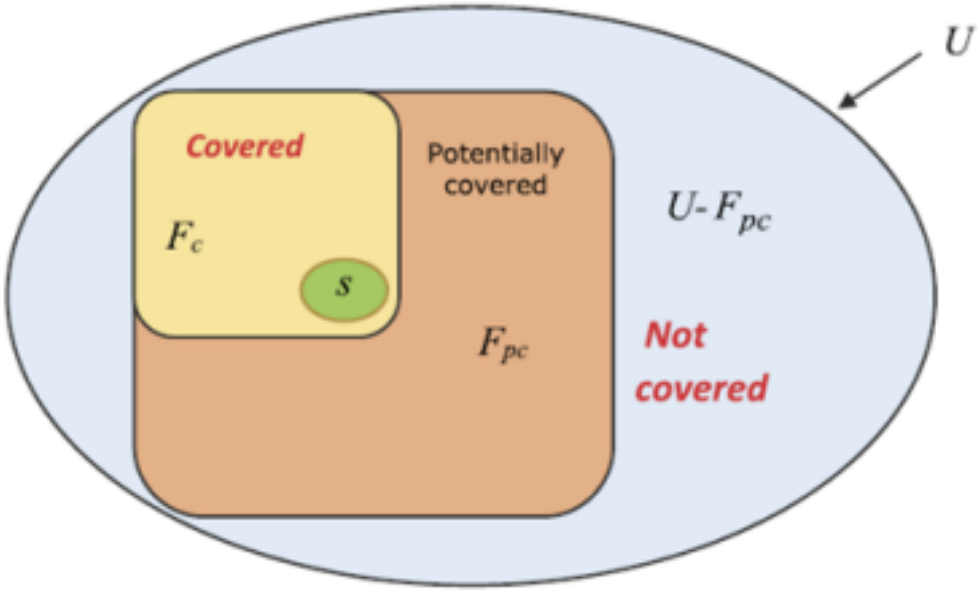


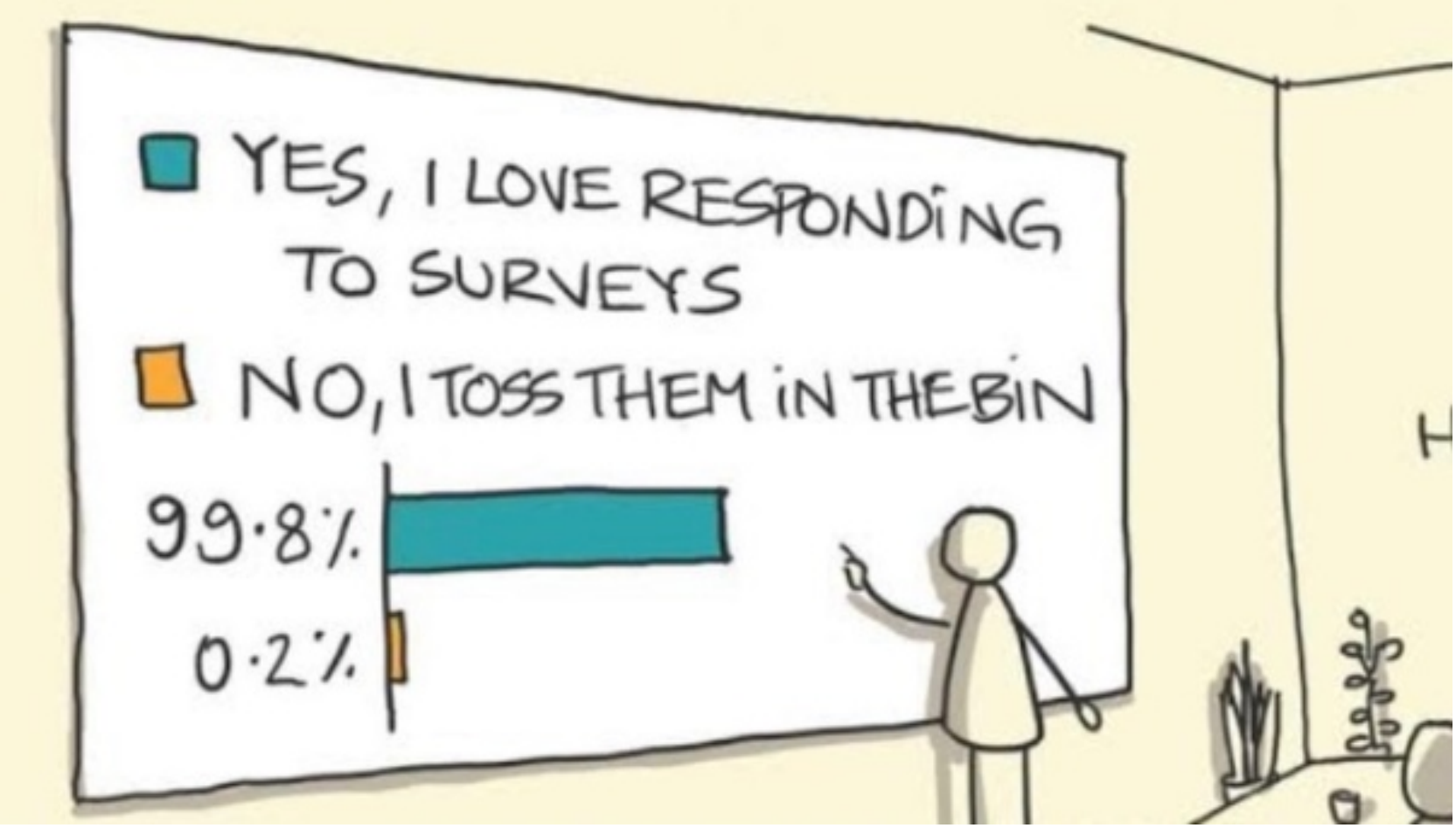
Figura 5.2: Adaptado de Inference for Nonprobability Samples (p.253) por M.Elliott y R.Valliant, Statistical Science. 32, 2, 249-264.

$$B(\overline{y}_v) = E_v(\overline{y}_v - \overline{Y}) = \frac{1}{f_v} E_v\{Cov(I_v, y)\}$$

$$I_{vi} = \begin{cases} 1 & i \in s_v \\ 0 & i \notin s_v \end{cases}, \quad f_v = n_v/N$$

Si $Corr(I_v, y) \neq 0$ habrá sesgo de selección.

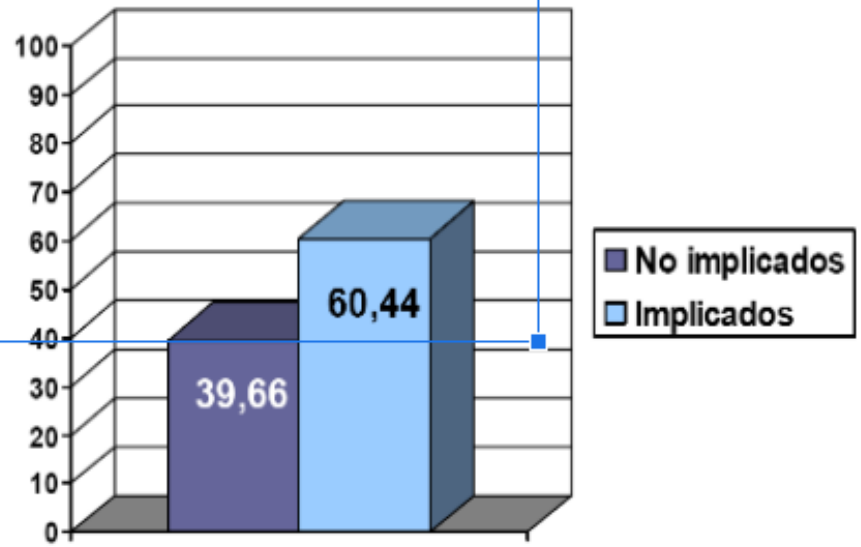
Un buen ejemplo de sesgo de selección



Otro ejemplo de sesgo de selección

FICHA TÉCNICA DE LA ENCUESTA

Diseño y Realización:	
Universo:	Trabajadores de la Universidad de
Tamaño de la muestra:	300 sujetos
Fecha trabajo campo:	20/11/02 a 01/02/03
Tipo de encuesta:	Acceso a cuestionario a través de Internet
Error muestra:	± 2,77
Informante:	Personal docente y personal de administración y servicios de la Universidad de
Supervisión e Informe:	Grupo de Investigación HUM#
Dirección y coordinación:	
Área geográfica:	



Gráfica 1. Porcentaje de trabajadores que confirman el acoso laboral en la Universidad.

Técnicas de reducción de sesgos

Depende de la información auxiliar disponible:

- Totales poblacionales de las variables auxiliares => Calibración
- Variables auxiliares para cada individuo de la población objetivo => Modelos de superpoblación (enfoque predictivo)
- Variables auxiliares de una muestra probabilística s_r =>
  - Ajuste de puntuación de propensión (PSA),
  - Predicción de probabilidad ajustada por propensión (PAPP),
  - Método de Ponderación Kernel (KW),
  - Emparejamiento Estadístico (SM) y
  - Estimación Doblemente Robusta (DR).

Calibración

$X_1, X_2, ...X_p$ conocidos
(totales de variables sociodemográficas obtenidas de censos)

Cuadro: Estructura de datos

	n	probabilidades	covariables	variable de estudio
M. no probabilística	n_v	no	si	si
Población	N	no	datos agregados	no

Calibración

Desarrollada por **Deville and Särndal (1992)**:

$$\hat{Y}_{CAL} = \sum_{i \in s_v} w_i^{\star} y_i$$

$w_i^{\star}$ de forma que:

$$\text{mín} \sum_{i \in s_v} G(w_i^{\star}, w_{vi}) \qquad \text{s.t.} \qquad \sum_{i \in s_v} w_i^{\star} x_i = X,$$

$w_{vi} = N/n_v$.

No se puede eliminar el sesgo de selección (**Ferri and Rueda, 2018**)

Modelos de superpoblación

$x_{1i}, , ...x_{pi}$ conocidos $\forall i = 1, ..., N$

Cuadro: Estructura de datos

	n	probabilidades	covariables	variable de estudio
M. no probabilística	n_v	no	si	si
Población	N	no	si	no

Modelos de superpoblación

Modeliza $y = m[y_i|x_i]$ y obtiene los valores predichos $\hat{y}_i = E_m[y_i|x_i]$

$$\hat{Y}_{MB} = \sum_{i \in s_v} y_i + \sum_{j \in u - s_v} \hat{y}_j$$

$$\hat{Y}_{MA} = \sum_{i \in U} \hat{y}_i + \sum_{j \in s_v} w_{vj}(y_j - \hat{y}_j)$$

$$\hat{Y}_{MC} = \sum_{i \in s_v} w_i^{MC} y_i \qquad \text{s.t.} \qquad \sum_{i \in s_v} w_i^{MC} \hat{y}_i = \sum_{i \in U} \hat{y}_i$$

con $w_{vi} = N/n_v$,

Ajuste de propensidades (PSA)

$x_{1i}, , ...x_{pi}$ conocidos en muestra probabilística

Cuadro: Estructura de datos

	n	π_i	covariables	variable de estudio
Muestra prob.	1	si	si	no
	⋮	⋮	⋮	⋮
s_r	n_r	si	si	no
Muestra no prob.	1	no	si	si
	⋮	⋮	⋮	⋮
s_v	n_v	no	si	si

Ajuste de propensidades (PSA)

Hipótesis

1. I_{vi} e y_i independientes dado $\mathbf{x}$: $\pi_{vi} = Pr(I_{vi} = 1|\mathbf{x}_i) = m(\mathbf{x}, \lambda), \quad i \in U$
2. $\pi_{vi} > 0, \forall i \in U$
3. I_{vi} e I_{vj} independientes dados $\mathbf{x}_i$ e $\mathbf{x}_j$

PSA proporciona $\hat{\pi}_v$ (propensity scores) usando $m(\mathbf{x}, \hat{\lambda})$:

$$\hat{\pi}_{vi} = E_m[I_{vi} = 1|\mathbf{x}_i], \quad i \in s_v \cup s_r$$

$\mathbf{x}$ se observa en s_v y s_r, y m puede ser cualquier modelo ML.

Ajuste de propensidades (PSA)

- MLE de π_{vi} es $m(\hat{\lambda}, \mathbf{x}_i)$
- $w_i^{PSA} = 1/\hat{\pi}_{vi}$ (**Valliant, 2019**)
- INVERSE PROPENSITY WEIGHTING

$$\hat{Y}_{PSA} = \sum_{i \in s_V} y_i/\hat{\pi}_{vi} = \sum_{i \in s_V} y_i w_i^{PSA}$$

Otras alternativas:

$w_i^{PSA} = (1 - \hat{\pi}_{vi})/\hat{\pi}_{vi}$ (**Schonlau and Couper, 2017**);

media de cada estrato (**Valliant and Dever, 2011**),...

Selección de variables

- La eficiencia del PSA depende en gran medida de las covariables utilizadas para la estimación de la propensión (**Lee, 2006**; **Valliant and Dever, 2011**).
- Si se desconoce la relación entre variables, usar algoritmos de selección para obtener un subconjunto óptimo de variables.
- Selección de variables antes de la estimación de las propensidades para eliminar variables redundates o irrelevantes (**Ferri-García and Rueda, 2022**).

Modificaciones a PSA

Propensity-adjusted probability prediction (PAPP) Elliot et al 2020

Versión modificada de PAS que asume s_v y s_r son disjuntas:

$$w_i^{PAPP} = \frac{1}{\hat{\pi}_{ri}} \cdot \frac{1 - \hat{\pi}_{vi}}{\hat{\pi}_{vi}}$$

Kernel Weighting Method (KW) Wang et. al. (2020)

$$w_i^{KW} = \sum_{j \in s_r} d_j k_{ij}$$

$$d(x_i, x_j) = \hat{\pi}_{vi} - \hat{\pi}_{vj}$$

Es menos sensible a las especificaciones erróneas del modelo y evita los pesos extremos.

Statistical Matching (SM)

$$y_i = E_m[y_i|x_i] + e_i, \quad e \sim N(0, \sigma)$$

- **Aproximación predictiva**:

$$\hat{y}_i = E_m[y_i|\mathbf{x}_i, I_{ri}] = M(\mathbf{x}_i) \qquad \text{y} \qquad \hat{Y}_{SM} = \sum_{i \in s_r} d_i \hat{y}_i$$

- **Nearest neighbour imputation**: Para $i \in s_r$ y $\forall j \in s_v$:

$$d(\mathbf{x}_{i(1)}, \mathbf{x}_i) \leq d(\mathbf{x}_i, \mathbf{x}_i) \qquad \text{y} \qquad \hat{Y}_{SM} = \sum_{i \in s_r} d_i y_{i(1)}$$

Estimador doble robusto (DR)

El estimador basado en PSA funciona bien si el modelo de propensidades está bien especificado, y el de SM si el modelo de predicción está bien especificado.

$$\hat{Y}_{DR} = \sum_{i \in s_v} \frac{(y_i - \hat{y}_i)}{\hat{\pi}_{vi}} + \sum_{i \in s_r} d_i \hat{y}_i$$

Estimador consistente si el modelo de PSA o el modelo de predicción se especifican correctamente, no necesariamente ambos.

Estimación de la varianza

Hay tres posibles fuentes de variación.

- d diseño muestral de s_r.
- El modelo m de la propensidad en s_v
- El modelo M para la regresión $y = M(x)$

Cada uno de los diferentes enfoques utilizados para ajustar el sesgo de selección requiere un marco de aleatorización conjunto.

No es sencillo obtener estimadores insesgados.

Técnicas de remuestreo (Jackknife y bootstrap) usadas por **Valliant (2020)** and **Chen et al (2019)** pero no se ha demostrado la consistencia de los estimadores

Asociación para la investigación de los medios de comunicación (AIMC)

Muestra no probabilística

Universo: usuarios de internet mayores de edad en España

16.482 cuestionarios obtenidos mediante anuncios en páginas internet y sitios web colaboradores (más de 200)

3000 entrevistas a un panel online de Dynata.

Muestra probabilística

Tendencias digitales COVID-19, (CIS, 2021).

3014 entrevistas mediante muestreo aleatorio de teléfonos fijos y móviles en estratos poblacionales resultantes de la división en 7 categorías según el número de habitantes de las comunidades autónomas.

Asociación para la investigación de los medios de comunicación (AIMC)

Muestra no probabilística

Universo: usuarios de internet mayores de edad en España

16.482 cuestionarios obtenidos mediante anuncios en páginas internet y sitios web colaboradores (más de 200)

3000 entrevistas a un panel online de Dynata.

Muestra probabilística

Tendencias digitales COVID-19, (CIS, 2021).

3014 entrevistas mediante muestreo aleatorio de teléfonos fijos y móviles en estratos poblacionales resultantes de la división en 7 categorías según el número de habitantes de las comunidades autónomas.

Variables

Covariables

Sexo. Recodificada

Edad

Comunidad Autónoma

Situación Laboral Recodificada

Estudios Recodificada

variables objeto de estudio

¿Cual cree usted que es su grado de conocimiento de la informática general?

Principiante, medio, avanzado o experto.

Variables

Covariables

Sexo. Recodificada

Edad

Comunidad Autónoma

Situación Laboral Recodificada

Estudios Recodificada

variables objeto de estudio

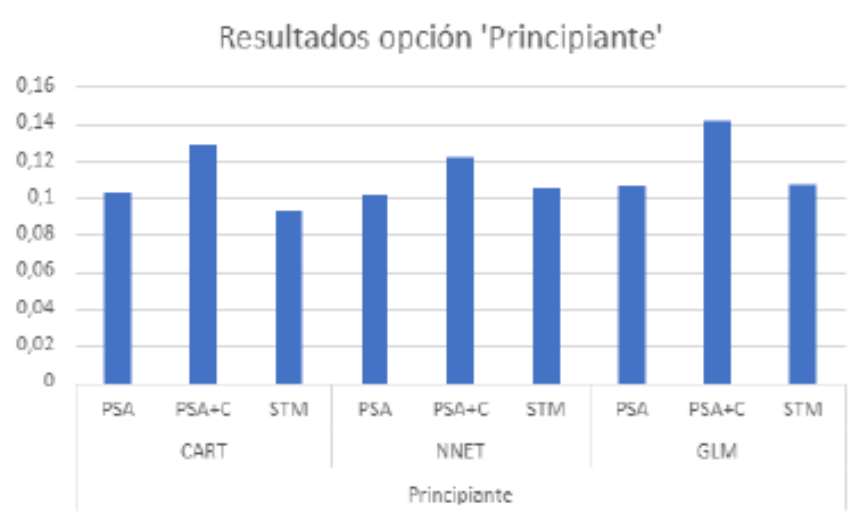
¿Cual cree usted que es su grado de conocimiento de la informática general?

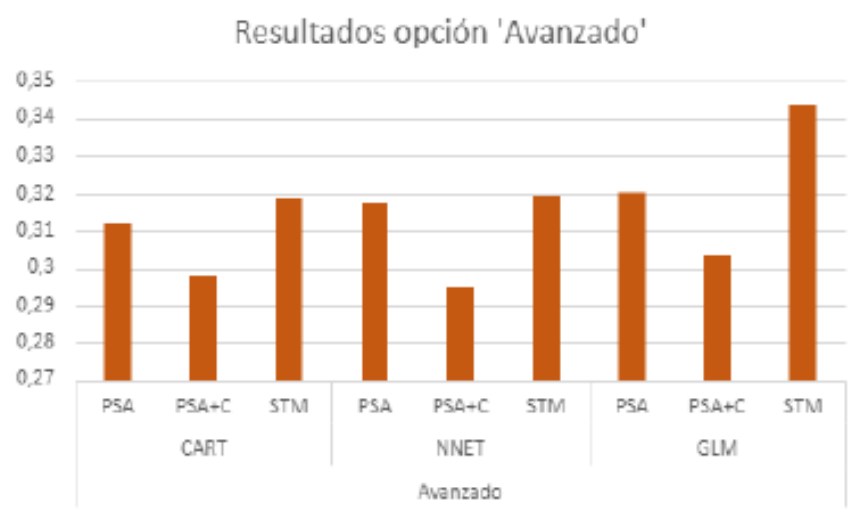
Principiante, medio, avanzado o experto.

Técnicas usadas

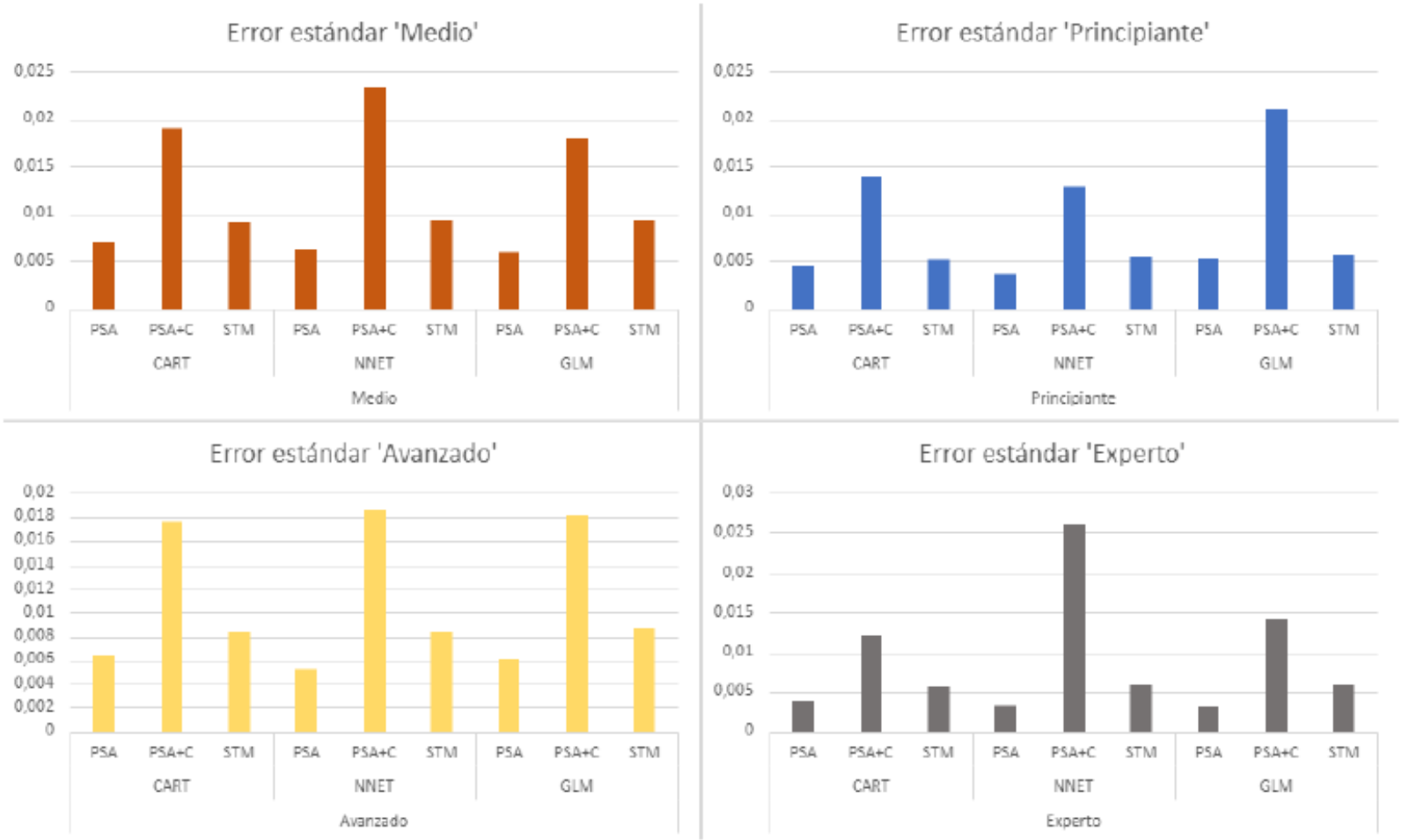
1. PSA con pesos de Valliant
2. PSA y calibración
3. Statistical Matching
4. ML: árboles de regresión y clasificación (CART), redes neuronales (NNET) y regresión lineal (GLM).

Resultados estimaciones





Resultados errores



PSA+ Calibración mayor error. La calibración aumenta la variabilidad de los pesos pero no disminuye el sesgo.

Conocimiento de los factores de riesgo y síntomas del cáncer

Muestra no probabilística

Universo: población española mayor de 18 años.

1029 entrevistas completas mediante Limesurvey difundidas online en listas de distribución de la EASP, colaboradores y diversas universidades.

Muestra probabilística

Oncobarómetro. AACC en colaboración con el CIS.

4769 entrevistas CATI obtenidas mediante muestreo aleatorio estratificado con asignación proporcional por CCAA, con cuotas de sexo y edad en hogares.

Conocimiento de los factores de riesgo y síntomas del cáncer

Muestra no probabilística

Universo: población española mayor de 18 años.

1029 entrevistas completas mediante Limesurvey difundidas online en listas de distribución de la EASP, colaboradores y diversas universidades.

Muestra probabilística

Oncobarómetro. AACC en colaboración con el CIS.

4769 entrevistas CATI obtenidas mediante muestreo aleatorio estratificado con asignación proporcional por CCAA, con cuotas de sexo y edad en hogares.

Variables

Covariables

Sociodemográficas: sexo, intervalo de edad, estado civil, situación laboral, y nivel de estudios completado.

Relacionadas con el estado de salud: estado de salud en los ultimos doce meses, tipo de vida que lleva, riesgo de cancer a lo largo de su vida, un médico le ha dicho alguna vez que sufre cancer y familiar cercano con cancer.

Variables objeto de estudio

1. Variables de síntomas del cáncer: (realmente relacionados con algún tipo de cáncer).
2. Variables de factores de riesgo confirmados del cáncer: (demostradas que incrementan el riesgo de cáncer).
3. Variables de factores de riesgo del cáncer que aún no han sido confirmados. (Algunas falsas creencias).

Variables

Covariables

Sociodemográficas: sexo, intervalo de edad, estado civil, situación laboral, y nivel de estudios completado.

Relacionadas con el estado de salud: estado de salud en los ultimos doce meses, tipo de vida que lleva, riesgo de cancer a lo largo de su vida, un médico le ha dicho alguna vez que sufre cancer y familiar cercano con cancer.

Variables objeto de estudio

1. Variables de síntomas del cáncer: (realmente relacionados con algún tipo de cáncer).

2. Variables de factores de riesgo confirmados del cáncer: (demostradas que incrementan el riesgo de cáncer).

3. Variables de factores de riesgo del cáncer que aún no han sido confirmados. (Algunas falsas creencias).

Técnicas usadas

1. Calibración
2. PSA con pesos de Valliant
3. PSA y calibración

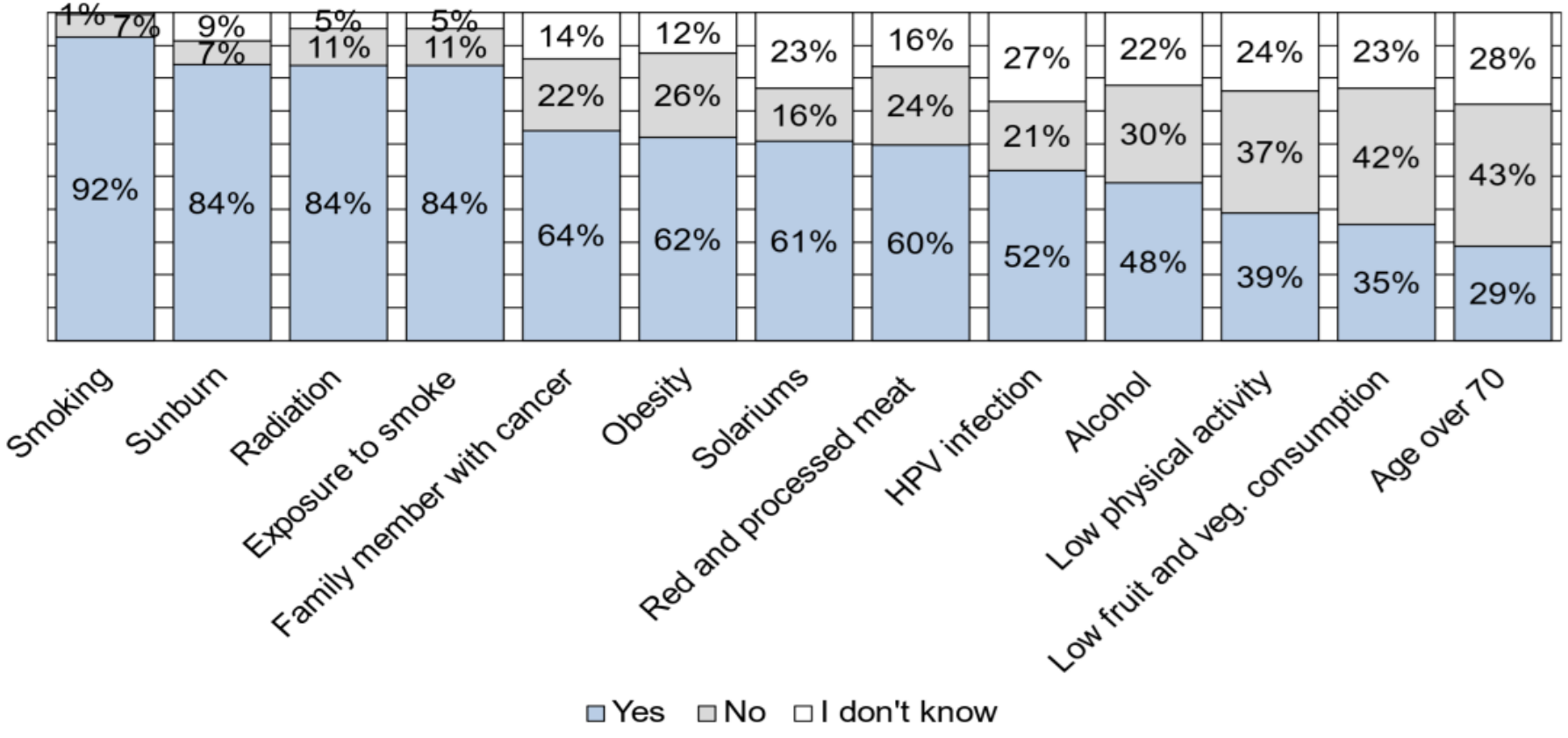
Comparativa muestras

		Probabilística	No prob.	Población
Sexo	Hombre	43.2%	32.7%	48.5%
	Mujer	56.8%	67.3%	51.5%
Edad	18-24	7%	19.9%	8.5%
	25-34	14.4%	11.7%	13.5%
	35-44	18%	15.3%	18.1%
	45-54	18.7%	17.7%	19.5%
	55-64	16.8%	25.5%	16.5%
	65 o más	25.1%	9.9%	23.9%
Estudios	No sabe leer o escribir	0.5%	0%	1.3%
	Primarios incompletos	4.5%	0.7%	4.4%
	Primarios completos	16.4%	2.4%	10.7%
	Primera etapa de educación	16.5%	5.4%	28.7%
	Segunda etapa de educación secundaria, con orientación general	21.1%	17.9%	14.2%
	Segunda etapa de educación secundaria con orientación profesional	17.3%	21.8%	8.4%
	Educación superior	23.7%	51.8%	32.3%

Tabla 3.1: Datos obtenidos a partir del INEbase del año 2021 sobre la variable sexo, intervalos de edad y nivel de estudios

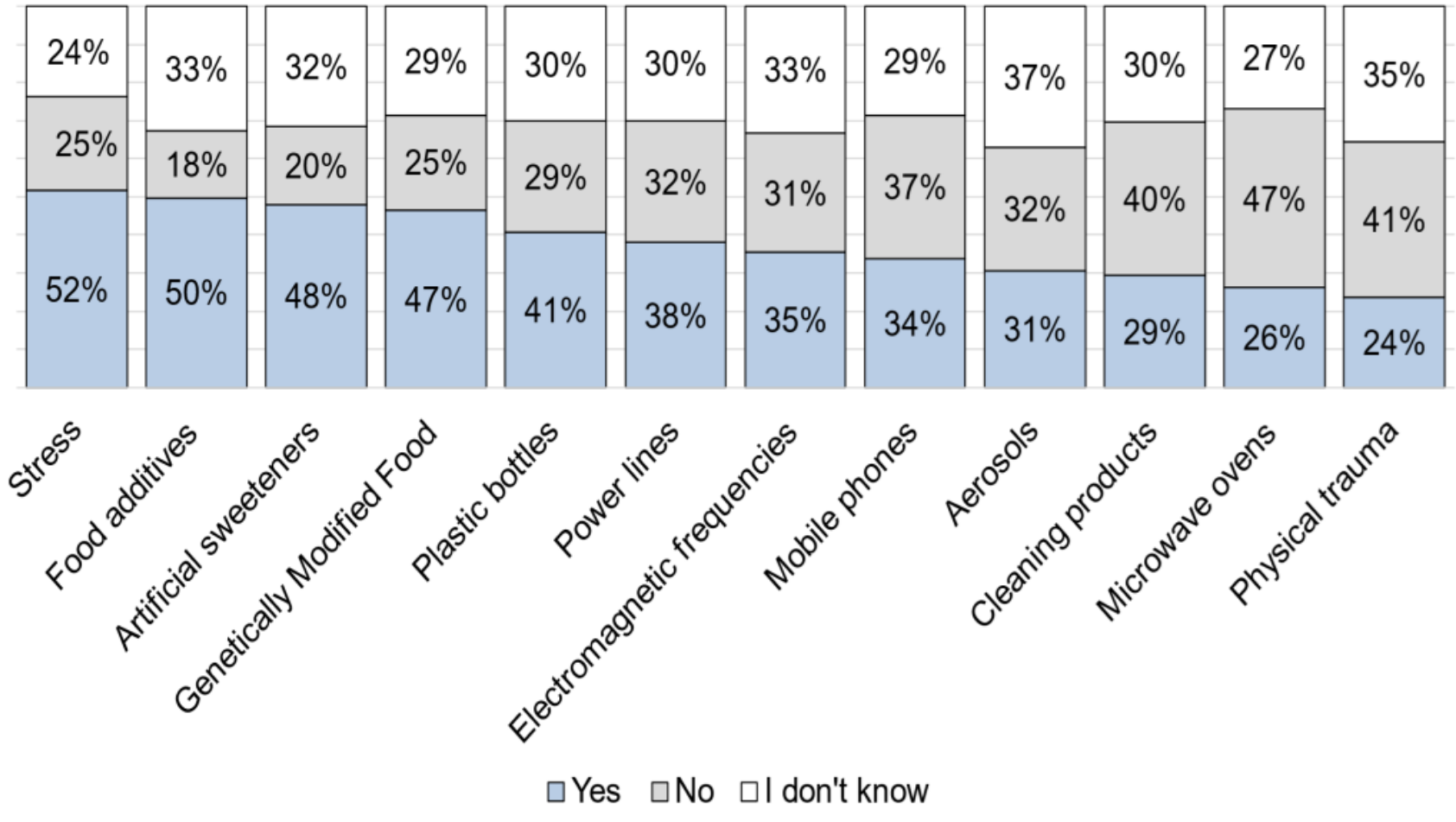
Resultados

Porcentaje de encuestados que reconoce cada factor de riesgo de cáncer



Resultados

Porcentaje de encuestados que respaldan cada causa mítica



Conocimiento de los factores de riesgo y síntomas del cáncer

Resultados de modelos de regresión lineal múltiple: a) el número de factores de riesgo reconocidos y b) el número de causas míticas

			a) # Risk factors		b) # Mythical causes	
	Parameter	N	B	p	B	p
	(Intercept)		8.37	< 0.001	2.54	< 0.001
Eucation level	Low	464	-1.66	< 0.001	0.87	0.01
	Medium	172	-1.35	< 0.001	1.75	< 0.001
	High	393	Ref.		Ref.	
Sex	Women	530	0.03	0.90	0.64	0.01
	Men	499	Ref.		Ref.	
Age	65+	247	-1.63	< 0.001	-0.04	0.92
	55-64	159	-0.32	0.46	0.54	0.21
	45-54	201	-0.35	0.41	0.27	0.48
	35-44	186	0.74	0.08	2.81	< 0.001
	25-34	139	0.59	0.20	1.85	< 0.001
	18-24	88	Ref.		Ref.	

Conocimiento de los factores de riesgo y síntomas del cáncer

Los participantes con educación alta (es decir, nivel universitario) reconocieron más factores de riesgo y respaldaron menos causas míticas.

Las mujeres apoyaron más causas míticas que los hombres.

Entre los diferentes grupos de edad, el reconocimiento de los factores de riesgo fue más bajo entre los de 65 años o más, y más alto entre los de 18 a 24 años. Los adultos jóvenes (de 25 a 44 años) respaldaron el mayor número de causas míticas.

Estimadores para combinar mustras probabilísticas con no probabilísticas

- Encuestas probabilísticas tienen que seguir usándose (**Beaumont y Rao (2021)**)
- Un volumen muy grande de datos puede hacer que la contribución relativa del sesgo al error total sea pequeña y puede reducir el tamaño efectivo de la muestra probabilística.
- Es necesario desarrollar enfoques que permitan incluir datos no probabilísticos a gran escala con datos de muestras probabilísticas.

Cuadro: Estructura de datos

	n	probabilidades de primer orden	x	y
s_r	n_r	si	si	si
s_v	n_v	no	si	si

Métodos para integrar datos

Combinando los estimadores de cada muestra

Elliot and Haviland (2007) $\hat{Y}_{EH} = \alpha \hat{Y}_r + (1-\alpha)\hat{Y}_v$

Rueda et. al.(2022, 2023)

$$\hat{Y}_{CPSA} = \alpha \hat{Y}_r + (1-\alpha)\hat{Y}_{PSA}, \qquad \alpha = \frac{MSE(\hat{Y}_{PSA})}{V(\hat{Y}_r) + MSE(\hat{Y}_{PSA})}$$

Combinando las muestra

$s = s_v \cup s_r$

$$\hat{Y}_{integrado} = \frac{1}{N}\sum_{k\in s} \hat{w}_k y_k$$

$$\hat{w}_k = \frac{n_v}{nr + n_v} w_k^Z \ (Z = PSA, DR, KW, \ ) \ k \in s_v$$

$$\hat{w}_k = \frac{n_r}{nr + n_v} d_k \ \ k \in s_r.$$

Software

Librerías de R

- *NonProbEst* (**Rueda et. al. 2020)**: MS, PSA, Cal, SM
- *nonprobsvy* (**Chrostowski and Beresewicz, 2023)**: PSA, SM, DR
- *KWML* (**Kern et al. 2020**):

Librerías de Python

- *NonProbEst* (**Rueda et. al. 2024**):

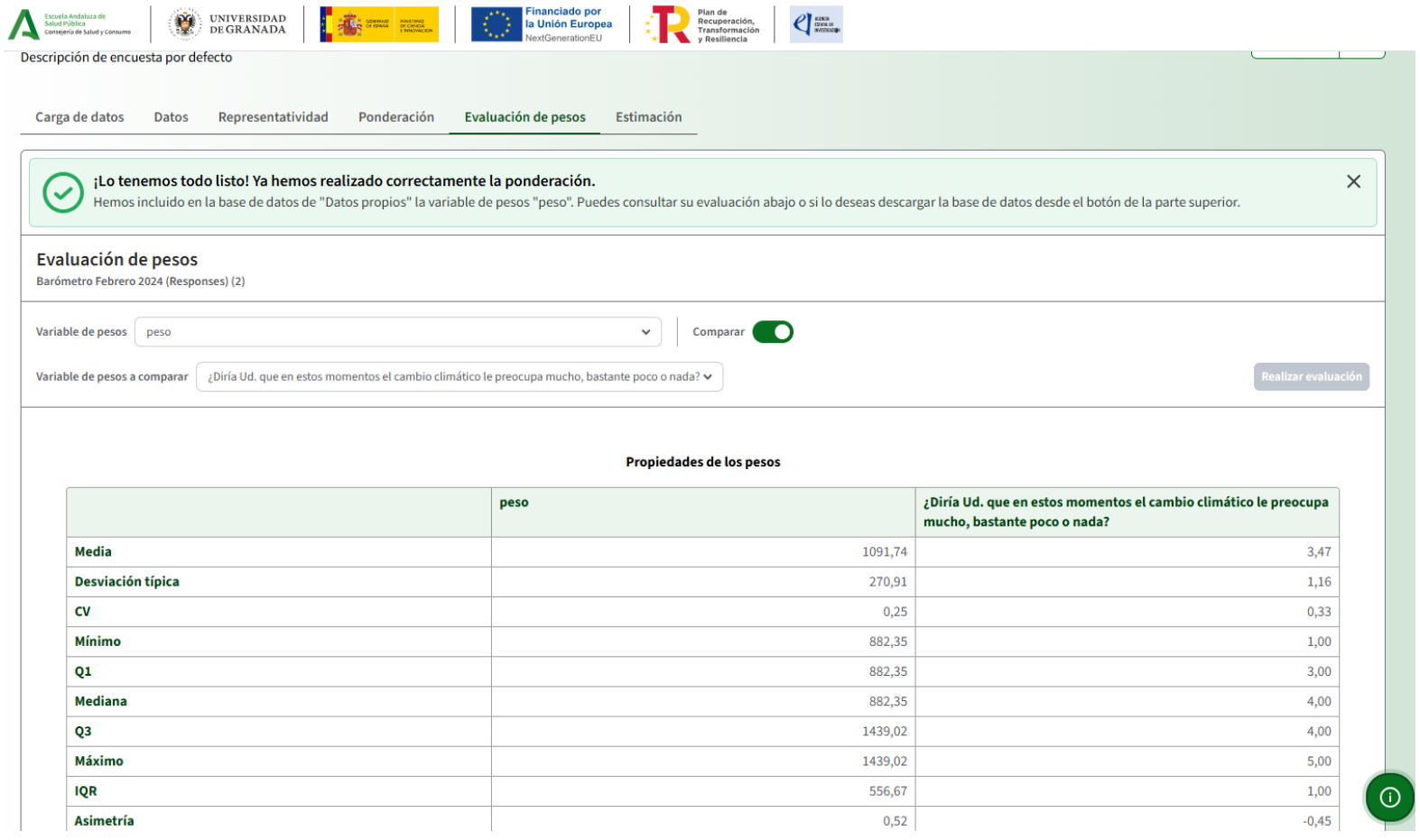
Disponible en Python Package Index (PyPI) `https://pypi.org/project/nonprobest/`.

Todos los estimadores con cualquier modelo ML implementado en Python.

DESARROLLO TECNOLÓGICO DE UNA PLATAFORMA WEB PARA EL AJUSTE DE ENCUESTAS

Objetivo: Diseñar, desarrollar y validar una plataforma para corregir los sesgos y mejorar la validez y precisión de las estimaciones de encuestas.
Encuestas: Uso combinado de datos provenientes de diversas fuentes: (muestras probabilísticas, muestras no probabilísticas y registros administrativos).
Técnicas: Métodos para tratamiento de sesgos de autoselección, cobertura y falta de respuesta, selección de variables, técnicas de aprendizaje automático.
Financiación:
1. Desarrollo tecnológico para mejorar la representatividad de encuestas mediante técnicas de reponderación estadística y de aprendizaje automático: plataforma BETTERSURVEYS. Instituto de Salud Carlos III. 2024-2025
2. Plataforma para ajustar la representatividad de las encuestas realizadas mediante web y redes sociales mediante reponderación con técnicas estadísticas y de ML avanzada. Ministerio de Ciencia e Innovación. Pruebas de Concepto. 2022-2024.

BETTERSURVEYS: https://bettersurveys.org/



Escuela Andaluza de Salud Pública · Consejería de Salud y Consumo · UNIVERSIDAD DE GRANADA · Financiado por la Unión Europea NextGenerationEU · Plan de Recuperación, Transformación y Resiliencia

Descripción de encuesta por defecto

Carga de datos | Datos | Representatividad | Ponderación | **Evaluación de pesos** | Estimación

¡Lo tenemos todo listo! Ya hemos realizado correctamente la ponderación.
Hemos incluido en la base de datos de "Datos propios" la variable de pesos "peso". Puedes consultar su evaluación abajo o si lo deseas descargar la base de datos desde el botón de la parte superior.

Evaluación de pesos

Barómetro Febrero 2024 (Responses) (2)

Variable de pesos: peso — Comparar

Variable de pesos a comparar: ¿Diría Ud. que en estos momentos el cambio climático le preocupa mucho, bastante poco o nada? — Realizar evaluación

Propiedades de los pesos

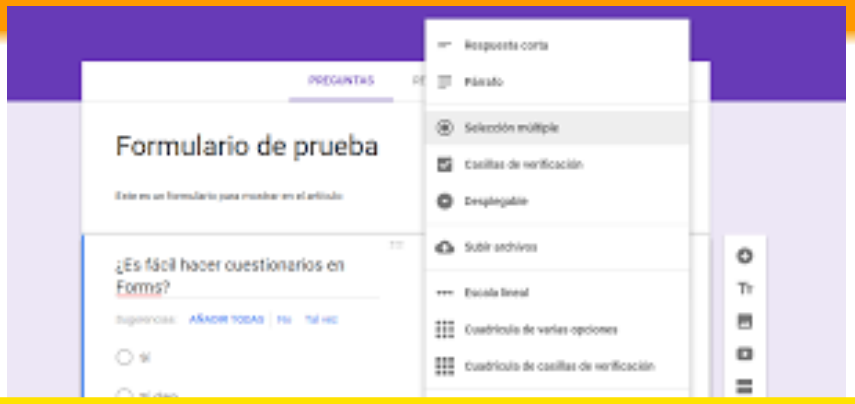
	peso	¿Diría Ud. que en estos momentos el cambio climático le preocupa mucho, bastante poco o nada?
Media	1091,74	3,47
Desviación típica	270,91	1,16
CV	0,25	0,33
Mínimo	882,35	1,00
Q1	882,35	3,00
Mediana	882,35	4,00
Q3	1439,02	4,00
Máximo	1439,02	5,00
IQR	556,67	1,00
Asimetría	0,52	-0,45

Conclusiones

- **Beaumont y Rao (2021)** Las encuestas no probabilísticas, aunque suelen tener tamaños muestrales grandes, pueden presentar importantes problemas de selección y cobertura
- Se han desarrollado nuevas y potentes metodologías para inferir parámetros utilizando datos de muestras no probabilísticas.
- Las encuestas web no probabilísticas pueden resultar útiles. Por ejemplo, cuando el grupo bajo estudio son subpoblaciones pequeñas, complementar una muestra probabilística pequeña con una muestra no probabilística más grande puede mejorar la eficiencia de las estimaciones.
- Existe software libre para poder aplicar estos métodos de forma sencilla.
- **Kalton (2023)** Este es un momento emocionante y desafiante para los metodólogos de encuestas.

References

- **Chen, Y., Li, P., & Wu, C. (2020)**. Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532), 2011-2021.
- **Elliott, M. R., & Valliant, R. (2017)**. Inference for non-probability samples. Statistical Science, 32(2), 249-264.
- **Rueda, M., Ferri-García, R., & Castro, L. (2020)**. The R package NonProbEst for estimation in non-probability surveys. The R Journal, 12(1), 406-418.
- **Ferri-García, R., & Rueda, M.D.M. (2020)**. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PloS one, 15(4), e0231500.
- **Rueda, M. D. M., Ferri-García, R., and Castro-Martın, L. (2022)**. Combining Big Data with probability survey data: a comparison of methodologies for estimation from non-probability surveys. Padua Research Archive-Institutional Repository 711.
- **Valliant, R., & Dever, J. A. (2011)**. Estimating propensity adjustments for volunteer web surveys. Sociological Methods & Research, 40(1), 105-137.



Técnicas para integración de muestras provenientes de diversas fuentes.

Gracias por su atención

Universidad de Granada

(PDC2022-133293-I00 financiada por MCIN/AEI/10.13039/501100011033, Acciones estratégicas de salud (DTS23/00032) e IMAG-María de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033.