



REVISTA LATINOAMERICANA DE
METODOLOGÍA DE
LA INVESTIGACIÓN
SOCIAL

**La problematización
como método en la
indagación social**

Nº29 – AÑO 15

ABRIL - SEPTIEMBRE 2025

PUBLICACIÓN ELECTRÓNICA SEMESTRAL

ISSN 1853-6190

ReLMIS
.com.ar

29

Revista Latinoamericana de Metodología de la Investigación Social



www.relmis.com.ar

Directora:

De Sena, Angélica | CONICET- UNLaM, UBA, CIES, Argentina.

Edición y Coordinación General:

Cervio, Ana Lucía | CONICET-IIGG UBA/CIES, Argentina.

Consejo Académico:

Barriga, Omar | Univ. Concepción, Chile

Cataldi, Silvia | Universidad Sapienza de Roma, Italia

Cohen, Néstor | FSOC-UBA, Argentina †

De Sena, Angélica | CONICET-UNLaM, UBA, CIES, Argentina

Ferreria, Andre | Univ. Fed. Pernambuco, Brasil

Hamlin, Cynthia | Univ. Fed. Pernambuco, Brasil

Henríquez, Guillermo | Univ. Concepción, Chile

Hernández, Marsiela | Univ. Simón Bolívar, Venezuela

Magallanes, Graciela | UNVM/CIES, Argentina

Mejía Navarrete, Julio | Univ. Nac. San Marcos, Perú

Mutzenberg, Remo | Univ. Fed. Pernambuco, Brasil

Padua, Jorge | COLMEX-CONACYT, México

Parra Saiani, Paolo | Univ. de Génova, Italia

Pastor Seller, Enrique | Univ. de Murcia, España

Piovani, Juan | Univ. Nac. La Plata, Argentina

Riella, Alberto | Univ. De la República, Uruguay

Rivera, Manuel | Univ. San Carlos, Guatemala

Salvia, Agustín | FSOC-UBA, Argentina

Sautu, Ruth | IIGG-UBA, Argentina

Scribano, Adrián | CONICET-IIGG-UBA/CIES,
Argentina.

Truda, Giovanna | Univ. de Salerno, Italia

Zacarías, Eladio | Univ. de El Salvador, El Salvador

Colaboradoras:

Chahbenderian, Florencia | IIGG- UBA, CIES.

D'hers, Victoria | CONICET- IIGG - UBA, CIES.

Faracce Macia, Constanza | CIC-UNLaM; UBA.

Paparone, Florencia Isaura | IIGG-UBA.

Diseño de Tapa:

Lucila Salvo

La problematización como método en la indagación social.

Nº 29. Año 15. Abril - Septiembre 2025.

relmis.com.ar

Publicación electrónica semestral



CENTRO DE INVESTIGACIONES Y ESTUDIOS SOCIOLÓGICOS. Arévalo 2357, piso 8, Dpto. 42 (CP 1425). Ciudad Autónoma de Buenos Aires, Argentina | E-mail: correo@relmis.com.ar | e-ISSN 1853-6190

Contenido

. Presentación	
. La problematización como método en la indagación social	4
<i>Por Camilo Agustín Martínez</i>	
. Artículos	
. Delimitación del tamaño muestral en estudios piloto	7
<i>Por Rodrigo Vargas Salomón (México)</i>	
. Distribución de Pareto: un modelo de probabilidad en investigación social	22
<i>Por José Moral de la Rubia (México)</i>	
. Uso de Procesamiento de Lenguaje Natural para procesar respuestas abiertas de una encuesta de opinión pública	51
<i>Por Esteban Martínez Porras, Adrián Ramírez Fernández, Laura Solís Bastos y José André Díaz-González (Costa Rica)</i>	
. ¿Por qué soy de mi clase social? Un análisis cuantitativo y cualitativo de respuestas sobre autopercepción de clase en un estudio por encuesta del Área Metropolitana de Buenos Aires (2016)	68
<i>Por Rodolfo Elbert (Argentina)</i>	
. Experiencia de uso de Chat GPT-3 como asistente de búsqueda en revisiones sistemáticas de literatura	83
<i>Por Claudia Cabrera Borges (Uruguay)</i>	
. Reseñas de publicaciones	
. Repensando la Metodología en Ciencias Sociales: desafíos de la creación y validación de indicadores	100
<i>Por Adriana Elisabet Romiti (Argentina)</i>	

Presentación:

La problematización como método en la indagación social

Camilo Agustín Martínez

A la hora de abordar los problemas de investigación la ciencia utiliza un método: procedimientos formales para la adquisición, comprensión y exposición del conocimiento científico. Este conocer requiere siempre de un sujeto y un objeto, así como de la relación que se establece entre ambos (Hessen, 1979). Aunque en las Ciencias Sociales estos elementos de conocimiento tienen la particularidad de ser una relación de sujeto-sujeto con su campo de estudio, ya que se ocupa de un mundo preinterpretado (Giddens, 1993), hay una compleja articulación entre el mundo social y el conocimiento científico, que pone en tensión dinámica los sentidos que los actores le asignan al mundo social. Al tiempo que este conocimiento, al esparcirse, condiciona a los agentes y a sus prácticas, reproduciendo o transformando a la sociedad (Andrade Carreño, 2015).

Para avanzar en este conocer y cómo conocer, debe existir un motor, es decir, una pasión por las problemáticas que se viven, ya que el investigador, con su posición, influye en el acto de investigar, ya desde la selección de un tema. Tanto el *corpus* teórico como la metodología utilizada para abordar un problema y hallar sus relaciones de causalidad se ubican en un tiempo. Problemas, preguntas e hipótesis de investigación se anclan en un espacio y en un tiempo (pasado-presente-futuro), en donde los métodos, productos histórico-culturales, se aplican sopesando ventajas y limitaciones (Aguilera Hintelholher, 2013).

Sin embargo, a lo largo de la investigación emergen obstáculos epistémicos que desafían al investigador a mantener una actitud reflexiva, crítica y vigilante sobre el *corpus* teórico, la metodología y las objetivaciones efectuadas, tanto en lo referente al sentido común como a la *doxa* académica. Debiendo evitar caer en el uso de prenociones y en abordajes que tomen al mundo como dado, ya que la familiaridad con el universo social es un gran y velado obstáculo epistemológico (Bourdieu, 1999). Logrando así la problematización de las relaciones y la ruptura con la imagen socialmente aceptada del objeto –que forma parte de la investigación en las Ciencias Sociales– en la cual se puede entrever una hermenéutica de las ausencias, que dejan ver, y de los síntomas, que exteriorizan (Scribano, 2009).

Actualmente es sugerente pensar cómo la digitalización de la vida social generó nuevas formas de ser, pensar y sentir tanto *online* como *offline*. En este contexto, no solo proliferaron grandes cantidades de datos, sino también una diversidad de métodos computacionales, que van más allá de los métodos de investigación tradicionales, para abordar las problemáticas que plantea el mundo social (Macanovic, 2022). Algunos de esos métodos son abordados y problematizados por los artículos que componen este nuevo número de ReLMIS.

El recorrido se inicia con el trabajo titulado *Delimitación del tamaño muestral en estudios piloto*, elaborado por Rodrigo Vargas Salomón. El artículo pone en tensión los criterios que subyacen a la hora de definir el tamaño de las muestras en los estudios pilotos, mostrando que éstos son fruto de una decisión razonada y crítica. A partir de sus argumentaciones, el autor afirma que la delimitación muestral en esta clase de estudios no se solventa simplemente en el supuesto de que muestras más grandes producen menos errores y más representatividad, sino que, por el contrario, deben atenderse otros factores de relevancia que participan en la aludida definición.

Seguidamente, en *Distribución de Pareto: un modelo de probabilidad en investigación social*, José Moral de la Rubia se detiene en la exposición de esta distribución, ofreciendo un desarrollo de utilidad para las y los investigadores sociales. En base a la exposición de funciones y la puesta en consideración de dos ejemplos –uno de cálculo de probabilidades y medidas descriptivas, y otro de estimación de parámetros y comprobación del ajuste– el autor argumenta

acerca de las posibilidades que la distribución de Pareto ofrece para el análisis de fenómenos sociales.

En tercer lugar, en el artículo *Uso de Procesamiento de Lenguaje Natural para procesar respuestas abiertas de una encuesta de Opinión Pública*, Esteban Martínez Porras, Adrián Ramírez Fernández, Laura Solís Bastos y José André Díaz-González desarrollan la utilidad del Procesamiento de Lenguaje Natural (PLN) para la investigación en Ciencias Sociales. Con base en datos emergentes de una encuesta sobre percepción sobre grupos migrantes entre la población costarricense, los autores prueban diversos algoritmos e identifican que el *random forest* es el que realiza una mejor clasificación automática de las respuestas, lo que contribuye a disminuir ambigüedades en la clasificación de información no estructurada.

A continuación, en el artículo *¿Por qué soy de mi clase social? Un análisis cuantitativo y cualitativo de respuestas sobre autopercepción de clase en un estudio por encuesta del Área Metropolitana de Buenos Aires (2016)*, Rodolfo Elbert reflexiona sobre las categorías utilizadas para indagar la autopercepción de clase en estudios por encuestas. Luego de efectuar una reflexión crítica en torno a las tradicionales categorías de clase utilizadas en estudios similares, el artículo se detiene en una pregunta abierta incluida en un instrumento administrado en el AMBA, y que es analizada siguiendo una estrategia mixta.

En el quinto artículo, titulado *Experiencia de uso de Chat GPT-3 como asistente de búsqueda en revisiones sistemáticas de literatura*, Claudia Cabrera Borges describe las ventajas y limitaciones de los chatbots GPT como apoyo para las revisiones de literatura en el ámbito de la investigación educativa. Asumiendo un posicionamiento crítico acerca del uso de inteligencia artificial generativa para la elaboración de revisiones sistemáticas, la autora concluye que se trata de una herramienta valiosa, sin embargo, advierte sobre la importancia de articular su uso con un enfoque crítico, considerando, además, aspectos éticos en la investigación educativa.

Cierra este número la reseña bibliográfica elaborada por Adriana Elisabet Romiti sobre el libro *Políticas sociales y microemprendimientos socio-productivos: Indicadores de sustentabilidad*, de Angélica De Sena (2018). En esta reseña se repasan y problematizan diversos desafíos asociados a la creación y validación de indicadores de sustentabilidad de microemprendimientos, ofreciendo una mirada crítica que posibilite dar cuenta del grado de desarrollo, así como de las posibilidades de reproducción de los mismos.

Desde la diversidad teórica, metodológica y epistemológica que ofrecen los artículos reunidos, el número 29 de ReLMIS no solo muestra que las viejas dicotomías cualitativo/cuantitativo deben dejarse de lado para incorporar mayor complejidad a la investigación de fenómenos sociales en el siglo XXI. En articulación con el origen etimológico del término "investigar", orientado a seguir las pistas o huellas, los artículos también ejemplifican, desde diversos registros, que no hay posibilidades de investigar sin problematizar, es decir, no hay oportunidades de producir conocimiento científico sobre el mundo social sin establecer una reflexión crítica y situada sobre los objetos de indagación, y sobre la posición de los investigadores, los métodos y las teorías que intervienen en el acto de conocer.

Referencias

- AGUILERA HINTELHOLHER, R.M. (2013). Identidad y diferenciación entre Método y Metodología. *Estudios Políticos*, 28, pp. 81-103.
- ANDRADE CARREÑO, A. (2015). Los Postulados Fundamentales de la Teoría de la Modernidad Reflexiva de Anthony Giddens. *Acta Sociológica*, 67, pp. 87-110.
- BOURDIEU P., CHAMBOREDON J.C. y PASSERON J.C. (2008). *El oficio del sociólogo. Presupuestos Epistemológicos*. México: Siglo XXI
- GIDDENS, A. (1993). *Las Nuevas Reglas del Método Sociológico*. Buenos Aires: Amarrortu.
- HESSEN, J. (1979). *Teoría del Conocimiento*. Buenos Aires: Losada.
- MACANOVIC, A. (2022). Minería de textos para las ciencias sociales: el estado y el futuro del análisis de textos computacional en sociología". *Social Science Research*, 108, Article Article 102784. <https://doi.org/10.1016/j.ssresearch.2022.102784>
- SCRIBANO, A. (2009). *Estudios sobre Teoría Social Contemporánea: Bhaskar, Bourdieu, Giddens, Habermas y Melucci*. Buenos Aires: CICCUS.

Autor.

Camilo Agustín Martínez

Facultad de Ciencias Sociales, Universidad de Buenos Aires, Argentina.

Licenciado en Sociología (UBA), maestrando en Sociología Económica en el Instituto de Altos Estudios en Ciencias Sociales, Universidad Nacional de San Martín (UNSAM). Docente de la Facultad de Ciencias Sociales (UBA). Integrante del Programa de Estudios sobre Cuerpos, Emociones y Sociedad, del Instituto de Investigaciones Gino Germani (IIGG-UBA).

Email: camilofuegos@hotmail.com

Citado.

MARTÍNEZ, Camilo Agustín (2025). Presentación. La problematización como método en la indagación social. *Revista Latinoamericana de Metodología de la Investigación Social - ReLMIS*, N°29, Año 15, pp. 4-6.



Delimitación del tamaño muestral en estudios piloto

Sample size delimitation in pilot studies

Rodrigo Vargas Salomón

Resumen

Una idea bastante difundida en las investigaciones, principalmente cuantitativas, es que las muestras deben ser muy amplias, asumiendo que, a mayor tamaño, menores posibilidades de error en los datos, y mejor representatividad estadística de los análisis inferenciales. El objetivo de este trabajo es enfatizar que este principio no es aplicable a todas las investigaciones, pues en algunas como los estudios piloto, el tamaño muestral, además de la representatividad estadística, requiere considerar otros criterios que pueden ser relevantes para un estudio en particular. Por ello, se analizan dos reglas generales al momento de determinar el tamaño muestral en los estudios piloto: las que establecen un número fijo de participantes; y las que sugieren delimitar el tamaño muestral a través de una serie de pasos, diseñados con base en las características propias de la investigación a realizar. Al final, ninguna de estas reglas es mejor que la otra, pues el tamaño muestral deberá ser una decisión ampliamente razonada y crítica por parte del o lo(a)s investigador(a)s, en relación con las características y fines de su investigación.

Palabras clave: pruebas piloto; estudios piloto; tamaño muestral; muestreo plano; muestreo escalonado.

Abstract

An idea that is quite widespread in research, mainly quantitative, is that the samples must be very large, assuming that the larger the size lower the chances of error in the data, and a better statistical representativeness of the inferential analyses. The objective of this work is to emphasize that this principle does not apply to all research since in some cases such as pilot studies, the sample size, in addition to statistical representativeness, requires considering other criteria that may be relevant for a particular study. Therefore, two general rules are analyzed when determining the sample size in pilot studies: those that establish a fixed number of participants; and those that suggest delimiting the sample size through a series of steps, designed based on the characteristics of the research to be carried out. After all, none of these rules is better than the other, because the size of a sample must be a widely reasoned and critical decision made by the researcher(s), related to the characteristics and purposes of their research.

Keywords: pretests; pilot studies; sample size; flat rules; stepped rules.

Introducción

Una perspectiva clásica en investigación, aunque no es la única, es que el muestreo debe utilizarse cuando se va a someter hipótesis a contrastación empírica, la cual se puede probar en un número reducido de datos relevantes para la propia hipótesis, elegidos sin tendencia ni criterio a partir de una infinitud potencial de datos (Bunge, 1983), que conformen la totalidad del estudio. No obstante, delimitar el tamaño de una muestra para una investigación no es tarea sencilla, y es probablemente uno de los primeros grandes problemas a los que se enfrentan lo(a)s investigador(a)s al desarrollar un proyecto, especialmente los nóveles, quienes al tener que definir cuántas personas o casos deben incluir en su investigación para que ésta sea representativa del total de la población, suelen encontrarse con una serie de fórmulas y procedimientos complejos para delimitar el tamaño muestral. También suelen toparse con datos que no siempre tienen al alcance o que desconocen cómo calcularlos, como el de población finita, los errores alfa (tipo I) y beta (tipo II), el margen de error absoluto permitido y los intervalos de confianza esperados, las puntuaciones z , la probabilidad de ocurrencia, el poder estadístico (potencia), el tamaño del efecto estandarizado, entre otros.

No obstante, la dificultad que pueda representar este cálculo, se convierte en una actividad primordial para la mayoría de las investigaciones, que pretenden utilizar estadísticas paramétricas, principalmente por los motivos que se resumen en la siguiente cita:

El tamaño de muestra permite a los investigadores saber cuántos individuos son necesarios estudiar, para poder estimar un parámetro determinado con el grado de confianza deseado, o el número necesario para poder detectar una determinada diferencia entre los grupos de estudio, suponiendo que existiese realmente. El cálculo del tamaño de la muestra es una función matemática que expresa la relación entre las variables, cantidad de participantes y poder estadístico (García-García *et al.*, 2013: 217).

Sin embargo, las opciones para calcular el tamaño muestral adecuado para una investigación son muy diversas. Actualmente existen incluso en internet calculadoras de tamaños de muestra gratuitas, que utilizan precisamente fórmulas estadísticas ya establecidas para tal efecto, principalmente la denominada “fórmula de Cochran”¹, que con algunas variantes, es la más utilizada para el cálculo de tamaños de muestras proporcionales y en términos porcentuales. La diferencia es que estas calculadoras en línea realizan los cálculos de manera automática, y no de forma manual como tradicionalmente se hacía, además de que facilitan la realización de cálculos previos inherentes a las fórmulas, como el valor de la puntuación z o probabilidad de ocurrencia, con lo que minimizan el número de datos requeridos para calcular el tamaño de una muestra.

Un ejemplo de este tipo de calculadoras es la del sitio web QuestionPro (2023), donde con tres sencillos pasos se puede calcular el tamaño muestral “de forma rápida y efectiva”:

1. Señalar el nivel de confianza (que puede ser 95% o 99%, aunque el sitio sugiere el 95%);
2. Seleccionar el margen de error (donde se sugiere 5% o menos); y
3. Asignar el total de la población (el cual dependerá del dato de población finita que se tenga para la investigación).

Ingresando esta información, lo único que resta hacer es dar *clic* en una casilla denominada “calcular muestra”, y se obtiene un número entero como resultado, que corresponde al tamaño muestral sugerido. Otros sitios similares son el de SurveyMonkey (2023) y el de Qualtrics (2023), en los cuales se piden exactamente los mismos tres datos para calcular, con la misma facilidad, el tamaño de una muestra. Cabe resaltar que el resto de los valores requeridos para aplicar la fórmula de Cochran, estos sitios los tienen predeterminados, por lo que no se solicitan a los usuarios.

¹ Existen varias versiones de esta fórmula, pero en general utiliza los mismos parámetros en su planteamiento cuando se trata de poblaciones finitas, y aunque se le conoce así porque está basada en las aportaciones de William G. Cochran, lo cierto es que en su obra principal *Sampling Techniques*, publicada originalmente en 1953, no establece directamente la fórmula que hoy se utiliza ampliamente, sino que realiza diversas propuestas dependiendo del tipo de estudio o de muestreo que se pretende realizar (Cochran, 1977).

Además, existen aplicaciones como GRANMO, diseñada por Marrugat *et al.* (1998) hace más de 25 años, precisamente considerando lo complicado que es aplicar manualmente algoritmos para el cálculo de muestras. Desde sus inicios se trató de un software de libre difusión (*shareware*), que evolucionó hasta contar actualmente con una calculadora muy avanzada de uso gratuito en línea, que permite calcular tamaños muestrales según diversos tipos o diseños de investigación (medidas apareadas, equivalencias, con dos proporciones independientes, estudios de casos y controles, estudios de cohortes, entre otros). Además de los datos básicos ya señalados, GRANMO permite otros como las desviaciones estándar, los riesgos de error alfa y beta, proporciones de grupos, estimaciones poblacionales, etc., según el diseño elegido.

Otro ejemplo similar destacable es Epi Info, desarrollada a finales del pasado siglo por el Centro para el Control y Prevención de Enfermedades de Atlanta, Estados Unidos, y distribuida desde entonces de manera gratuita. Esta herramienta también evolucionó de un programa en MS-DOS, hasta contar actualmente con un software para computadoras y una aplicación para teléfonos inteligentes –ambos de uso libre– que también permiten calcular de manera automatizada tamaños de muestra con distintos diseños e indicadores estadísticos según se requiera (CDC, 2022).

No obstante la existencia de estas herramientas en línea, u otras como los softwares estadísticos tipo SPSS, ECHANT, PASS, STATS, EPIDAT, etc., de los que se puede hacer uso en la actualidad, y que son muy efectivos porque se apegan a los procedimientos tradicionales de cálculos de tamaño muestral, lo cierto es que tanto en los sitios en línea como en los softwares, y actualmente incluso con el uso de sistemas de Inteligencia Artificial como el ChatGPT, Claude o Gemini, no se consideran características particulares que pueden tener algunos tipos o situaciones de investigación, en los cuales no es tan simple definir cuál es tamaño de muestra más conveniente.

En algunos casos, apegarse a una fórmula no es la mejor opción, menos aún si se toman en cuenta aspectos como los que atinadamente advierten García-García *et al.* (2013), para quienes calcular el tamaño de una muestra implica mucho más que realizar una simple operación aritmética que proporcione un valor, pues se trata más bien de una función matemática, donde cambiar el valor de una variable de la ecuación, afecta a otros valores, teniendo un efecto no sólo sobre el número obtenido como tamaño muestral, sino también sobre la potencia² de la misma, y por tanto, sobre la calidad de los datos que se obtengan.

Un ejemplo de investigaciones donde el tamaño muestral no es tan simple de definir mediante una fórmula matemática, son las investigaciones cualitativas, donde más que buscar la representatividad en función de un valor paramétrico e inferencial, lo que se busca es la representatividad teórica (no paramétrica), en función del tema o fenómeno de estudio. Esto hace que no sea tan relevante a cuántas personas incluir, sino más bien a quiénes, de modo que sean lo(a)s más idóneo(a)s para la investigación, generando un sesgo deliberado en la selección de informantes, al que se le suele denominar como muestreo teórico, intencional, crítico o por juicio. No obstante, si bien algunos estudios cualitativos podrían ser útiles como prueba piloto en estudios cuantitativos e híbridos, esta alternativa no es abordada en este artículo, dada la complejidad y espacio que requiere su discusión.

En el campo de las investigaciones cuantitativas también suele haber estudios donde tampoco es tan simple establecer un tamaño muestral idóneo. Al respecto, Kerlinger & Lee (2002) señalan que “un [sic] regla dura pero eficaz, que se enseña a estudiantes principiantes de investigación es: utilizar una muestra tan grande como sea posible” (p. 157). Según estos autores, esto no es sólo un asunto de que los números grandes sean mejores en sí mismos, sino que las muestras grandes permiten que realmente funcione el principio de aleatorización, para que sean realmente representativas de la población estudiada ya que, en las muestras pequeñas, la posibilidad de generar sesgos en la selección de casos o informantes es mayor, y con ello, también la posibilidad de errores en los análisis inferenciales. Por tanto, otro principio básico al respecto es

² Potencia o poder estadístico, es la probabilidad de que una prueba aplicada, identifique de forma correcta un resultado o efecto de interés para el estudio. Se vincula de manera directa con la sensibilidad de los instrumentos, la efectividad de los resultados, y a su vez, con la probabilidad de no cometer un error de tipo II o beta.

que, “a menor tamaño de la muestra, mayor será el error, y que, a mayor tamaño de la muestra, menor será el error resultante” (Kerlinger & Lee, 2002: 157).

No obstante, estos mismos autores señalan que esta regla de tener muestras grandes, no es necesariamente apropiada para todas las investigaciones, pues mientras en algunos casos puede ser muy simple obtenerlas, en otros no lo es, de modo que la determinación estadística del tamaño muestral, debe estimarse en relación con la situación particular de la investigación a realizar (Kerlinger & Lee, 2002). Además, las muestras grandes tampoco son sinónimo de ausencias de errores o sesgos en los resultados, en algunas investigaciones, “una muestra demasiado grande hará que una diferencia muy pequeña resulte estadísticamente significativa, sin tener necesariamente una significancia práctica” (Kerlinger & Lee, 2002: 266), aspecto en el que también coinciden Babbie (1990) y Stevens (2009). De tal modo que, aun usando fórmulas o procedimientos para simplificar el cálculo de un tamaño de muestra, esta tarea rara vez es algo sencillo.

Los estudios piloto

Un tipo de estudios que suelen presentar problemas al delimitar el tamaño de una muestra, son los estudios piloto, los cuales suelen llevarse a cabo para “probar en menor escala aspectos logísticos de la ejecución del estudio, lo que evitará cometer errores en los estudios posteriores y de mayor envergadura” (Díaz-Muñoz, 2020: 100). No obstante, como bien señala este mismo autor, este tipo de estudios, a pesar de ser publicables en algunas revistas, por lo general reciben poca o nula importancia cuando se enseña investigación, lo que genera prácticas inadecuadas en su realización y diseño, a pesar de que pueden ser muy útiles para evitar la aparición de defectos que podrían ser graves y costosos en términos de recursos, si es que se replican en el estudio final para el cual se diseñó el piloto.

Es tan poca la relevancia que se le suele dar a los estudios piloto, que por lo general se les considera sólo como una fase del apartado metodológico de los proyectos de investigación, con la única y exclusiva finalidad de probar instrumentos o procedimientos que no cuentan con adecuados criterios de confiabilidad y validez, o que fueron diseñados específicamente para el estudio final que se pretende realizar, usándolos erróneamente como sinónimo de las pruebas piloto o pretest. Este poco valor otorgado a los estudios piloto, puede observarse por ejemplo en los libros o manuales generales de metodología de la investigación que se utilizan con mayor frecuencia en México, y probablemente también en otros países de Latinoamérica, donde ni siquiera suele hablarse de estudios piloto, sino únicamente de pruebas piloto.

Por ejemplo, para Hernández *et al.* (2014) y Hernández y Mendoza (2018), la prueba piloto es sólo una fase para construir instrumentos de medición, en la cual, se administra el instrumento a una muestra pequeña [aunque no se señala qué tan pequeña], y a partir de los hallazgos se puede probar la confiabilidad y validez iniciales de los instrumentos, así como otros aspectos como sus instrucciones, condiciones de aplicación y procedimientos involucrados. Para McMillan y Schumacher (2005), las pruebas piloto también son necesarias para identificar cambios en los instrumentos y sus instrucciones, pero además pueden servir para proporcionar una idea de los patrones de respuesta más probables, de los tiempos de aplicación, así como de la claridad y el formato de aplicación. Aunque para esto consideran muy útil que se cuente con al menos diez personas dispuestas a completar estas pruebas y proporcionar sugerencias para mejorarlas, y que se cuente no con una, sino con dos pruebas piloto: una inicial, y otra más formal con una muestra mayor.

Dieterich (2021) también señala algo similar sobre la relevancia de las pruebas piloto (las cuales el autor denomina erróneamente como estudios piloto), para recabar información previa que ayude a planificar adecuadamente una encuesta. Aunque este autor tampoco señala un número específico de personas a incluir, y su recomendación va más bien en un sentido cualitativo, que implica incluir a personas que se consideren como representativas de la población a estudiar, o que puedan proporcionar datos o pautas relevantes que ayuden a mejorar el diseño de los instrumentos definitivos. Kerlinger & Lee (2002) abordan incluso menos el tema, pues no lo incluyen en lo que exponen sobre muestreo y aleatoriedad, y únicamente se refieren a las pruebas

piloto de manera muy general cuando hablan de las investigaciones por encuesta, señalando que una vez que se han construido los instrumentos a utilizar, “se prueban con antelación en una muestra pequeña representativa del universo” (Kerlinger & Lee, 2002: 548), para luego revisarlos y darles la forma final.

Como puede advertirse en los ejemplos señalados, existen diversas posturas sobre la relevancia de los estudios piloto y la forma en que deben desarrollarse, aunque se relacionan especialmente con poner a prueba instrumentos o procedimientos que se incluirán en una investigación más amplia. No obstante, sobre lo cual no hay realmente un consenso, es sobre la determinación de su tamaño muestral, o bien, la delimitación de cuántos casos, personas o sujetos de estudio se deben incluir en un pilotaje.

Al respecto, ya se ha señalado que sí existen diversas fórmulas y herramientas para dicha tarea, lo curioso es que, en muchas de estas fórmulas y herramientas, se señala la importancia de definir el error de estimación o la variabilidad esperada de los datos, no de forma arbitraria, sino de estudios previos similares, o bien, a partir de los datos obtenidos mediante una prueba piloto; sin aclarar cómo se obtendría el tamaño de la muestra para llevar a cabo ese pilotaje. En el fondo parece generarse una especie de error recursivo, donde el cálculo del tamaño de una muestra requiere un dato de una prueba o un estudio piloto, en la que a su vez hay que calcular un tamaño de muestra. Además, los criterios también suelen diferir dependiendo del campo disciplinar desde el cual se realicen las investigaciones.

Tamaño de muestra para estudios piloto en la investigación médica

Lewis *et al.* (2021) señalan que en las investigaciones que incluyen Pruebas o Ensayos Controlados Aleatorizados (RCTs, por sus siglas en inglés), que son experimentos científicos usados principalmente en el área médica, definir el tamaño de una muestra para los estudios piloto es uno de los aspectos que aún no quedan claros, pues las recomendaciones al respecto varían desde 10-12, hasta 60-75 personas por grupo. Esto depende sobre todo del objetivo principal de la investigación, o de criterios como la viabilidad, un tipo de parámetro clínico, la tasa de eventos esperados, el tamaño del efecto estandarizado³, si los datos se requieren para validar una escala, o incluso si se trata de un estudio piloto interno o externo al ensayo principal.

Whitehead *et al.* (2016), al respecto de la estimación del tamaño de la muestra en estudios piloto, que son por lo general parte de todo protocolo de investigación médica o de estudios clínicos, señalan que existen dos reglas generales que suelen utilizarse: las de tamaño de muestra plano o fijo (*flat rules*), y las de tamaño de muestra escalonada o establecida mediante una serie de pasos (*stepped rules*). De estas dos, lo más simple, según estos autores, es recurrir a una regla general donde el investigador tiene que elegir un tamaño de muestra para su estudio piloto, de entre las propuestas ya existentes al respecto, como si se tratara de un número fijo [*flat rule*].

Sobre este punto, existen diversas propuestas (todas ellas citadas por Whitehead *et al.*, 2016), como la de Browne (1995), quien sugiere al menos 30 sujetos para poder estimar un parámetro; la de Julious (2005), quien propone un mínimo de 12 sujetos por cada tratamiento empleado; la de Teare *et al.* (2014), quienes recomiendan mínimo 70 sujetos para lograr reducir la imprecisión en la estimación de la desviación estándar; la de Kieser & Wassmer (1996), para quienes 20 a 40 sujetos serían adecuados en estudios cuya muestra final incluya entre 80 a 250 sujetos; y la de Sim & Lewis (2011), quienes propusieron que 55 o más sujetos minimizaría el tamaño de efecto en poblaciones pequeñas a medianas.

No obstante, estas reglas planas tienen sus limitaciones, ya que suelen aplicarse de manera independiente al tamaño de la población final de la prueba que se está diseñando, por lo que el riesgo en la simplicidad de este enfoque es generar una muestra mayor [o no correspondiente] que el tamaño de muestra final requerido para estudios grandes o pequeños

³ El tamaño del efecto estandarizado, es una medida estadística referida a la magnitud o fuerza que tienen los resultados obtenidos de un instrumento o una intervención experimental, muy útil para la estadística inferencial, pues permite medir la fuerza de relación existente, por ejemplo, entre dos o más variables o entre distintos grupos de estudio.

(Whitehead et al., 2016). Además, existen evidentes diferencias entre los diversos estimados, los cuales no sólo difieren en el número plano que establecen como idóneo, sino también en los planteamientos en los que se basan para su establecimiento, que en algunos casos el número estimado parece haberse calculado de manera un tanto arbitraria.

La segunda regla a la que se refieren Whitehead et al. (2016), sobre la estimación del tamaño de la muestra en estudios piloto, es la de tamaño de muestra escalonada [*stepped rules*] con posibilidades de mejorarse, la cual consideran una mejor opción que la elección de un número plano. Para este tipo de muestra, es recomendable tomar en consideración al menos un estimado sobre la probabilidad de que el tamaño del efecto estandarizado que se espera en el estudio principal, pueda ser pequeño, mediano o grande dentro de un rango preestablecido. Para ello, estos autores retoman la propuesta de los estándares de Cohen, que varían dependiendo de si el tamaño del efecto estandarizado para el estudio principal, se estima como pequeño, mediano o grande (a la cual Whitehead et al. agregan la categoría de extra pequeño), con niveles de potencia estadística ya sea del 80% o 90%. A partir de estos parámetros que pueden estimarse al menos como probables, sería más factible determinar un tamaño de muestra para un estudio piloto con base en lo establecido en la Tabla 1.

Tabla 1. Estimación de reglas generales escalonadas (*stepped rules*) para calcular el tamaño muestral de una prueba piloto, requerido por cada tratamiento o grupo que se utilizará en el estudio principal

Diferencia estandarizada	Prueba principal con 80% de potencia esperada	Prueba principal con 90% de potencia esperada
Extra pequeña ($\delta < 0.1$)	50	75
Pequeña ($0.1 = \delta < 0.3$)	20	25
Mediana ($0.3 = \delta < 0.7$)	10	15
Grande ($\delta = 0.7$)	10	10

Tabla retomada de Whitehead et al (2016, p. 1071). Traducción propia.
El símbolo δ se refiere al tamaño del efecto estandarizado según la clasificación de Cohen.

Este tipo de cálculos son especialmente relevantes en los estudios clínicos y las investigaciones médicas, que es precisamente el foco general de discusión del trabajo de Whitehead et al. (2016), y en los cuales también son ampliamente recomendados los estudios experimentales y cuasi experimentales con al menos un grupo control, ya que el tamaño del efecto estandarizado, permite medir el efecto de una intervención o tratamiento a partir de la diferencia de las medias de los grupos de estudio. Este valor estadístico se puede calcular mediante pruebas como la “d” de Cohen o la “g” de Hegdes, y de manera general permitiría estimar si un tratamiento, medicamento o cualquier tipo de intervención, es o no efectiva, a partir de las medias estadísticas de los grupos estudiados.

Tamaño de muestra para estudios piloto en ciencias sociales

De manera similar a los tipos de muestreo sugeridos en las investigaciones médicas, en las ciencias sociales, donde los estudios piloto suelen utilizarse también para probar algunos procedimientos, pero con mayor frecuencia para verificar las propiedades psicométricas de algún instrumento a partir de una prueba piloto, es posible encontrar algunas reglas similares de tamaños fijos de muestras. Una regla de este tipo es la propuesta de McMillan y Schumacher (2005), quienes señalan que aun cuando el tamaño de la muestra para un estudio piloto debería ser mayor de 20, lo importante es que en toda investigación se realice un estudio piloto, aunque para ello se cuente con al menos 10 sujetos.

Otros autores como García-García *et al.* (2013) y Chavarria y Pulgarín (2020), señalan que el tamaño de muestra para pruebas piloto debe ser entre 30 y 50 participantes, quienes además deben poseer los atributos que se pretende medir en la población objetivo; estimado que supuestamente retoman de las recomendaciones de Earl Babbie, particularmente de su obra *Fundamentos de investigación social*. No obstante, Babbie (2000) en dicha obra, aunque sí desarrolla un capítulo completo sobre la lógica del muestreo, donde expone diversos cálculos de tamaños muestrales dependiendo del tipo o diseño de investigación, lo cierto es que nunca se refiere directamente a los estudios piloto. La referencia más cercana a este aspecto es una anotación que hace este autor en relación de la prueba previa (pretest) de los cuestionarios, señalando que “no es esencial que los sujetos de la prueba previa comprendan una muestra representativa, pero debe acudir a personas para quienes el cuestionario sea por lo menos pertinente” (Babbie, 2000: 138), mas no establece de manera directa ninguna cifra estimada para tal efecto.

En otra de sus obras clásicas, específicamente sobre los instrumentos de medición, Babbie (1990) resalta la relevancia de las pruebas previas como una estrategia útil para identificar posibles errores antes de un estudio final, ahorrando dinero y esfuerzos innecesarios, en caso de que se considere que no es posible lograr los objetivos esperados. Babbie (1990) sí distingue aquí entre las pruebas previas de los instrumentos (*pretests*) y los estudios piloto (*pilot studies*), señalando que los primeros deben ser flexibles y variados, lo que implica ignorar por completo el muestreo controlado y la representatividad estricta, para seleccionar sujetos que estén relacionados con el diseño de investigación y los instrumentos bajo consideración, de modo que puedan contribuir a su mejora. En cambio, los estudios piloto, que para este autor son una versión en miniatura del diseño del estudio final, sí deben incluir una muestra representativa de la población objetivo, seleccionada de la misma forma en que se pretende seleccionar la muestra final, mediante una estrategia que contemple o no la superposición de individuos, según lo permita el estudio planteado.

Aunque Babbie tampoco especifica en esta obra cómo determinar el tamaño de muestra para un estudio piloto, es posible inferir que al considerar a este tipo de estudios con la misma rigurosidad que a los estudios finales, se debe utilizar un procedimiento de determinación del tamaño de la muestra, similar al que se utilizará dependiendo el diseño de la investigación final. Para ello se debería utilizar el mismo principio que este autor denomina como PPS (*Probability Proportionate to Size Sampling*), para resaltar la importancia de que, a menor tamaño de la muestra, mayor probabilidad de error muestral y viceversa (Babbie, 1990). Con esta acotación, se regresa a dos reglas generales ya discutidas en este trabajo a partir de los planteamientos de Kerlinger & Lee (2002) –sobre los que sin duda coincide Babbie (1990, 2000)– respecto de que muestras grandes son necesariamente mejores en cualquier tipo de estudio y que, por tanto, entre más grande sea la muestra, menores serán los errores resultantes, principios que al menos en ciencias sociales, pueden ser muy cuestionados.

No obstante, sí existen autores que consideran algunas reglas o principios a seguir para determinar el tamaño de una muestra, ya sea en las pruebas piloto o en los estudios piloto. Recientemente Roco *et al.* (2021), elaboraron un breve, pero muy interesante, escrito donde enfatizan un problema similar al antes planteado, que es precisamente la falta de fundamentación que respalde la delimitación del tamaño muestral en las investigaciones que se realizan para la validación de instrumentos, aun cuando dichas investigaciones que se publican, sí cuentan con un análisis detallado de validación a partir de los resultados obtenidos de las muestras.

Lo que se cuestiona, por tanto, no es la calidad de los análisis realizados para validar instrumentos, sino la falta de fundamento sobre la determinación del tamaño de la muestra a partir de la cual se elaboran dichos análisis. Si se es muy crítico al respecto, esto sí implica cuestionar los resultados, debido a las inconsistencias o lagunas en el origen y selección de lo(a)s sujetos de estudio, que como ya ha sido señalado, en muchas ocasiones se realizan de forma arbitraria. Bajo este cuestionamiento, Roco *et al.* (2021: 877) aclaran que: “Si bien no existe un único criterio para determinar el tamaño muestral en el proceso de validación de un cuestionario, sí existen una serie de recomendaciones que creemos que es importante tener siempre a la vista”. Entre las cuales, usando las aportaciones de otros autores, destacan las siguientes propuestas, cuyas fuentes fueron cotejadas y revisadas para este escrito:

1. Si se pretenden realizar análisis estadísticos de los ítems a partir de la media, desviación típica y coeficiente de correlación, puede ser suficiente contar con una muestra de entre 50 y 100 sujetos para una primera aplicación, previendo que se contará con una segunda aplicación ya con el instrumento mejorado, donde se pueden incluir entre 5 y 10 sujetos por ítem, con un mínimo de 300 sujetos, para así poder tener mejores garantías sobre la validez del instrumento (Osterlind, 2002 y Martínez-Arias, 1995).
2. Si se van a utilizar procedimientos como el análisis factorial confirmatorio, para evaluar la estructura interna de los instrumentos, la recomendación es incluir entre 5 y 10, o incluso entre 2 y 3 participantes por ítem, siempre y cuando el número total no sea inferior a 200 (Stevens, 2009 y Kline, 1986).
3. Snook & Gorsuch (1989), proponen a 300 participantes como regla general, para obtener soluciones fiables en cualquier tipo de análisis factorial realizado.
4. Si se pretende estimar la fiabilidad de un instrumento mediante estadísticos test-retest, formas paralelas o alfa de Cronbach; el estimado de participantes puede ser de entre 200 y 300 (Clark & Watson, 1995).
5. Muestras más pequeñas podrían ser aceptables si la metodología [o diseño] del estudio de validación, tiene contemplado replicar la medición en diferentes grupos, donde el estudio final considere incluir al menos al doble de sujetos que el número de ítems del instrumento a validar, con un mínimo de 100 participantes por grupo (Kline, 1986).

A partir de esta revisión de autores, Roco *et al.* (2021) también concuerdan en la existencia de diversos criterios de tamaños muestrales para validar instrumentos, concluyendo que antes de decidir un número:

Debemos primero asegurarnos de que el tamaño que estimemos permita que todos los segmentos de la población que sean de nuestro interés estén correctamente representados, evitando sesgos por exceso o deficiencia. Si esta consideración no se cumple, el instrumento no logrará su finalidad, a pesar de que se nos señale estadísticamente que estamos en presencia de un instrumento fiable (Roco *et al.*, 2021: 878).

No obstante lo categóricas que parezcan algunas de las propuestas mencionadas, es importante señalar que incluso los mismos autores citados por Roco *et al.* (2021), tampoco son tan dogmáticos al proponer un número fijo o un valor estimado para el tamaño muestral [*flat rule*], como si siempre fuera ésta la mejor opción. Por ejemplo, Osterlind (2002), señala que, al momento de construir un instrumento, no existe un número específico de personas que deban participar en la determinación de una respuesta “correcta” para los ítems de una prueba. Para algunos ítems, sólo dos o tres personas pueden ser suficientes, mientras que, en otros contextos o ítems, se deben incluir más personas en el consenso. Aunque sí se recomienda que para programas de evaluación a gran escala se incluya en el consenso de cada ítem, un mínimo de tres personas, y preferiblemente cuatro o incluso cinco. Esto no implica que este número deba multiplicarse por el total de los ítems de la prueba, sino más bien considerar las respuestas de estas personas en conjunto, buscando establecer un consenso a partir las mismas.

Osterlind (2002) también señala que hay que tomar en cuenta que la validez no se consigue a través del instrumento en sí mismo, sino que es la interpretación de las puntuaciones obtenidas luego de su aplicación, lo que le da significado. En tal sentido, la validez de una prueba se construye como un proceso escalonado de reunir evidencia, para generar una interpretación específica de las puntuaciones obtenidas [*stepped rule*]. Éstas deberán evaluarse de manera concreta a partir de diferentes métodos que justifiquen las inferencias particulares realizadas a partir de los datos recabados, tomando en cuenta que algunas pruebas pueden ser apropiadas para ciertos tipos de inferencias, pero no para otras.

Diversos autore(a)s como Fox *et al.* (2009), Lewis *et al.* (2021) y Stevens (2009), enfatizan la importancia del tamaño de la muestra en términos paramétricos, por su relación directa con la potencia estadística y los errores alfa y beta o de tipo I y II (la probabilidad de rechazar una hipótesis nula cuando es verdadera, o de aceptarla cuando es falsa), lo cual según estos

autore(a)s, en muestras muy grandes (>200) no genera un problema. No obstante, aclara Stevens, si el tamaño de la muestra no es lo suficientemente grande, se puede recurrir a un agrupamiento lógico basado en fundamentos teóricos o sustantivos, por ejemplo en los análisis factoriales, donde a pesar de que muchos autores han recomendado tamaños de muestra mínimos de 200, lo cierto es que muestras más pequeñas también pueden ser adecuadas para modelos con menos parámetros a estimar, e incluso un tamaño de muestra grande puede derivar en la retención de demasiados factores: “un tamaño de muestra grande puede llevar a la retención de demasiados factores” (Stevens, 2009: 329; la traducción es del autor).

Stevens (2009) enfatiza además que, entre los factores retenidos con muestras demasiado amplias, pueden incluirse algunos que en realidad son triviales, como también lo advierte Goodwyn: “Los mismos problemas asociados con el tamaño de la muestra en las pruebas de significancia estadística tradicionales, se presentan cuando estas pruebas se aplican a la retención de factores. Pues utilizar muestras grandes en el análisis factorial, es una estrategia que a menudo suele considerar factores triviales como estadísticamente significativos” (2012: 5; la traducción es del autor).

Como puede observarse, tanto en los estudios piloto, como en las pruebas piloto o pruebas previas de los instrumentos a utilizar en una investigación con aplicación final, no existe una fórmula preestablecida ni un número “mágico” que represente el tamaño ideal de una muestra paramétrica o no paramétrica, por lo que tampoco son totalmente aplicables las reglas generales de números fijos o *flat rules*. Más bien, se plantean diversas opciones de muestras, generalmente reducidas, que pueden funcionar dependiendo del propio diseño y fines de la investigación.

En la literatura existente al respecto en ciencias sociales, parece predominar la sugerencia de que la selección de los casos no se realice de manera aleatoria, sino por su pertinencia en relación con los atributos o características que se pretendan medir, o bien, por los elementos que se pretendan probar con los instrumentos que se estén piloteando, sugiriendo que la muestra de participantes, sea seleccionada de forma no necesariamente paramétrica, sino más bien intencional, deliberada o a juicio de lo(a)s investigadore(a)s, lo que corresponde con tamaños de muestra que se pueden ir mejorando desde una perspectiva más escalonada, tipo *stepped rules*.

Otro aspecto a considerar son las diferencias entre los estudios piloto internos y externos, ya que según lo propuesto por Herbert *et al.* (2019, citado por Lewis *et al.*, 2021), esta distinción es muy importante al momento de definir el tamaño de una muestra, debido a que los estudios piloto externos suelen ser ligeramente mayores y diferentes a los internos, que se pueden basar en un porcentaje del tamaño total de la muestra y ser además una parte ya programada de un estudio mayor.

A pesar de que los estudios piloto internos pueden ser muy atractivos, porque es posible incluir los datos de lo(a) participantes en el mismo análisis que se realice para el estudio principal, de modo que el cálculo de la muestra del estudio piloto interno puede no ser necesario, pues se incluirían a una porción de los sujetos que conformarían la muestra total del estudio, lo cierto es que los estudios piloto internos presentan también sus propios riesgos. Basta con citar lo señalado por In (2017), quien aclara que, para realizar un estudio piloto interno, éste debe planificarse minuciosamente como parte de la etapa de diseño del estudio principal, e incluirlo como parte de su procedimiento, de modo que al menos una de las muestras del estudio piloto interno, debe ser idéntica a la del estudio principal, y por tanto se tienen que utilizar los mismos criterios de inclusión y exclusión, que en ocasiones no se conocen bien desde el estudio piloto.

Hay que considerar también como lo señala este autor (In, 2017), que no sería posible realizar cambios sugeridos a partir del estudio piloto, si es que estos afectan a otras categorías del estudio principal. Además, el empleo de estudios piloto internos genera una probabilidad ligeramente mayor de cometer errores tipo I (o alfa), debido a que las hipótesis que se establecen tanto para el estudio piloto como para el estudio principal, suelen ser independientes entre sí. También por el hecho de que los estudios piloto no tienen como finalidad el probar hipótesis, sino comprender mejor la viabilidad del reclutamiento de los participantes o del diseño del estudio principal [o de los instrumentos], por lo que el tamaño de la muestra o muestras, debe establecerse con base en ello, y no con la finalidad de probar hipótesis.

Conclusiones

Ante esta diversidad de propuestas respecto de un tamaño idóneo de muestras, tanto para las pruebas piloto como para los estudios piloto; o bien, ante la posible aplicación de reglas generales para definir dicho tamaño, prácticamente en cualquier campo de investigación, pero especialmente en las ciencias sociales, parece ser que la única certeza, es que no existe tal certeza. Por eso es que en los libros clásicos de metodología de la investigación, se abordan fórmulas y procedimientos específicos para delimitar el tamaño muestral dependiendo el tipo o diseño de investigación, pero rara vez se incluyen a los estudios piloto o a las pruebas previas, que suelen realizarse para validar instrumentos, determinar errores muestrales, generar una idea previa del comportamiento que podrían tener los resultados en un estudio mayor, determinar la necesidad de ajustes en los procedimientos, estimar el costo o viabilidad de un estudio, o cualquier otra finalidad para la cual, se considere necesario realizar un estudio piloto.

La respuesta en estos casos no está en la tendencia general de trabajar siempre con muestras grandes. Autores como Babbie (1990), Stevens (2009), Goodwyn (2012) o Kerlinger & Lee (2002), aunque resaltan las ventajas estadísticas de las muestras grandes, también señalan que éstas no están necesariamente libres de sesgos, por lo que no deberían ser una regla fija para todos los tipos de estudios. Esto es especialmente aplicable a las pruebas o estudios piloto internos, donde la muestra utilizada es en realidad una pequeña porción de la muestra global del estudio principal, la cual necesariamente será más grande, debiendo ser calculada en relación con la población total de estudio con la cual se trabaje.

La respuesta parece ir más bien en el sentido planteado por Lewis *et al.* (2021), quienes destacan precisamente que no existe un consenso respecto de la determinación del tamaño de muestra requerido en los estudios piloto, pues las recomendaciones suelen ser muy variadas dependiendo incluso de quien las realice. Ante estas variaciones, Lewis *et al.* (2021) advierten que, antes de cualquier requisito de muestra fija, el tamaño de la misma puede basarse en el nivel de precisión de un parámetro particular a utilizar para calcular la viabilidad, en la tasa de eventos a incluir en el estudio principal, así como en el tamaño del efecto esperado.

A estos, se podrían agregar otros criterios no tan mencionados, como la posibilidad de aceptación de participar de los sujetos seleccionados, la fidelidad del tratamiento o instrumentos a aplicar, y la retención esperada de lo(a)s participantes. También se sugieren opciones complementarias para disminuir los errores de estimación en los análisis estadísticos, que pueden ser especialmente útiles para los estudios piloto, entre las que se destacan principalmente dos: 1) una adecuada y exhaustiva revisión bibliográfica sobre los temas o variables en los que se va a fundamentar el estudio piloto y, 2) la consulta previa de opiniones de expertos; que pueden complementar la determinación de criterios teóricos y estadísticos para asignar un tamaño de muestra más apropiado, contribuyendo a una mejor confiabilidad y validez.

Es importante mencionar algunos otros supuestos previos, que pueden complicar el empleo de muestras grandes en los estudios piloto, entre los que se destacan:

1. Cuando la población a investigar no es muy extensa, y una muestra amplia puede originar que en el estudio final se dupliquen muchos casos que ya participaron en el pilotaje, sobre todo si éste se realiza como un estudio piloto externo. Ante esta situación, aunque tampoco existe un criterio respecto del tiempo que debe pasar entre la aplicación de una prueba y otra (test-retest), las estimaciones fluctúan entre seis y doce meses cuando se trata de encuestas o pruebas psicológicas, debido a que se asume que existe aprendizaje o habituación a los reactivos, que puede afectar los resultados de la segunda aplicación.
2. Cuando cada caso o sujeto que se estudie supone un costo alto en términos de recursos, generalmente financieros, pero también de otros tipos como humanos o técnicos. Un ejemplo sería el pilotaje de un procedimiento médico innovador altamente costoso, como los derivados de la ingeniería genética o la ingeniería de células madre, donde además de los conflictos éticos que puede suponer su desarrollo, los procedimientos pueden ser muy complejos, por lo que su pilotaje suele hacerse generalmente en especies animales, o con grupos muy reducidos de personas.

3. Cuando los casos o sujetos de estudio son muy reducidos, debido a ciertas características especiales o muy particulares. Por ejemplo, si se realizara una investigación con gemelos monocigóticos, que además por cualquier circunstancia hayan sido separados en sus primeras semanas o meses de vida y criados de manera distinta, lo cual podría ser muy útil para profundizar sobre la diferencia entre lo biológico (genética) y lo social (aprendizaje) en la formación de la personalidad. Las posibilidades de detectar este tipo de casos serían de una entre miles, por lo que habría que conformarse tal vez con algunos pocos casos tanto para los estudios piloto que se requieran (si es que se requieren), como para el estudio final.
4. Cuando se realizarán estudios destructivos y las muestras agotan la posibilidad de experimentación, por lo que no hay forma de replicar los procedimientos con el resto de la población, o bien, su replicación puede ser muy compleja. Aquí se incluye el empleo de procedimientos químicos, físicos o mecánicos que suelen emplearse en ingenierías para probar la duración o calidad de algunos materiales, aunque podrían existir casos similares en otras disciplinas, donde un procedimiento implique dañar o destruir las muestras elegidas para los ensayos.
5. En tipos o diseños de estudio particulares, generalmente de corte cualitativo y no paramétrico, donde el diseño mismo no requiere incluir muchos casos para el estudio final, y en ocasiones tampoco es necesaria la selección de una muestra, como podrían ser las historias de vida, los estudios etnográficos, los diseños fenomenológicos y algunos estudios de casos cualitativos. Estos estudios son especialmente útiles en ciencias sociales, y los estudios piloto en este ámbito, en muchas ocasiones ni siquiera son necesarios, aunque también existen perspectivas teóricas distintas sobre el muestreo cualitativo y no paramétrico, que deberían ser revisadas a detalle cuando se desarrollen este tipo de estudios.
6. Cuando la disponibilidad de sujetos para integrar las muestras es muy limitada, como podría ocurrir en investigaciones que abordan temas muy sensibles donde puede ser difícil conseguir informantes, no por sus características personales, sino por las del tema a investigar que puede generar un muy bajo nivel de participación. Ejemplos de estos temas pueden ser las situaciones vinculadas con el narcotráfico, la tortura, la discriminación, los crímenes violentos, las violaciones graves a los derechos humanos, entre otros, donde una muestra pequeña se puede justificar a partir de la relevancia teórica de los sujetos de estudio incluidos, y no necesariamente por su representatividad estadística.

En estos casos u otras situaciones de investigación particulares, donde no es posible cumplir la tendencia general de trabajar con muestras grandes, la decisión no puede depender de un criterio plano de muestreo o *flat rule*, sino que el tamaño de la muestra, si es que es requerida, deberá ser una decisión ampliamente razonada de manera crítica, por parte de quienes dirijan la investigación. Esta decisión se deberá construir desde una perspectiva sistemática de muestreo escalonado tipo *stepped rule*, la cual considere principalmente las características de la población a incluir en el estudio, así como el diseño y los objetivos de la investigación a realizar, y del estudio piloto en caso de que se requiera, pues como se ha señalado en algunos de los ejemplos anteriores, existen situaciones de investigación donde los estudios piloto pueden no ser necesarios.

Algunas consideraciones sobre el tamaño muestral, también serán distintas en un estudio piloto realizado para validar un instrumento, donde se requerirá un número más amplio de casos para poder efectuar ciertos tipos de análisis estadísticos (como los análisis factoriales exploratorios y confirmatorios), que cuando se realiza simplemente para calcular un error estándar necesario para una muestra final, o bien, para conocer otros aspectos de los instrumentos, como la claridad en sus instrucciones, estructura de ítems, opciones de respuesta, tiempo de aplicación, etc., para lo cual puede no ser necesaria una muestra tan amplia en el pilotaje, y podría incluso ser recomendable alguna forma de muestreo no paramétrico.

Por último, todo proceso de determinación del tamaño muestral, debería considerar las características propias de cada investigación, por lo que antes de tomar una decisión sobre el tamaño de una muestra en un estudio piloto o una prueba previa, es importante analizar de

manera detallada también otros aspectos como la disponibilidad de recursos, el tiempo que se tiene para realizar la investigación, la accesibilidad a los sujetos de estudio, el tipo de datos a recabar para su análisis, entre otros. Una reflexión profunda de estos y otros aspectos propios de las investigaciones, no sólo serviría para determinar de mejor manera un tamaño muestral, sino que podría evitar posibles errores o sesgos en el estudio final, lo cual, es precisamente una de las finalidades principales de los estudios piloto.

Referencias

- BABBIE, E. (2000). *Fundamentos de investigación social*. México: International Thomson Editores.
- _____ (1990). *Survey Research Methods*. Belmont, California: Wadsworth Publishing Company.
- BROWNE, R. H. (1995). On the use of a pilot sample for sample size determination. *Stat Med*, 14(17), 1933-1940. DOI: [10.1002/sim.4780141709](https://doi.org/10.1002/sim.4780141709)
- BUNGE, M. (1983). *La investigación científica. Su estrategia y su filosofía*. México: Editorial Ariel S. A.
- Centers for Disease Control and Prevention (CDC). (2022). *The Epi Info™ Story*. Recuperado de <https://www.cdc.gov/epiinfo/story.html>
- CHAVARRIA, C. T. M. y PULGARÍN, M. S. A. (2020). Construcción y validación de un instrumento para caracterizar el nivel de innovación en instituciones prestadoras de servicios de salud. *Revista de métodos cuantitativos para la economía y la empresa*, (30), 258-278. <https://doi.org/10.46661/revmetodoscuanteconempresa.3519>
- CLARK, L. A. & WATSON, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319. <https://doi.org/10.1037/1040-3590.7.3.309>
- COCHRAN, W. G. (1977). *Técnicas de muestreo*. México: Compañía Editorial Continental.
- Díaz-Muñoz, G. (2020). Metodología del estudio piloto. *Revista chilena de radiología*, 26(3), 100-104. <http://dx.doi.org/10.4067/S0717-93082020000300100>
- DIETERICH, H. (2021). *Nueva guía para la investigación científica*. México: Grupo Editor Orfila Valentini.
- FOX, N., HUNN, A. & MATTERS, N. (2009). *Sampling and Sample Size Calculation*. National Institute for Health Research. Recuperado de: <https://www.studocu.com/row/document/haramaya-university/community-based-training-program/sampling-and-sample-size-calculation/107278420>
- GARCÍA-GARCÍA, J. A.; REDING-BERNAL, A. y LÓPEZ-ALVARENGA, J. C. (2013). Cálculo del tamaño de la muestra en investigación en educación médica. *Investigación en Educación Médica*, 2 (8), 217-224.
- GOODWYN, F. (2012). Question Number Two: How Many Factors? *Annual meeting of the Southwest Educational Research Association*, 1-16.
- HERNÁNDEZ, S. R., FERNÁNDEZ, C. C. y BAPTISTA, L. M. P. (2014). *Metodología de la investigación*. México: McGraw-Hill.
- HERNÁNDEZ, S. R. y MENDOZA, T. C. P. (2018). *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*. México: McGraw-Hill.
- IN, J. (2017). Introduction of a pilot study. *Korean Journal of Anesthesiology*, 70(6), 601-605. <https://doi.org/10.4097/kjae.2017.70.6.601>
- JULIOUS, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics*. 4(4), 287-291. <https://doi.org/10.1002/pst.185>
- KERLINGER, F. N. & LEE, H. B. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales*. México: McGraw-Hill.
- KIESER, M. & WASSMER, G (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal*, 38(8), 941-949. <https://doi.org/10.1002/bimj.4710380806>
- KLINE, P. (1986). *A Handbook of test Construction*. New York: Methuen.

- LEWIS, M.; BROMLEY, K.; SUTTON, C. J.; MCGRAY, G.; MYERS, H. L. & LANCASTER, G. A. (2021). Determining sample size for progression criteria for pragmatic pilot RCTs: the hypothesis test strikes back! *Pilot and Feasibility Studies*, 7(40), 1-14. <https://doi.org/10.1186/s40814-021-00770-x>
- MARRUGAT, J.; VILA, J.; PAVESI, M. y SANZ, F. (1998). Estimación del tamaño de la muestra en la investigación clínica y epidemiológica. *Medicina clínica*, 111, 267-76.
- MARTÍNEZ-ARIAS, R. (1995). *Psicometría: Teoría de los test psicológicos y educativos*. Madrid: Síntesis.
- MCMILLAN, J. H. y SCHUMACHER, S. (2005). *Investigación educativa. Una introducción conceptual*. Madrid: Pearson.
- OSTERLIND, S. J. (2002). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats*. New York: Kluwer Academic Publishers.
- QUALTRICS (2023). *Sample size calculator*. Recuperado de: <https://www.qualtrics.com/blog/calculating-sample-size/>
- ROCO V., A.; HERNÁNDEZ O., M. y SILVA G., O. (2021) ¿Cuál es el tamaño muestral adecuado para validar un cuestionario? *Nutrición Hospitalaria*, 38(4), 877-878. <https://dx.doi.org/10.20960/nh.03633>
- SIM, J. & LEWIS, M. (2011). The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *Journal of Clinical Epidemiology*, 65(3) 301-308. <https://doi.org/10.1016/j.jclinepi.2011.07.011>
- SNOOK, S. C. & GORSUCH, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, 106(1), 148-154. <https://doi.org/10.1037/0033-2909.106.1.148>
- SOFTWARE PARA ENCUESTAS QUESTIONPRO (2023). *Calculadora de muestras*. Recuperado de: https://www.questionpro.com/es/calculadora-de-muestra.html#calculadora_de_muestra
- SURVEY MONKEY (2023). *Calculadora del tamaño de muestra*. Recuperado de: <https://es.surveymonkey.com/mp/sample-size-calculator/>
- STEVENS, J. P. (2009). *Applied multivariate statistics for the social sciences*. New York: Routledge.
- TEARE, M. D.; DIMAIRO, M.; SHEPHARD, N.; HAYMAN, A.; WHITEHEAD, A. & WALTERS, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, 15(264), 1-13. DOI: [10.1186/1745-6215-15-264](https://doi.org/10.1186/1745-6215-15-264)
- WHITEHEAD, A. L.; JULIOUS, S. A.; COOPER, C. L. y CAMPBELL, M. J. (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research*, 25(3), 1057-1073.

Autor.

Rodrigo Vargas Salomón

Departamento de Ciencias de la Salud del Centro Universitario de Los Altos, Universidad de Guadalajara, México.

Doctor en Ciencias Sociales por el Colegio de Jalisco y Maestro en Psicología por la Universidad de Guadalajara. Profesor-Investigador Asociado C de la Universidad de Guadalajara.

E-mail: rodrigo.vargas0848@academicos.udg.mx

Citado.

VARGAS SALOMÓN, Rodrigo (2025). Delimitación del tamaño muestral en estudios piloto. *Revista Latinoamericana de Metodología de la Investigación Social – ReLMIS*, N°29, Año 15, pp. 7-21.

Plazos.

Recibido: 02/06/2024. Aceptado: 01/11/2024.



Distribución de Pareto: un modelo de probabilidad en investigación social

Pareto distribution: a probability model in social research

José Moral de la Rubia

Resumen

Este artículo metodológico tiene como objetivo exponer la distribución de Pareto de una forma clara y ejemplificada para su comprensión entre investigadores sociales. Esta distribución continua con forma de J invertida, asimetría y cola pesada a la derecha constituye un buen modelo de probabilidad para diversas variables sociales, como riqueza e ingreso, y de conductas que son altamente frecuentes en pocas personas y poco frecuentes en la mayoría. Se desarrolla la distribución tipo I que tiene un parámetro de escala x_m y otro de forma α . Se inicia con una breve reseña histórica. Se muestran las funciones de densidad, distribución acumulativa, de cola, de momentos y característica. Se sigue con las medidas descriptivas, estimadores por los métodos de momentos y máxima verosimilitud, su relación con otras distribuciones y las pruebas de bondad de ajuste. Se aplica todo este material con dos ejemplos: uno de cálculo de probabilidades y medidas descriptivas, y otro de estimación de parámetros y comprobación del ajuste. Finalmente, se hacen sugerencias para su uso.

Palabras clave: distribución de probabilidad continua; estimación de parámetros; medidas descriptivas; índice de cola de Pareto; índice de concentración de Gini.

Abstract

This methodological article aims to expose the Pareto distribution in a clear and exemplified way for its understanding among social researchers. This continuous distribution with an inverted J shape, skewness, and a heavy tail to the right constitutes a good probability model for various social variables, such as wealth and income, and for behaviors that are highly frequent in a few people and infrequent in the majority. The type-I distribution is developed, which has a scale parameter x_m and another of shape α . It begins with a brief historical review. The density, cumulative distribution, tail, moment, and characteristic functions are shown. It continues with the descriptive measures, estimators by the methods of moments and maximum likelihood, its relationship with other distributions, and the goodness-of-fit tests. All this material is applied with two examples: one of calculating probabilities and descriptive measures and the other of parameter estimation and fit testing. Finally, suggestions for its use are made.

Keywords: continuous probability distribution; parameter estimation; descriptive measures; Pareto tail index; Gini concentration index.

1. Introducción¹

La distribución de Pareto es una de las distribuciones continuas no normales que puede ser de gran utilidad en el área de las Ciencias Sociales y de la Conducta (Barnoy y Reich, 2022; Feng et al., 2020). Cabe señalar que la exposición explicativa de esta distribución suele ser muy teórica y confusa, sobre todo por ser presentada bajo distintas parametrizaciones (Sarabia, Jorda y Prieto, 2019) y muy enfocada al área de la economía (Barczy, Nedényi y Sütő, 2023), de ahí que el objetivo de este artículo es presentar esta distribución de probabilidad de una forma clara, comprensible y ejemplificada para distintos tipos de investigadores sociales, como psicólogos, sociólogos y trabajadores sociales.

La distribución fue desarrollada entre 1896 y 1897 por el ingeniero civil, economista y sociólogo italiano Vilfredo Federico Damaso Pareto (1848–1923) durante su curso de economía política impartido en la universidad de Lausana (UNIL), Suiza. Surge para describir el reparto de la tierra, riqueza e ingresos en una sociedad. De estos estudios, se infiere el principio 20-80 que enuncia que el 80% de la riqueza se concentra en el 20% de la población. Cabe señalar que tanto la distribución como el principio de desigualdad hoy en día se aplican como modelos de probabilidad para estas mismas variables (Charpentier y Flachaire, 2022; Feng et al., 2020), aunque la desigualdad suele ser menor, con un máximo de 30-70 (McCarthy y Winer, 2019). Además, la distribución y la regla de desigualdad sirven como modelos en otros ámbitos distintos a la economía, como en psicología (Campbell y Brauer, 2021; Rajeev, 2022), ingeniería (Chen et al., 2019; Sudharson y Prabha, 2019), climatología (Le Gall, et al., 2022) y física (Rácz et al., 2023; Xiong et al., 2019). Las reclamaciones a las aseguradoras por accidentes y enfermedades siguen una distribución de Pareto, lo que lleva a estas compañías a castigar excesivamente las primas de seguros de determinados sectores de la población, como adultos mayores (Diawara et al., 2021; Zhang, Wu y Yao, 2022).

2. Caracterización de la distribución

2.1. Parámetros y soporte

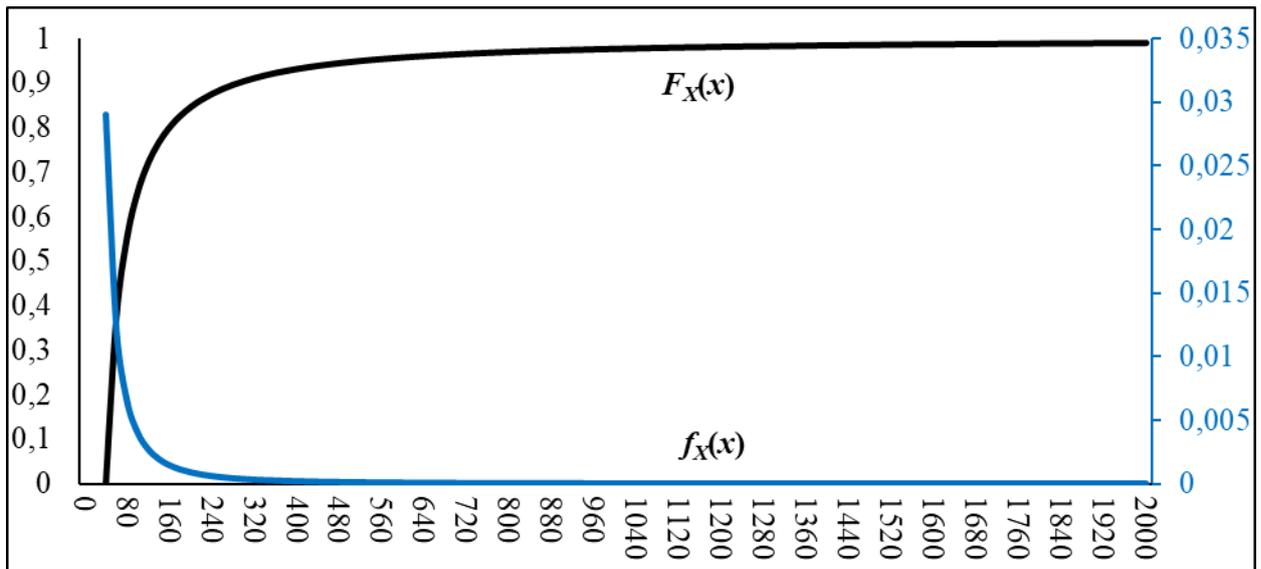
Esta distribución continua de la familia de distribuciones de potencia está determinada por dos parámetros en su presentación más sencilla o tipo I (Ahmad y Almetwally, 2020). Un parámetro es de escala, se puede denotar por x_m , aunque su notación es muy variable, y corresponde al pico o moda y al valor mínimo de la distribución (Chattamvelli y Shanmugam, 2021). El otro parámetro es de forma, se denota por α (Fedotenkov, 2020) y se suele denominar índice de cola de Pareto (Andria, 2022). El espacio parametral de ambos abarca el intervalo $(0, \infty)$. El soporte de la distribución corresponde al intervalo $[x_m, \infty)$. El perfil de su función de densidad describe una curva con un descenso rápido desde su valor máximo hasta hacerse asíntota al eje de abscisas, por lo que presenta asimetría positiva y leptocurtosis con cola larga a la derecha (Figura 1). El valor de α que modela con exactitud la ley 20-80 es $\log_4(5) = \ln(5)/\ln(4) \approx 1.161$, el cual es un valor demasiado extremo para la distribución de la riqueza que, usualmente, es mayor que 1.5 (Yang y Zhou, 2022).

$$X \sim \text{Pareto}(x_m, \alpha)$$

$$x_m \text{ y } \alpha \in (0, \infty) \text{ y } x \in [x_m, \infty)$$

¹ Esta es la versión en español del artículo: Moral de la Rubia, J. (2025). Pareto distribution: a probability model in social research. *Open Journal of Social Sciences*, 13(1), 86-121. <https://doi.org/10.4236/jss.2025.131007>

Figura 1. Función de densidad $f_X(x)$ y probabilidad acumulativa $F_X(x)$ del ingreso por hora trabajada de una distribución de Pareto con parámetro de escala $x_m = 40$ y forma $\alpha = \log_4(5)$.



Fuente: elaboración propia.

2.2. Funciones

Función de densidad:

$$f_X(x) = \alpha \frac{x_m^\alpha}{x^{\alpha+1}}, \text{ para valores de } x \geq x_m$$

Función de distribución acumulativa:

$$F_X(x) = P(X \leq x) = \int_{x_m}^x f_X(x) dx = 1 - \left(\frac{x_m}{x}\right)^\alpha, \text{ para valores de } x \geq x_m$$

Función de distribución acumulativa complementaria o de cola:

$$\bar{F}_X(x) = P(X > x) = \int_{x_m}^{\infty} f_X(x) dx = \left(\frac{x_m}{x}\right)^\alpha, \text{ para valores de } x \geq x_m$$

Función cuantil:

$$Q_X(p) = \frac{x_m}{\sqrt[\alpha]{1-p}} = x_m(1-p)^{-\frac{1}{\alpha}}, \text{ donde } p \in (0, 1)$$

Función generadora de momentos: Para un intervalo no trivial de t en torno a 0 no está definida, pero sí para valores de $t \leq 0$.

$$M_X(t) = E(e^{tx}) = \int_{x_m}^{+\infty} e^{tx} f_X(x) dx = \begin{cases} \alpha(-x_m t)^\alpha \Gamma(-\alpha, -x_m t) & t < 0 \\ 1 & t = 0 \\ \text{Indefinida} & t > 0 \end{cases}$$

Función gamma incompleta superior:

$$\Gamma(-\alpha, -x_m t) = \int_{-x_m t}^{\infty} u^{-\alpha-1} e^{-u} du$$

Función característica:

$$i = \sqrt{-1} \text{ y } t \in \mathbb{R}$$

$$C_X(t): \mathbb{R} \rightarrow \mathbb{C}; C_X(t) = E(e^{itx}) = \int_{x_m}^{\infty} e^{itx} f_X(x) dx = \alpha(-ix_m t)^{\alpha} \Gamma(-\alpha, -ix_m t)$$

2.3. Medidas descriptivas de tendencia central, variación y forma

Esperanza matemática o media aritmética:

$$\mu(X) = E(X) = \int_{x_m}^{\infty} x f_X(x) dx = \frac{\alpha}{\alpha-1} x_m, \alpha > 1$$

Media geométrica:

$$\mu_g(X) = G(X) = e^{E(\ln(X))} = x_m \sqrt[\alpha]{e} = x_m e^{1/\alpha}$$

Media armónica:

$$\mu_h(X) = H(X) = \frac{1}{E(1/X)} = x_m \left(1 + \frac{1}{\alpha}\right)$$

Mediana:

$$Md_n(X) = Q_X(p = 0.5) = \frac{x_m}{\sqrt[\alpha]{1/2}} = x_m \sqrt[\alpha]{2}$$

Moda:

$$Mo(X) = \{x | \max[f_X(x)]\} = x_m$$

Momento no central de orden k :

$$E(X^k) = \int_{x_m}^{\infty} x^k f_X(x) dx = \frac{\alpha}{\alpha-k} x_m^k, \alpha > k$$

Varianza:

$$Var(X) = \sigma^2(X) = E[(X - E(X))^2] = \int_{x_m}^{\infty} (x - E(X))^2 f_X(x) dx = \frac{\alpha x_m^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2$$

Desviación estándar:

$$DE(X) = \sigma(X) = \sqrt{Var(X)} = \sqrt{\frac{\alpha x_m^2}{(\alpha - 1)^2(\alpha - 2)}} = \frac{x_m}{\alpha - 1} \sqrt{\frac{\alpha}{\alpha - 2}}, \alpha > 2$$

Entropía: Al estar el logaritmo con base neperiana, la entropía viene en nats.

$$H(X) = E(-\ln[f_X(x)]) = - \int_{x_m}^{\infty} f_X(x) \ln[f_X(x)] dx = \ln\left[\frac{x_m}{\alpha} e^{1+\frac{1}{\alpha}}\right] = \ln(x_m) - \ln(\alpha) + 1 + 1/\alpha$$

Asimetría y exceso de curtosis basadas en momentos centrales con la notación original de Karl Pearson:

$$\sqrt{\beta_1(X)} = \frac{E[(X - E(X))^3]}{(E[(X - E(X))^2])^{3/2}} = \frac{2(\alpha + 1)}{\alpha - 3} \sqrt{\frac{\alpha - 2}{\alpha}}, \alpha > 3$$

$$\beta_2(X) - 3 = \frac{E[(X - E(X))^4]}{(E[(X - E(X))^2])^2} = \frac{6(\alpha^3 + \alpha^2 - 6\alpha - 2)}{\alpha(\alpha - 3)(\alpha - 4)}, \alpha > 4$$

3. Estimación de parámetros

Sea una muestra aleatoria de tamaño n de una variable cuantitativa continua X que sigue una distribución de Pareto tipo I con parámetros x_m y α . Por ejemplo, los datos muestrales podrían ser registros del salario mensual de trabajadores elegidos al azar en una empresa grande.

$$\{x_1, x_2, \dots, x_n\} \subseteq X \sim \text{Pareto}(x_m, \alpha)$$

3.1. Estimador de α por el método de momentos

El estimador del parámetro de forma α por el método de momentos, se obtiene a partir de la esperanza matemática o media aritmética de X , si se conoce el valor del parámetro de localización x_m , que es el pico y valor mínimo de la distribución. Dicha esperanza matemática corresponde al cociente entre la media muestral (numerador) y la diferencia entre la media muestral y el parámetro conocido x_m (denominador). Es válido cuando $\alpha > 2$.

$$E(X) = \frac{\alpha}{\alpha - 1} x_m$$

$$\alpha E(X) - E(X) - \alpha x_m = 0$$

$$\alpha(E(X) - x_m) = E(X)$$

$$\alpha = \frac{E(X)}{E(X) - x_m} = \frac{\mu(X)}{\mu(X) - x_m}$$

$$\hat{\alpha} = \frac{\hat{E}(X)}{\hat{E}(X) - x_m} = \frac{\hat{\mu}(X)}{\hat{\mu}(X) - x_m} = \frac{\bar{x}}{\bar{x} - x_m}, \text{ donde } \hat{\mu}(X) = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

El estimador puntual converge a una distribución normal, al tener X una distribución con momentos finitos (Rao, 1973), lo que permite obtener un estimador por intervalo asintótico.

$$\lim_{n \rightarrow \infty} \hat{\alpha}_n = \frac{\bar{x}}{\bar{x} - x_m} \xrightarrow{d} N\left(\alpha, \frac{\alpha(\alpha - 1)^2}{n(\alpha - 2)}\right)$$

$$P\left(\hat{\alpha} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\alpha(\alpha - 1)^2}{n(\alpha - 2)}} \leq \alpha \leq \hat{\alpha} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\alpha(\alpha - 1)^2}{n(\alpha - 2)}}\right) = 1 - \alpha$$

3.2. Estimadores de X_m y α por el método de máxima verosimilitud

El estimador por el método de máxima verosimilitud del parámetro de escala x_m es el valor mínimo muestral y el del parámetro de forma α es el inverso de la media aritmética de los logaritmos de los cocientes entre cada valor x_i y el mínimo muestral de los valores (Siudem, Nowak y Gagolewski, 2022).

$$\hat{x}_m = \min(x_1, x_2, \dots, x_n) = \min(\{x_i\}_{i=1}^n)$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^n \ln\left(\frac{x_i}{\hat{x}_m}\right)} = \frac{n}{\sum_{i=1}^n \ln\left(\frac{x_i}{\min(\{x_i\}_{i=1}^n)}\right)}$$

La información de Fisher para n datos de estos dos parámetros viene dada por la siguiente matriz cuadrada 2×2 :

$$I(x_m, \alpha | x) = n \begin{pmatrix} -E\left[\frac{\delta}{\delta x_m \delta x_m} \ln f_X(x | x_m, \alpha)\right] & -E\left[\frac{\delta}{\delta x_m \delta \alpha} \ln f_X(x | x_m, \alpha)\right] \\ -E\left[\frac{\delta}{\delta \alpha \delta x_m} \ln f_X(x | x_m, \alpha)\right] & -E\left[\frac{\delta}{\delta \alpha \delta \alpha} \ln f_X(x | x_m, \alpha)\right] \end{pmatrix}$$

$$= n \begin{pmatrix} \alpha/x_m^2 & -1/x_m \\ -1/x_m & 1/\alpha^2 \end{pmatrix} = \begin{pmatrix} n\alpha/x_m^2 & -n/x_m \\ -n/x_m & n/\alpha^2 \end{pmatrix}$$

La varianza de un parámetro θ , $Var(\theta)$, siempre es mayor o igual que su cuota inferior de Cramer y Rao, $CICR(\theta)$ o inverso de su información de Fisher para n datos, $1/I(\theta)$ (Xu, Sedory y Singh, 2022).

$$Var(\hat{\theta}) = \sigma_{\hat{\theta}}^2 \geq CICR(\hat{\theta}) = 1/I(\theta)$$

Los estimadores obtenidos por el método de máxima verosimilitud tienen como propiedades asintóticas la insesgidez, consistencia, eficiencia y normalidad, lo que los hacen muy útiles (Song, Ryoung-Park y Yoon, 2022).

$$\hat{\theta}_n \text{ es un estimador asintóticamente insesgado de } \theta, \text{ si } \lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$$

$$\hat{\theta}_n \text{ es un estimador asintóticamente consistente de } \theta, \text{ si } \lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

$\hat{\theta}_n$ es un estimador asintóticamente eficiente de θ , si $\lim_{n \rightarrow \infty} \frac{CICR(\hat{\theta})}{Var(\hat{\theta}_n)} = 1$

$\hat{\theta}_n$ como estimador θ converge a la normalidad, si $\lim_{n \rightarrow \infty} \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{DE(\hat{\theta})} \xrightarrow{d} N(0, 1)$

Desde estas propiedades asintóticas, se puede definir un error asintótico para los estimadores por máxima verosimilitud de x_n y α ; así como un intervalo de confianza asintótico.

$$EEA(\hat{\theta}) = \sigma_{\hat{\theta}} = \sqrt{1/I(\hat{\theta})}$$

$$P\left(\hat{\theta} - z_{1-\frac{\alpha}{n}}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{1-\frac{\alpha}{n}}\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

$$EEA(\hat{x}_m) = \sqrt{\frac{1}{n\alpha/x_m^2}} = \sqrt{\frac{x_m^2}{n\alpha}} = \frac{x_m}{\sqrt{n\alpha}}; \widehat{EEA}(\hat{x}_m) = \frac{\hat{x}_m}{\sqrt{n\hat{\alpha}}}$$

$$P\left(\hat{x}_m - z_{1-\frac{\alpha}{n}}\frac{\hat{x}_m}{\sqrt{n\hat{\alpha}}} \leq x_m \leq \hat{x}_m + z_{1-\frac{\alpha}{n}}\frac{\hat{x}_m}{\sqrt{n\hat{\alpha}}}\right) = 1 - \alpha$$

$$EEA(\hat{\alpha}) = \sqrt{\frac{1}{n/\alpha^2}} = \sqrt{\frac{\alpha^2}{n}} = \frac{\alpha}{\sqrt{n}}; \widehat{EEA}(\hat{\alpha}) = \frac{\hat{\alpha}}{\sqrt{n}}$$

$$P\left(\hat{\alpha} - z_{1-\frac{\alpha}{n}}\frac{\hat{\alpha}}{\sqrt{n}} \leq \alpha \leq \hat{\alpha} + z_{1-\frac{\alpha}{n}}\frac{\hat{\alpha}}{\sqrt{n}}\right) = 1 - \alpha$$

El estimador de α por el método de máxima verosimilitud es más eficiente que el del método de momentos, pero no es insesgado. Ante esta salvedad, se puede definir un estimador corregido de sesgo que tiene una varianza menor que el estimador sesgado (Rytgaard, 1990) y se puede usar para lograr una estimación asintótica más eficiente. Como las anteriores definiciones, el uso de estas fórmulas asintóticas requiere una muestra grande, a saber, un tamaño mayor que 30 y de preferencia de al menos 100 (Mateus y Caeiro, 2022).

$$\hat{\alpha}_c = \frac{n-1}{\sum_{i=1}^n \ln\left(\frac{x_i}{\min(\{x_i\}_{i=1}^n)}\right)}$$

$$\bar{V}ar(\hat{\alpha}_c) = \hat{\alpha}_c^2/n - 2 < Var(\hat{\alpha})$$

$$\widehat{EE}(\hat{\alpha}_c) = \sqrt{\hat{\alpha}_c^2/(n-2)} = \hat{\alpha}_c/\sqrt{n-2}$$

$$P\left(\hat{\alpha}_c - z_{1-\frac{\alpha}{n}}\frac{\hat{\alpha}_c}{\sqrt{n-2}} \leq \alpha \leq \hat{\alpha}_c + z_{1-\frac{\alpha}{n}}\frac{\hat{\alpha}_c}{\sqrt{n-2}}\right) = 1 - \alpha$$

3.3. Distribuciones exactas

Sea X una variable aleatoria con distribución de Pareto (x_m, α) . La distribución exacta de la suma de los logaritmos de cada uno de los n datos muestrales de X divididos por el valor mínimo muestral sigue una distribución exponencial de parámetro de tasa igual que α (Rytgaard, 1990).

$$\sum_{i=1}^n \ln\left(\frac{x_i}{\min(\{x_i\}_{i=1}^n)}\right) = \sum_{i=1}^n \ln\left(\frac{x_i}{\hat{x}_m}\right) \sim \text{Exponencial}(\lambda = \alpha)$$

La distribución exacta del estimador por máxima verosimilitud del parámetro de escala x_m es una distribución Pareto con parámetro de escala x_m y parámetro de forma $n\alpha$ y la distribución exacta del estimador por máxima verosimilitud de parámetro de forma α es una distribución gamma inversa con parámetro de forma $n - 1$ y parámetro de escala $n\alpha$ (Qian, Chen y He, 2021).

$$\hat{x}_m = \min(\{x_i\}_{i=1}^n) \sim \text{Pareto}(x_m, n\alpha)$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^n \ln\left(\frac{x_i}{\min(\{x_i\}_{i=1}^n)}\right)} \sim \text{Inv - Gamma}(\alpha = n - 1, \beta = n\alpha)$$

4. Relación con otras distribuciones e índice de concentración de Gini

Existe una forma generalizada de la distribución de Pareto de cuatro parámetros: de localización μ , escala σ y dos de forma (índice de cola de Pareto α e índice de desigualdad γ). Su función de distribución acumulativa corresponde a la siguiente expresión (Arnold, 2015):

Soporte: $x \geq \mu$

Espacio parametral: $\mu \in (-\infty, \infty)$ y σ, α y $\gamma \in (0, \infty)$.

$$F(X) = 1 - \frac{1}{\left[1 + \sqrt[\gamma]{\frac{x - \mu}{\sigma}}\right]^\alpha}$$

El parámetro de localización μ no es la esperanza matemática o media aritmética de la distribución y el parámetro de escala σ no es la desviación estándar. La media depende de los parámetros σ, α y γ y unas funciones gamma y requiere que $\gamma < \alpha$.

$$E(X) = \frac{\sigma \Gamma(\alpha - \gamma) \Gamma(1 + \gamma)}{\Gamma(\alpha)}, \gamma < \alpha$$

En la distribución de Pareto tipo I, $\mu = \sigma = x_m$ (mínimo y moda) y $\gamma = 1$.

$$F(X) = 1 - \frac{1}{\left[1 + \left(\frac{x - \mu}{\sigma}\right)^{1/\gamma}\right]^\alpha} = 1 - \frac{1}{\left[1 + \left(\frac{x - x_m}{x_m}\right)^{1/1}\right]^\alpha} = 1 - \frac{1}{\left(1 + \frac{x}{x_m} - 1\right)^\alpha} = 1 - (x_m/x)^\alpha$$

En la distribución de Pareto tipo II, $\gamma = 1$. Cuando el parámetro μ de la distribución de Pareto tipo II es igual que 0, recibe el nombre de distribución de Lomax (1954).

$$F(X) = 1 - \frac{1}{\left[1 + \left(\frac{x - \mu}{\sigma}\right)^{1/\gamma}\right]^\alpha} = 1 - \frac{1}{\left[1 + \left(\frac{x - \mu}{\sigma}\right)^{1/1}\right]^\alpha} = 1 - \frac{1}{\left[1 + \frac{x - \mu}{\sigma}\right]^\alpha}$$

En la distribución de Pareto tipo III, $\alpha = 1$.

$$F(X) = 1 - \frac{1}{\left[1 + \left(\frac{x - \mu}{\sigma}\right)^{1/\gamma}\right]^\alpha} = 1 - \frac{1}{\left[1 + \left(\frac{x - \mu}{\sigma}\right)^{1/\gamma}\right]^1} = 1 - \frac{1}{1 + \left(\frac{x - \mu}{\sigma}\right)^{1/\gamma}} = 1 - \frac{1}{1 + \sqrt[\gamma]{\frac{x - \mu}{\sigma}}}$$

Sea X una variable aleatoria con distribución de Pareto con parámetros x_m y α . El producto del parámetro de escala x_m y la exponencial con base el número de Euler ($e = 2.718281\dots$) yponente Y sigue una distribución exponencial de parámetro de tasa o inverso de escala $\lambda = \alpha$. A la inversa, sea Y una variable aleatoria con distribución exponencial de parámetro de tasa o inverso de escala λ . El logaritmo neperiano del cociente entre la variable aleatoria Y y el valor x_m sigue una distribución de Pareto tipo I de parámetro de escala x_m y parámetro de forma $\alpha = \lambda$. Esto implica que las funciones de distribución acumulativas de X e Y son iguales e intercambiables cuando se calculan las probabilidades acumuladas.

$$X \sim \text{Pareto}(x_m, \alpha); x_m e^X = Y \sim \text{Exponencial}(\lambda = \alpha)$$

$$Y \sim \text{Exp}(\lambda); \ln\left(\frac{Y}{x_m}\right) = X \sim \text{Pareto}(x_m, \alpha = \lambda)$$

$$P(X \leq x) = F_X(x|x_m, \alpha) = F_Y[y = \ln(x/x_m) | \lambda = \alpha] = 1 - (x_m/x)^\alpha = 1 - e^{-\alpha y}$$

Cabe señalar que entre las distribuciones exponencial y de Pareto hay una relación análoga a la existente entre las distribuciones normal y lognormal. Las distribuciones de Pareto y lognormal se pueden aplicar a los mismos datos, aunque una probablemente tenga mejor ajuste que la otra, y ambas son una transformación exponencial de una variable más ampliamente conocida y usada, la primera de una distribución exponencial y la segunda de una distribución normal (Feng et al., 2020).

La distribución zeta es el equivalente discreto de la distribución de Pareto. Si el soporte de la distribución zeta está acotado en un intervalo de números naturales constituye la denominada distribución de Zipf (Arnold, 2015).

$$X \sim Z(s)$$

$$s \in (1, \infty), x \in \mathbb{N}$$

$$P(X = x) = f_X(x) = \frac{1}{x^s \zeta(s)} = \frac{1}{x^s \sum_{n=1}^{\infty} 1/n^s}$$

$$P(X \leq x) = F_X(x) = \frac{\sum_{n=1}^x 1/n^s}{\zeta(s)} = \frac{\sum_{n=1}^x 1/n^s}{\sum_{n=1}^{\infty} 1/n^s}$$

$$E(X) = \frac{\zeta(s-1)}{\zeta(s)} = \frac{\sum_{n=1}^{\infty} 1/n^{s-1}}{\sum_{n=1}^{\infty} 1/n^s}; s > 2$$

$$E(X^k) = \frac{\zeta(s-k)}{\zeta(s)} = \frac{\sum_{n=1}^{\infty} 1/n^{s-k}}{\sum_{n=1}^{\infty} 1/n^s}; s > k + 1$$

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{\zeta(s)\zeta(s-2) - [\zeta(s-1)]^2}{[\zeta(s)]^2}; s > 3$$

$$Mo(X) = 1$$

El índice de cola de Pareto (1897) está relacionado con el índice de concentración de Gini (Safari *et al.*, 2019). Corrado Gini en 1936 creó este índice para resumir en un solo número la información contenida en la curva de Lorenz (Mojiri y Ahmadi, 2022). La curva de Lorenz (1905) se obtiene a partir de un diagrama bidimensional que representa el número de personas en el eje horizontal de abscisas (x) y el ingreso o la riqueza acumulada en el eje vertical o de ordenadas, $F(x)$. Dentro de esta gráfica se dispone una recta con una pendiente de 45 grados que se denomina línea de igualdad. En la medida que la curva definida por los pares coordenados $(x, F(x))$ se separa de esta recta, hay más desigualdad. El índice de Gini es la proporción del área entre la línea de igualdad y la curva de Lorenz y el área del triángulo recto inferior. Un valor de 0 indica total igualdad, y aparece cuando todos los individuos tienen el mismo ingreso; un valor de 1 refleja la máxima diferencia, y surge cuando una persona recibe todo el ingreso y las demás personas serían mano de obra no remunerada o sin ingresos (Sitthiyot y Holasut, 2021). En países con más igualdad de ingresos, como los escandinavos, el índice Gini es menor que 0.3. En los países con más desigualdad, como en algunos del continente africano (Sudáfrica, Namibia, Surinam, Zambia Eswatini, Botsuana, Angola y Zimbabue) y americano (Bélice, Brasil, Colombia y Panamá), el índice Gini es mayor que 0.5 (World Bank, 2022). Cabe señalar que los países con sistemas de economía comunista no poseen índices de Gini bajos y las políticas socialdemócratas que implican aumento del gasto y deuda públicos lo incrementan en lugar de disminuirlo (Tokhirov, 2021). Así, se ha denominado a las políticas populistas de izquierdas políticas de igualación en la pobreza con la emergencia de una oligarquía política (Aránguez-Sánchez, 2019; Landoni y Villegas, 2022).

El valor en la curva de Lorenz en el punto x para una variable aleatoria continua X con función de densidad $f_x(x)$ se obtiene mediante la siguiente proporción:

$$L(F(x)) = \frac{\int_{-\infty}^x xf(x) dx}{\int_{-\infty}^{\infty} xf(x) dx} = \frac{\int_{-\infty}^x xf(x) dx}{E(X)}$$

Aplicado a la distribución de Pareto, la expresión anterior queda del siguiente modo:

$$X \sim \text{Pareto}(x_m, \alpha)$$

$$F(x) = 1 - (x_m/x)^\alpha$$

$$L(F(x)) = 1 - (1 - F(x))^{1-\frac{1}{\alpha}} = 1 - \left[1 - \left(1 - \left(\frac{x_m}{x}\right)^\alpha\right)\right]^{1-\frac{1}{\alpha}} = 1 - \left(\frac{x_m}{x}\right)^{\alpha-1}$$

$$L(F(x)) = 0, 0 < \alpha < 1$$

El índice de Gini se puede expresar como uno menos el doble de la integral de 0 a 1 de la curva de Lorenz. Aplicado a la distribución de Pareto, queda como el inverso del doble del parámetro de cola menos uno, $1/(2\alpha-1)$, siempre que $\alpha > 1$. Cuando $\alpha < 1$, el índice de Gini toma su valor mínimo que es 1. En la Tabla 1, se muestra la correspondencia entre valores de los índices de cola de Pareto y concentración de Gini y su interpretación.

$$G(X) = 1 - 2 \int_0^1 L(F(x)) dx = 1 - 2 \int_0^1 \left(1 - \left(\frac{x_m}{x}\right)^{\alpha-1}\right) dx = \frac{1}{2\alpha-1}, \alpha \geq 1$$

$$G(X) = 1, 0 < \alpha < 1$$

$$\alpha = 0.5(1 + 1/G(X))$$

Tabla 1. Correspondencia entre los índices de Gini y Pareto y su interpretación

G	α	Interpretación
0.05	10.5	Muy poca
0.1	5.5	desigualdad
0.15	3.833	
0.2	3	Poca
0.25	2.5	desigualdad
0.3	2.167	
0.35	1.929	
0.4	1.75	Desigualdad
0.45	1.61	Bastante
0.5	1.5	desigualdad
0.55	1.409	
0.6	1.333	Mucha
0.65	1.269	desigualdad
0.7	1.214	
0.75	1.167	
0.8	1.125	
0.85	1.088	Muchísima
0.9	1.056	desigualdad
0.95	1.026	
1	1	

Nota. G = índice de concentración de Gini y α = índice de cola de Pareto.

Fuente: elaboración propia

5. Generación de muestras aleatorias paretianas y contraste de bondad de ajuste

Sea U una variable aleatoria con distribución uniforme estándar, la transformación $x_m(1-U)^{-1/\alpha}$ sigue una distribución de Pareto con parámetro de escala x_m y de forma α . Este procedimiento permite obtener una muestra aleatoria con distribución de Pareto a partir de una muestra aleatoria de una variable uniforme estándar y se denomina muestreo por transformada inversa (Gass y Fu, 2016).

Para comprobar si los datos muestrales se ajustan a la distribución de Pareto, Chu, Dickin y Nadarajah (2019) recomiendan la prueba de Kolmogorov y Smirnov con base en un estudio de simulación. La forma más sencilla para aplicar esta prueba inferencial es transformar los datos con supuesta distribución de Pareto a valores con una distribución exponencial a través del logaritmo neperiano del cociente entre cada dato y el valor mínimo muestral y , a continuación, aplicar la prueba de Kolmogorov y Smirnov adaptada a muestras exponencialmente distribuidas (Stephens, 1974). El contraste inferencial se puede complementar con una comprobación gráfica a través del diagrama cuantil-cuantil y el histograma. En el primero, se busca una alineación de puntos en una recta con pendiente de 45 grados y, en el segundo, un perfil con forma de J invertida (Bhoj y Chandra, 2021).

Hipótesis estadísticas: $H_0: X \sim \text{Pareto}(x_m, \alpha) \equiv Y = \ln[X/\min(X)] \sim \text{Exponencial}(\lambda = \alpha)$ y $H_1: X \not\sim \text{Pareto}(x_m, \alpha)$.

Supuestos: Muestra aleatoria de tamaño n de la variable cuantitativa continua X .

Estadístico de contraste: Se ordenan los datos en sentido ascendente, se les asignan rangos u órdenes y se transforman para que sigan una distribución exponencial bajo el supuesto de que la hipótesis nula es cierta.

$X_{(1)}$	$X_{(2)}$...	$X_{(i)}$...	$X_{(n-1)}$	$X_{(n)}$
1	2	...	i	...	$n-1$	n
$\ln[X_{(1)}/X_{(1)}]$	$\ln[X_{(2)}/X_{(1)}]$		$\ln[X_{(i)}/X_{(1)}]$		$\ln[X_{(n-1)}/X_{(1)}]$	$\ln[X_{(n)}/X_{(1)}]$

Se calcula la probabilidad acumulada para cada valor transformado y los estadísticos D^+ , D^- y D .

$$y_i = \ln[x_i/x_{(1)}] \in Y \sim \text{Exponencial} \left(\hat{\lambda} = \frac{n}{\sum_{i=1}^n \ln(x_i)} \right)$$

$$F_Y(y_{(i)}) = 1 - e^{-\hat{\lambda}y_{(i)}}$$

$$D^+ = \max[(i)/n - F_Y(y_{(i)})]$$

$$D^- = \max \left[F_Y(y_{(i)}) - \frac{(i) - 1}{n} \right]$$

$$D = \max(D^+, D^-)$$

Se aplica la corrección de estimación muestral al estadístico D (Stephens, 1974).

$$D_c = (D - 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n})$$

Se toma la decisión desde el estadístico D transformado. Si $D_c \leq D_\alpha$, se mantiene H_0 y si $D_c > D_\alpha$, se rechaza H_0 con un nivel de significación de α . Los valores críticos D_α dependen del nivel de significación α (Stephens, 1974), el cual usualmente se fija en 0.05 ($D_{0.05} = 1.094$). Con muestras pequeñas ($n < 20$), se puede subir a 0.10 ($D_{0.10} = 0.990$) y, con muestras grandes ($n > 500$), se puede bajar a 0.01 ($D_{0.01} = 1.308$).

Otra opción para el contraste inferencial es la prueba de Anderson y Darling (1952). Al igual que con el test de Kolmogorov y Smirnov se transforman los datos para que sigan una distribución exponencial y se aplica una corrección al estadístico de contraste cuando los parámetros son estimados desde la muestra (Stephens, 1986).

$$A^2 = -n - \sum_{i=1}^n \frac{2i-1}{n} [\ln(F_Y(y_i)) + \ln(1 - F_Y(y_{n+1-i}))]$$

$$A_c^2 = A \left(1 + \frac{0.6}{n} \right)$$

Como en la prueba de Kolmogorov y Smirnov, se toma la decisión desde el estadístico A^2 corregido. Si $A_c^2 \leq A_\alpha^2$ (valor crítico), se mantiene H_0 y si $A_c^2 > A_\alpha^2$, se rechaza H_0 con un nivel de significación de α . Los valores críticos A_α^2 dependen del tipo de distribución (exponencial), método de estimación (estimador de máxima verosimilitud) y nivel de significación α (Stephens, 1986), el

cual usualmente se fija en 0.05 ($D_{0.05} = 1.321$). Con muestras pequeñas ($n < 20$), el nivel de significación α se puede subir a 0.10 ($D_{0.10} = 1.062$) y, con muestras grandes ($n > 500$), se puede bajar a 0.01 ($D_{0.01} = 1.959$).

Existe una modificación de la prueba de Anderson y Darling para distribuciones con asimetría positiva y casos muy atípicos en la cola derecha, la cual fue desarrollada por Sinclair, Spurr y Ahmad (1990). La distribución de Pareto corresponde a este tipo de distribución, por lo que la prueba se aplica a la muestra original en caso de usarse. Su estadístico de contraste se denota por AU_n^2 .

$$AU_n^2 = \frac{n}{2} - 2 \sum_{i=1}^n F_X(x_{(i)}) - \sum_{i=1}^n \left(2 - \frac{2i-1}{n}\right) \ln(1 - F_X(x_{(i)}))$$

$$F_X(x_{(i)}) = 1 - (x_i/\hat{x}_m)^2 = 1 - (x_i/\min(x))^2$$

Cálculo del valor crítico para la toma de decisión:

$${}_pAU_n^2 = 1 - \frac{1}{1 + e^{G(p)}}$$

$$G(p) = 0.1170 - 0.03791t + 0.06318u + 0.09878tu + 0.009184t^2u - 0.00009742t^4u$$

$$t = \ln \frac{p}{1-p}, u = \frac{1}{1 + 0.3/\sqrt{n}}$$

Si $AU_n^2 \leq {}_pAU_n^2$, se mantiene $H_0: X \sim \text{Pareto}(x_m, \alpha)$ con el nivel de significación de p y $AU_n^2 > {}_pAU_n^2$, se rechaza. Cuando el tamaño muestral tiende a infinito, el valor de u se hace unitario y el valor crítico ${}_pAU_n^2$ alcanza su valor asintótico: ${}_pAU_n^2 = 0.356$ con $p = 0.1$, ${}_pAU_n^2 = 0.432$ con $p = 0.05$ y ${}_pAU_n^2$ es 0.610 con $p = 0.01$.

6. Ejemplos de cálculo con la distribución Pareto

6.1. Cálculo de probabilidades y medidas descriptivas

Instrucciones: Sea X una variable aleatoria con distribución de Pareto de parámetros $x_m = 500$ y $\alpha = 4.1$ que mide el salario quincenal en dólares en una empresa. Calcule la probabilidad de tener un ingreso menor o igual que 592, menor que 1000, entre 800 y 1400, mayor o igual que 600 y mayor que 1500. Compute la esperanza matemática o media aritmética $\mu(X)$, media geométrica $\mu_g(X)$, media armónica $\mu_h(X)$, mediana $Mdn(X)$ y moda $Mo(X)$ como medidas de tendencia central; la varianza $Var(X)$, desviación estándar $DE(X)$ y entropía $H(X)$ como medidas de variación; así como las medidas de la forma basadas en momentos centrales estandarizados de asimetría $\sqrt{\beta_1(X)}$ y exceso de curtosis $\beta_2(X)$. Obtenga el índice de concentración de Gini $G(X)$. Finalmente, represente su función de densidad $f_X(x)$ y distribución acumulativa $F_X(x)$.

$$X \sim \text{Pareto}(x_m = 500, \alpha = 4.1)$$

Probabilidad de tener un ingreso menor o igual que 592:

$$P(X \leq 592) = \int_{500}^{592} f_X(x) dx = 1 - (x_m/x)^\alpha = 1 - (500/592)^{4.1} = 0.4997$$

Probabilidad de tener un ingreso menor que 1000:

$$P(X < 1000) = \int_{500}^{1000} f_X(x) dx = 1 - (500/1000)^{4.1} = 0.9417$$

Probabilidad de tener un ingreso entre 800 y 1400:

$$P(800 \leq X \leq 1400) = P(X \leq 1400) - P(X \leq 800) = 0.9853 - 0.8544 = 1309$$

$$P(X \leq 1400) = 1 - (500/1400)^{4.1} = 0.9853$$

$$P(X \leq 800) = 1 - (500/800)^{4.1} = 0.8544$$

Probabilidad de tener un ingreso mayor o igual que 600:

$$P(X \geq 600) = (500/600)^{4.1} = 0.4735$$

Probabilidad de tener un ingreso mayor que 1500:

$$P(X > 1500) = (500/1500)^{4.1} = 0.0111$$

Medidas de tendencia central:

$$\mu(X) = E(X) = \frac{\alpha}{\alpha - 1} x_m = \frac{4.1}{3.1} \times 500 = 661.29$$

$$\mu_g(X) = e^{E(\ln(X))} = x_m e^{1/\alpha} = 500 \times e^{1/4.1} = 638.11$$

$$\mu_h(X) = H(X) = \frac{1}{E(1/X)} = \left(1 + \frac{1}{\alpha}\right) x_m = \left(1 + \frac{1}{4.1}\right) \times 500 = 621.95$$

$$Mdn(X) = Q_X(p = 0.5) = \sqrt[4]{2} \times x_m = 2^{1/4.1} \times 500 = 592.10$$

$$Mo(X) = \{x | \max(f_X(x))\} = x_m = 500$$

Medidas de variación:

$$Var(X) = \mu_2(X) = E[(X - E(X))^2] = x_m^2 \frac{\alpha}{(\alpha - 1)^2(\alpha - 2)} = 500^2 \frac{4.1}{3.1^2 \times 2.1} = 50790.35$$

$$DE(X) = \mu_2^{1/2}(X) = \sqrt{E[(X - E(X))^2]} = \frac{x_m}{\alpha - 1} \sqrt{\frac{\alpha}{\alpha - 2}} = \frac{500}{3.1} \sqrt{\frac{4.1}{2.1}} = 225.37$$

$$H(X) = E[-\ln(f_X(x))] = \ln \left[\frac{x_m}{\alpha} e^{1 + \frac{1}{\alpha}} \right] = \ln \left[\frac{500}{4.1} e^{1 + \frac{1}{4.1}} \right] = 6.05 \text{ nats}$$

Asimetría y exceso de curtosis basadas en momentos centrales:

$$\sqrt{\beta_1(X)} = \frac{\mu_3(X)}{\mu_2^{3/2}(X)} = \frac{2(\alpha + 1)}{\alpha - 3} \sqrt{\frac{\alpha - 2}{\alpha}} = \frac{2 \times 5.1}{1.1} \sqrt{\frac{2.1}{4.1}} = 6.64$$

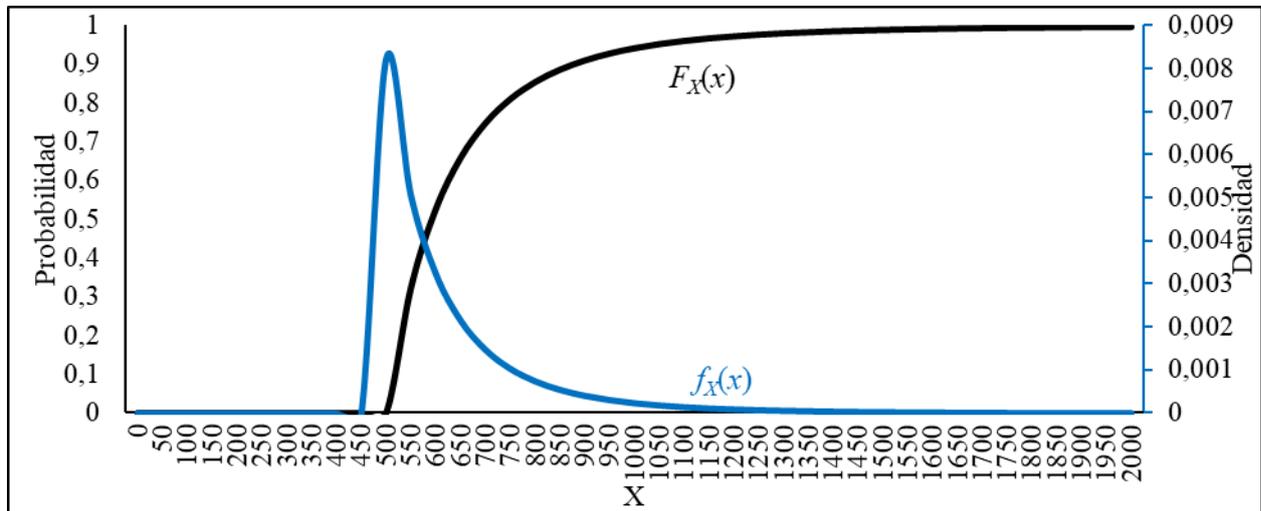
$$\beta_2(X) - 3 = \frac{\mu_4(X)}{\mu_2^2(X)} - 3 = \frac{6(\alpha^3 + \alpha^2 - 6\alpha - 2)}{\alpha(\alpha - 3)(\alpha - 4)} = \frac{6(4.1^3 + 4.1^2 - 6 \times 4.1 - 2)}{4.1 \times 1.1 \times 0.1} = 786.67$$

Índice de Gini:

$$G(X) = 1/(2\alpha - 1) = 1/(2 \times 4.1 - 1) = 0.1389$$

El índice de Gini está próximo a 0, lo que indica bastante igualdad en los salarios dentro de la empresa. En la Figura 2, se representa la función de densidad $f_X(x)$ y distribución acumulativa $F_X(x)$ de la variable aleatoria X de salario quincenal (en dólares) que sigue una distribución Pareto($x_m = 500$, $\alpha = 4.1$).

Figura 2. Función de densidad $f_X(x)$ y distribución acumulativa $F_X(x)$ de $X \sim \text{Pareto}(500, 4.1)$.



Fuente: elaboración propia.

6.2. Generación de muestra aleatoria, estimación de parámetros y bondad de ajuste

Instrucciones: Genere una muestra aleatoria de 40 datos desde la variable $X \sim \text{Pareto}(500, 4.1)$. Estime puntualmente los parámetros x_m y α , usando los estimadores de máxima verosimilitud, y por intervalos con un nivel de confianza al 95% mediante el error y distribución asintóticos. Finalmente, compruebe la bondad de ajuste con la gráfica de los cuantiles teóricos versus los cuantiles empíricos y el histograma, así como con las pruebas de Kolmogorov y Smirnov, y de Anderson y Darling.

En la primera columna de la Tabla 2, se dispone una muestra aleatoria de 40 datos u_i extraída de una distribución uniforme estándar $U \sim U[0, 1]$, para lo cual se usó el generador de números aleatorios de Excel versión 2021. A través de la función cuantil, los datos u_i se transforman en una muestra aleatoria con un tamaño de 40 de una variable X, la cual sigue una distribución de Pareto con parámetros poblacionales: $x_m = 500$ (de escala) y $\alpha = 4.1$ (de forma). Este proceso generativo de datos constituye el denominado muestreo por transformada inversa.

$$x_i = x_m(1 - u_i)^{-\frac{1}{\alpha}} = 500(1 - u_i)^{-\frac{1}{4.1}}$$

$$x_1 = 500(1 - 0.226)^{-\frac{1}{4.1}} = 532.239$$

Tabla 2. Generación de una muestra aleatoria y contraste del ajuste

Generación de X		Prueba de Kolmogorov y Smirnov					Gráfica C-C		
u_i	x_i	(i)	$x_{(i)}$	$y_{(i)}$	$F_Y(y_{(i)})$	D^+	D^-	$p_{(i)}$	$x_{t(i)}$
0.226	532.239	1	502.595	0	0	0.025	0	0.017	504.716
0.678	659.182	2	506.685	0.008	0.032	0.018	0.007	0.041	507.983
0.754	703.919	3	509.868	0.014	0.055	0.020	0.005	0.066	511.358
0.791	732.467	4	513.713	0.022	0.083	0.017	0.008	0.091	514.847
0.427	572.739	5	516.552	0.027	0.103	0.022	0.003	0.116	518.457
0.989	1502.032	6	519.474	0.033	0.123	0.027	-0.002	0.140	522.196
0.956	1071.113	7	524.641	0.043	0.156	0.019	0.006	0.165	526.073
0.356	556.651	8	529.753	0.053	0.188	0.012	0.013	0.190	530.097
0.651	646.362	9	532.239	0.057	0.203	0.022	0.003	0.215	534.278
0.592	622.201	10	540.797	0.073	0.252	-0.002	0.027	0.240	538.628
0.275	540.797	11	540.979	0.074	0.253	0.022	0.003	0.264	543.159
0.885	847.360	12	550.305	0.091	0.302	-0.002	0.027	0.289	547.886
0.921	928.627	13	551.504	0.093	0.308	0.017	0.008	0.314	552.824
0.524	599.242	14	556.651	0.102	0.333	0.017	0.008	0.339	557.991
0.145	519.474	15	560.951	0.110	0.353	0.022	0.003	0.364	563.406
0.726	685.652	16	569.140	0.124	0.389	0.011	0.014	0.388	569.093
0.053	506.685	17	571.286	0.128	0.398	0.027	-0.002	0.413	575.075
0.325	550.305	18	572.739	0.131	0.404	0.046	-0.021	0.438	581.384
0.421	571.286	19	588.961	0.159	0.466	0.009	0.016	0.463	588.050
0.578	617.102	20	598.325	0.174	0.498	0.002	0.023	0.488	595.114
0.489	588.961	21	599.242	0.176	0.501	0.024	0.001	0.512	602.620
0.021	502.595	22	609.166	0.192	0.533	0.017	0.008	0.537	610.619
0.924	937.437	23	617.102	0.205	0.556	0.019	0.006	0.562	619.175
0.331	551.504	24	622.201	0.213	0.570	0.030	-0.005	0.587	628.359
0.622	633.900	25	633.900	0.232	0.601	0.024	0.001	0.612	638.261
0.723	683.834	26	646.362	0.252	0.630	0.020	0.005	0.636	648.989

0.211	529.753	27	659.182	0.271	0.658	0.017	0.008	0.661	660.674
0.412	569.140	28	683.834	0.308	0.704	-0.004	0.029	0.686	673.482
0.077	509.868	29	685.652	0.311	0.707	0.018	0.007	0.711	687.625
0.276	540.979	30	703.919	0.337	0.736	0.014	0.011	0.736	703.373
0.521	598.325	31	721.765	0.362	0.761	0.014	0.011	0.760	721.090
0.822	761.718	32	732.467	0.377	0.775	0.025	0.000	0.785	741.265
0.125	516.552	33	761.718	0.416	0.807	0.018	0.007	0.810	764.590
0.971	1185.755	34	804.876	0.471	0.845	0.005	0.020	0.835	792.076
0.858	804.876	35	847.360	0.522	0.873	0.002	0.023	0.860	825.283
0.105	513.713	36	928.627	0.614	0.912	-0.012	0.037	0.884	866.783
0.376	560.951	37	937.437	0.623	0.915	0.010	0.015	0.909	921.249
0.555	609.166	38	1071.113	0.757	0.950	0.000	0.025	0.934	998.449
0.179	524.641	39	1185.755	0.858	0.967	0.008	0.017	0.959	1124.364
0.778	721.765	40	1502.032	1.095	0.987	0.013	0.012	0.983	1417.313
Σ			10.108						

Nota. Generación de X: u_i = muestra en su orden aleatorio extraída de una distribución uniforme estándar y $x_i = 500(1-u_i)^{-1/4.1}$ = datos transformados para seguir una distribución de Pareto(500, 4.1) mediante muestreo por transformada inversa. Transformación logarítmica de X y prueba de Kolmogorov y Smirnov: (i) = orden del dato x_i dentro de la muestra de 40 elementos, $x_{(i)}$ = datos x_i ordenados en sentido ascendente o cuantiles empíricos, $y_{(i)} = \ln(x_{(i)}/502.595)$ = transformación logarítmica de los datos x_i divididos por el mínimo muestral, $F_Y(y_{(i)})$ = función de distribución acumulativa de la variable Y que sigue una distribución exponencial con parámetro de tasa $\lambda = 3.957$, $D^+ = (i)/40 - F_Y(y_{(i)})$ = diferencia entre las funciones de distribución acumulativas empírica y teórica y $D^- = F_Y(y_{(i)}) - ((i)-1)/40$ = diferencia entre la función de distribución empírica teórica y la empírica con regazo de 1. Gráfica Cuantil-Cuantil: $p_{(i)} = ((i)-1/3)/(n+1/3)$ = orden del cuantil teórico y $x_{(i)} = Q_X[p_{(i)}]$ = cuantiles teóricos calculados mediante la función cuantil de una distribución de Pareto, dando a los parámetros sus valores estimados. Σ = suma por columna, Max = valor máximo por columna.

Fuente: elaboración propia.

A continuación, se estiman puntualmente los dos parámetros mediante los estimadores del método de máxima verosimilitud y por intervalo con los errores asintóticos.

$$\hat{x}_m = \min(\{x_i\}_{i=1}^{40}) = 502.595$$

$$\widehat{EEA}(\hat{x}_m) = \frac{\hat{x}_m}{\sqrt{n\hat{\alpha}}} = \frac{502.595}{\sqrt{40 \times 3.957}} = 39.948$$

$$P\left(\hat{x}_m - z_{1-\frac{\alpha}{n}} \frac{\hat{x}_m}{\sqrt{n\hat{\alpha}}} \leq x_m \leq \hat{x}_m + z_{1-\frac{\alpha}{n}} \frac{\hat{x}_m}{\sqrt{n\hat{\alpha}}}\right) = 1 - \alpha$$

$$P(502.595 - 1.96 \times 39.948 \leq x_m \leq 502.595 + 1.96 \times 39.948) = 0.95$$

$$P(x_m \in [424.299, 580.891]) = 0.95$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^n \ln\left(\frac{x_i}{\hat{x}_m}\right)} = \frac{n}{\sum_{i=1}^n \ln\left(\frac{x_i}{\min(\{x_i\}_{i=1}^n)}\right)} = \frac{40}{\sum_{i=1}^{40} \ln\left(\frac{x_i}{502.595}\right)} = \frac{40}{10.108} = 3.957$$

$$\widehat{EEA}(\hat{\alpha}) = \hat{\alpha}/\sqrt{n} = 3.957/\sqrt{40} = 0.626$$

$$P\left(\hat{\alpha} - z_{1-\frac{\alpha}{n}} \frac{\hat{\alpha}}{\sqrt{n}} \leq \alpha \leq \hat{\alpha} + z_{1-\frac{\alpha}{n}} \frac{\hat{\alpha}}{\sqrt{n}}\right) = 1 - \alpha$$

$$P(3.957 - 1.96 \times 0.626 \leq \alpha \leq 3.957 + 1.96 \times 0.626) = 0.95$$

$$P(\alpha \in [2.731, 5.184]) = 0.95$$

También se acude a la fórmula corregida de sesgo en la estimación del parámetro de forma α . En un principio, la estimación puntual de α parece menos exacta, pero la estimación por intervalo es más eficiente y adecuada.

$$\hat{\alpha}_c = \frac{n-2}{\sum_{i=1}^n \ln\left(\frac{x_i}{\hat{x}_m}\right)} = \frac{38}{\sum_{i=1}^{40} \ln\left(\frac{x_i}{502.595}\right)} = \frac{38}{10.108} = 3.759$$

$$\widehat{EE}(\hat{\alpha}_c) = \hat{\alpha}_c/\sqrt{n-2} = 3.759/\sqrt{38} = 0.610$$

$$P\left(\hat{\alpha}_c - z_{1-\frac{\alpha}{n}} \frac{\hat{\alpha}_c}{\sqrt{n-2}} \leq \alpha \leq \hat{\alpha}_c + z_{1-\frac{\alpha}{n}} \frac{\hat{\alpha}_c}{\sqrt{n-2}}\right) = 1 - \alpha$$

$$P(3.759 - 1.96 \times 0.610 \leq \alpha \leq 3.759 + 1.96 \times 0.610) = 0.95$$

$$P(\alpha \in [2.564, 4.955]) = 0.95$$

El contraste de la bondad de ajuste se inicia con las pruebas inferenciales. Los 40 datos con distribución de Pareto x_i se transforman para que sigan una distribución exponencial y_i y sobre esta muestra se aplica la prueba de Kolmogorov y Smirnov con el procedimiento de Stephens (1974). En la Tabla 2, se encuentra la transformación de los datos y los cálculos para obtener los estadísticos D^+ (diferencia máxima entre la función de distribución acumulativa empírica y la teórica) y D^- (diferencia máxima entre la función de distribución acumulativa teórica y la empírica con rezago de 1) y D (máximo de D^+ y D^-). Una vez transformado D , se observa que es menor que el valor crítico para un nivel de significación del 5%, por lo que se mantiene la hipótesis nula de bondad de ajuste. Los datos muestrales transformados y_i siguen una distribución exponencial y los datos sin transformar x_i siguen una distribución de Pareto.

$$H_0: X \sim \text{Pareto}(x_m, \alpha) \equiv Y = \ln\left(\frac{X}{\min(X)}\right) \sim \text{Exp}(\lambda = \alpha) \text{ y } H_1: X \not\sim \text{Pareto}(x_m, \alpha)$$

$$y_i = \ln[x_i/502.595] \in Y \sim \text{Exponencial}\left(\hat{\lambda} = n/\sum_{i=1}^{40} \ln(x_i) = 3.957\right)$$

$$F_Y(y_{(i)}) = 1 - e^{-\hat{\lambda} \times y_{(i)}} = 1 - e^{-3.957 \times y_{(i)}}$$

$$D^+ = \max[(i)/n - F_Y(y_{(i)})] = 0.046$$

$$D^- = \max[F_Y(y_{(i)}) - ((i) - 1)/n] = 0.037$$

$$D = \max(D^+, D^-) = \max(0.046, 0.037) = 0.046$$

Se transforma el estadístico de contraste D (Stephens, 1974)

$$D_c = \left(D - \frac{0.2}{n}\right) \left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right) = \left(0.046 - \frac{0.2}{40}\right) \left(\sqrt{40} + 0.26 + \frac{0.5}{\sqrt{40}}\right) = 0.275$$

$$D_c = 0.275 < D_{\alpha=0.05} = 1.094, \text{ se mantiene } H_0.$$

$$H_0: Y \sim \text{Exp}(\lambda) \Rightarrow X = 502.595e^Y \sim \text{Pareto}(x_m, \alpha)$$

Los datos y_i también se usan en la aplicación de la prueba de Anderson y Darling (1952). Se ordenan en sentido ascendente. Para el cálculo de estadístico A^2 , se requiere eliminar el primero, cuyo valor es 0, ya que le corresponde un valor de probabilidad acumulada de 0 y el logaritmo de esta probabilidad no es un número real. Para el cálculo de las probabilidades acumuladas, se estima el parámetro de tasa λ a través de su estimador de máxima verosimilitud que es el inverso de la media muestral de los 39 datos restantes (Tabla 3). El estadístico A^2 corregido es menor que el valor crítico para un nivel de significación del 5% (Stephens, 1986), por lo que se mantiene la hipótesis nula de que los datos $y_i = \ln(x_i/502.595)$ siguen una distribución exponencial y, por consiguiente, los datos $x_i = 502.595 \times e^{y_i}$ una distribución de Pareto.

$$\hat{\lambda} = \frac{39}{\sum_{i=1}^{39} y_{(i)}} = \frac{39}{10.108} = 3.858$$

$$F_Y(y_{(i)}) = \hat{\lambda} e^{-\hat{\lambda} y_{(i)}} = 3.858 e^{-3.858 y_{(i)}}$$

$$\sum_{i=1}^{39} A_i = \sum_{i=1}^{39} \frac{2i-1}{39} \left[\ln(F_Y(y_{(i)})) + \ln(1 - F_Y(y_{(n+1-i)})) \right] = -39.080$$

$$A^2 = -n - \sum_{i=1}^{39} A_i = -39 + 39.080 = 0.080$$

$$A_c^2 = A^2 \left(1 + \frac{0.6}{n}\right) = 0.080 \left(1 + \frac{0.6}{39}\right) = 0.081$$

$$A_c^2 = 0.081 < A_{0.05^2} = 1.321, \text{ se mantiene } H_0: Y = \ln(X/x_m) \sim \text{Exp}(\alpha); X \sim \text{Pareto}(x_m, \alpha)$$

Tabla 3. Cálculos para obtener el estadístico A^2 de Anderson y Darling desde los valores y_i

i	$y_{(i)}$	$(2i-1)/39$	$F_Y(y_{(i)})$	$\ln[F_Y(y_{(i)})]$	$y_{(n+1-i)}$	$F_Y[y_{(n+1-i)}]$	$\ln[1-F_Y(y_{(n+1-i)})]$	A_i
	0	0	0					
1	0.008	0.026	0.031	-3.481	1.095	0.985	-4.224	-0.198
2	0.014	0.077	0.054	-2.920	0.858	0.964	-3.312	-0.479
3	0.022	0.128	0.081	-2.514	0.757	0.946	-2.919	-0.697
4	0.027	0.179	0.100	-2.300	0.623	0.910	-2.405	-0.844
5	0.033	0.231	0.120	-2.123	0.614	0.906	-2.369	-1.037
6	0.043	0.282	0.153	-1.880	0.522	0.867	-2.015	-1.099
7	0.053	0.333	0.184	-1.694	0.471	0.837	-1.817	-1.170
8	0.057	0.385	0.198	-1.618	0.416	0.799	-1.604	-1.239
9	0.073	0.436	0.246	-1.402	0.377	0.766	-1.453	-1.244
10	0.074	0.487	0.247	-1.398	0.362	0.753	-1.396	-1.361
11	0.091	0.538	0.295	-1.220	0.337	0.727	-1.300	-1.357
12	0.093	0.590	0.301	-1.200	0.311	0.698	-1.198	-1.415
13	0.102	0.641	0.326	-1.122	0.308	0.695	-1.188	-1.481
14	0.110	0.692	0.345	-1.063	0.271	0.649	-1.046	-1.460
15	0.124	0.744	0.381	-0.965	0.252	0.621	-0.971	-1.439
16	0.128	0.795	0.390	-0.942	0.232	0.592	-0.896	-1.460
17	0.131	0.846	0.396	-0.927	0.213	0.561	-0.824	-1.481
18	0.159	0.897	0.458	-0.782	0.205	0.547	-0.792	-1.412
19	0.174	0.949	0.490	-0.714	0.192	0.524	-0.742	-1.381
20	0.176	1.000	0.493	-0.708	0.176	0.493	-0.679	-1.387
21	0.192	1.051	0.524	-0.647	0.174	0.490	-0.673	-1.387
22	0.205	1.103	0.547	-0.603	0.159	0.458	-0.612	-1.340
23	0.213	1.154	0.561	-0.578	0.131	0.396	-0.504	-1.248
24	0.232	1.205	0.592	-0.525	0.128	0.390	-0.494	-1.228
25	0.252	1.256	0.621	-0.476	0.124	0.381	-0.480	-1.201
26	0.271	1.308	0.649	-0.433	0.110	0.345	-0.424	-1.120
27	0.308	1.359	0.695	-0.364	0.102	0.326	-0.394	-1.030

28	0.311	1.410	0.698	-0.359	0.093	0.301	-0.358	-1.012
29	0.337	1.462	0.727	-0.318	0.091	0.295	-0.350	-0.977
30	0.362	1.513	0.753	-0.284	0.074	0.247	-0.284	-0.860
31	0.377	1.564	0.766	-0.266	0.073	0.246	-0.283	-0.859
32	0.416	1.615	0.799	-0.224	0.057	0.198	-0.221	-0.720
33	0.471	1.667	0.837	-0.177	0.053	0.184	-0.203	-0.634
34	0.522	1.718	0.867	-0.143	0.043	0.153	-0.166	-0.530
35	0.614	1.769	0.906	-0.098	0.033	0.120	-0.127	-0.399
36	0.623	1.821	0.910	-0.095	0.027	0.100	-0.106	-0.365
37	0.757	1.872	0.946	-0.055	0.022	0.081	-0.084	-0.262
38	0.858	1.923	0.964	-0.037	0.014	0.054	-0.055	-0.178
39	1.095	1.974	0.985	-0.015	0.008	0.031	-0.031	-0.091
Σ	10.108							-39.08

Nota. i = orden o rango del dato $y = \ln[x_i/\min(x)]$ cuando se orden en sentido ascendente, $y_{(i)}$ = valor y en el orden i , $(2i-1)/39$ = primer factor de A_i , $F_Y(y_{(i)})$ = probabilidad acumulada por el valor $y_{(i)}$ bajo un modelo de distribución exponencial de parámetro de tasa $\lambda = 3.858$, $\ln[F_Y(y_{(i)})]$ = logaritmo de la probabilidad acumulada por el valor $y_{(i)}$, $y_{(n+1-i)}$ = valor y en el orden $n+1-i$, $F_Y[y_{(n+1-i)}]$ = probabilidad acumulada por el valor $y_{(n+1-i)}$ bajo el mismo modelo de probabilidad, $\ln[1-F_Y(y_{(n+1-i)})]$ = logaritmo neperiano del complemento de la probabilidad acumulada por el valor $y_{(n+1-i)}$, $A_i = ((2i-1)/39)(\ln[F_Y(y_{(i)})]+\ln[1-F_Y(y_{(n+1-i)})])$, Σ = suma por columna.

Fuente: elaboración propia.

La prueba de Anderson y Darling modificada por Sinclair, Spurr y Ahmad (1990) para una muestra con asimetría positiva se aplica con los datos originales ordenados en sentido ascendente $x_{(i)}$. Se usa la función de distribución acumulativa de la distribución de Pareto, $F_X(x_{(i)}) = 1-(x_{(i)}/502.595)^{3.858}$; no obstante, las probabilidades son exactamente las mismas que las dadas por la función de distribución acumulativa de una distribución exponencial con parámetro de tasa $\lambda = 3.858$. El valor del estadístico de contraste AU_n^2 es menor que el valor crítico para un tamaño de muestra de 40 y un nivel de significación del 5%, por lo que se mantiene la hipótesis nula de ajuste a una distribución de Pareto. Véase a continuación los cálculos, parte de los cuales se hallan en la Tabla 4.

$$AU_n^2 = \frac{n}{2} - 2 \sum_{i=1}^n F_X(x_{(i)}) - \sum_{i=1}^n \left(2 - \frac{2i-1}{n}\right) \ln(1 - F_X(x_{(i)}))$$

$$AU_{40}^2 = \frac{40}{2} - 2 \times 19.636 - (-19.306) = 0.035$$

$$t = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{0.05}{0.95}\right) = -2.944$$

$$u = \frac{1}{1 + 0.3/\sqrt{n}} = \frac{1}{1 + 0.3/\sqrt{40}} = \frac{1}{1.0480} = 0.955$$

$$G(p) = 0.1170 - 0.03791t + 0.06318u + 0.09878tu + 0.009184t^2u - 0.00009742t^4u$$

$$G(p = 0.05) = 0.1170 - 0.03791 \times (-2.944) + 0.06318 \times 0.955 + 0.09878 \times (-2.944) \times 0.955 + 0.009184 \times (-2.944)^2 \times 0.955 - 0.00009742 \times (-2.944)^4 \times 0.955 = 0.080$$

$${}_pAU_n^2 = 1 - 1/[1 + e^{G(p)}]$$

$${}_{0.05}AU_{40}^2 = 1 - 1/[1 + e^{0.080}] = 0.520$$

$$AU_{40}^2 = 0.035 < {}_{0.05}AU_{40}^2 = 0.520, \text{ se mantiene } H_0: X \sim \text{Pareto}(x_m, \alpha)$$

Tabla 4. Cálculos para obtener el estadístico de contraste de la prueba de Anderson y Darling modificada por Sinclair Spurr y Ahmad (1990)

i	$x_{(i)}$	$F_X(x_{(i)})$	$2 - [(2i-1)/40]$	$\ln[1 - F_X(x_{(i)})]$	AU_i
1	502.595	0	1.975	0	0
2	506.685	0.031	1.925	-0.031	-0.060
3	509.868	0.054	1.875	-0.055	-0.104
4	513.713	0.081	1.825	-0.084	-0.154
5	516.552	0.100	1.775	-0.106	-0.188
6	519.474	0.120	1.725	-0.127	-0.220
7	524.641	0.153	1.675	-0.166	-0.277
8	529.753	0.184	1.625	-0.203	-0.330
9	532.239	0.198	1.575	-0.221	-0.348
10	540.797	0.246	1.525	-0.283	-0.431
11	540.979	0.247	1.475	-0.284	-0.419
12	550.305	0.295	1.425	-0.350	-0.499
13	551.504	0.301	1.375	-0.358	-0.493
14	556.651	0.326	1.325	-0.394	-0.522
15	560.951	0.345	1.275	-0.424	-0.540
16	569.140	0.381	1.225	-0.480	-0.588
17	571.286	0.390	1.175	-0.494	-0.581
18	572.739	0.396	1.125	-0.504	-0.567
19	588.961	0.458	1.075	-0.612	-0.658
20	598.325	0.490	1.025	-0.673	-0.690

21	599.242	0.493	0.975	-0.679	-0.662
22	609.166	0.524	0.925	-0.742	-0.686
23	617.102	0.547	0.875	-0.792	-0.693
24	622.201	0.561	0.825	-0.824	-0.680
25	633.900	0.592	0.775	-0.896	-0.694
26	646.362	0.621	0.725	-0.971	-0.704
27	659.182	0.649	0.675	-1.046	-0.706
28	683.834	0.695	0.625	-1.188	-0.743
29	685.652	0.698	0.575	-1.198	-0.689
30	703.919	0.727	0.525	-1.300	-0.682
31	721.765	0.753	0.475	-1.396	-0.663
32	732.467	0.766	0.425	-1.453	-0.618
33	761.718	0.799	0.375	-1.604	-0.602
34	804.876	0.837	0.325	-1.817	-0.590
35	847.360	0.867	0.275	-2.015	-0.554
36	928.627	0.906	0.225	-2.369	-0.533
37	937.437	0.910	0.175	-2.405	-0.421
38	1071.113	0.946	0.125	-2.919	-0.365
39	1185.755	0.964	0.075	-3.312	-0.248
40	1502.032	0.985	0.025	-4.224	-0.106
Σ		19.636			-19.306

Nota. i = orden o rango del dato muestral de X cuando los 40 datos se ordenan en sentido ascendente, $x_{(i)}$ = dato muestral de X en el orden o rango i , $F_X(x_{(i)})$ = probabilidad acumulada de $x_{(i)}$ bajo una distribución de Pareto ($x_m = 502.595$, $\alpha = 3.858$), $2 - [(2i-1)/40]$ = primer factor de AU_i , $\ln[1 - F_X(x_{(i)})]$ = logaritmo neperiano del probabilidad a la cola derecha de $x_{(i)}$ o segundo factor de AU_i , $AU_i = (2 - [(2i-1)/40]) \times \ln[1 - F_X(x_{(i)})]$, Σ = suma por columna.

Fuente: elaboración propia

Para obtener la gráfica cuantil-cuantil, se ordenan los datos generados en sentido ascendente $x_{(i)}$ y se asigna a cada dato $x_{(i)}$ su correspondiente orden (i). Con el orden del dato se calcula el orden del cuantil teórico $p_{(i)} = ((i+1/3)/(n+1/3))$, usando la fórmula correspondiente a la mediana de los estadísticos de orden de una distribución beta, $Mdn = (\alpha-1/3)/(\alpha+\beta-2/3)$, de parámetros de forma: $\alpha = i$ y $\beta = n+1-i$. Debe considerarse que el estadístico de orden i de una distribución uniforme estándar sigue dicha distribución beta. Para calcular los cuantiles teóricos, se usa la función cuantil dando como valores para los parámetros las estimaciones obtenidas (Tabla 2).

$$x_{(1)} = 502.595$$

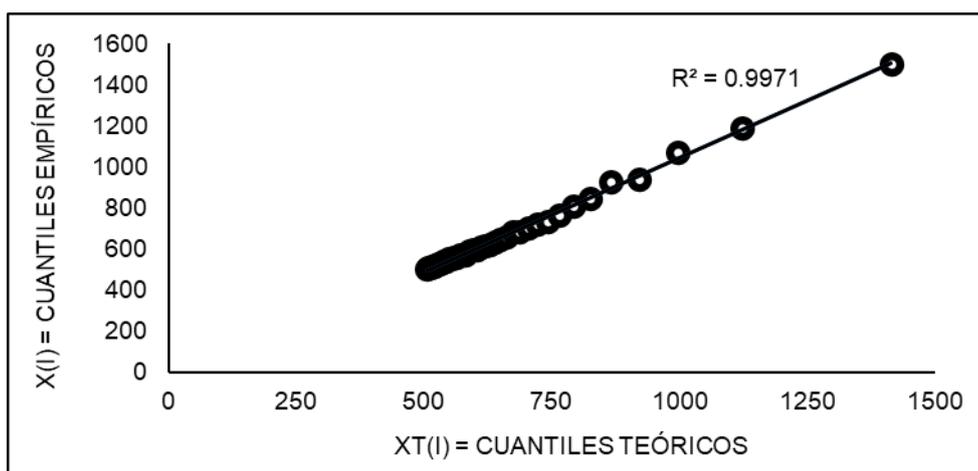
$$p_{(1)} = \frac{(i) - 1/3}{n + 1/3} = \frac{1 - 1/3}{40 + 1/3} = 0.017$$

$$x_{\tau_{(1)}} = \hat{x}_m (1 - p_{(i)})^{\frac{1}{\alpha}} = 502.595 (1 - 0.017)^{\frac{1}{3.957}} = 504.716$$

$$(x_{\tau_{(1)}} \quad x_{(1)}) = (502.595 \quad 504.716)$$

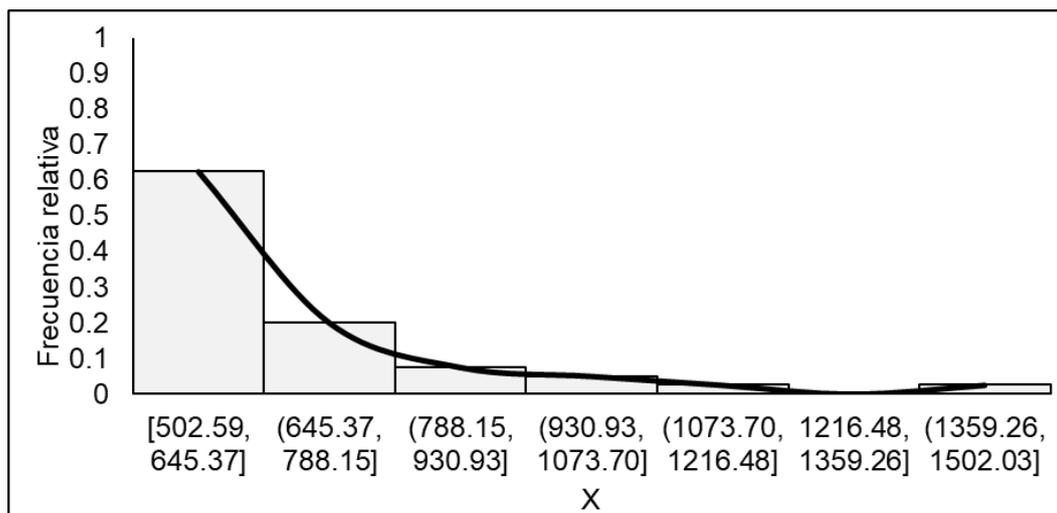
En la Figura 3, aparecen los pares coordenados de cuantiles teóricos $x_{(i)}$ y empíricos $x_{(i)}$. Los primeros se disponen en el eje horizontal o de abscisas porque predicen los cuantiles empíricos que se colocan en el eje vertical o de ordenadas. En la gráfica, los puntos coordenados se alinean a 45 grados, lo que indica que los datos empíricos siguen el modelo teórico. Esta alineación se puede cuantificar por el coeficiente de correlación lineal; a su vez, el cuadrado de la correlación informa de la varianza compartida. En este ejemplo, la correlación lineal es de 0.999 y la varianza compartida es del 99.7%. Adicionalmente, el histograma muestra el perfil de J invertido esperado para una muestra que sigue una distribución de Pareto (Figura 4), con lo que se puede concluir que los datos muestrales siguen un modelo de distribución de Pareto.

Figura 3. Gráfica cuantil-cuantil.



Fuente: elaboración propia.

Figura 4. Histograma de frecuencias relativas con el polígono de frecuencias superpuesto.



Fuente: elaboración propia.

7. Conclusiones

La presentación ejemplificada de la distribución de Pareto con parámetro de escala x_m y forma α permite constatar que el cálculo de probabilidades es muy simple, al igual que el cómputo de las medidas descriptivas. Como en otras distribuciones, los mejores estimadores de sus parámetros son los definidos mediante el método de máxima verosimilitud (Warsono et al., 2019), los cuales son sencillos y rápidos de calcular. El estimador de x_m es el mínimo muestral y el estimador de α es el inverso del promedio de los datos muestrales transformados logarítmicamente (Siudem et al., 2022). Naturalmente, se requiere contrastar el ajuste de los datos empíricos al modelo de probabilidad combinando una estrategia gráfica a través del histograma (forma de J invertida) y diagrama cuantil-cuantil (alineación a 45 grados) y una estrategia inferencial mediante las pruebas de Kolmogorov y Smirnov (Chu et al., 2019), Anderson y Darling (Stephens, 1986) o Anderson y Darling modificada (Sinclair et al., 1990). Con las dos primeras se transforman los datos para que sigan una distribución exponencial y con la tercera prueba no se requiere transformación.

Dentro de las Ciencias Sociales, esta distribución es un buen modelo de probabilidad para la distribución del ingreso, renta, recursos adquiridos acumulables en una empresa, región o país y reclamos de seguros (Feng et al., 2020), así como para la frecuencia promedio de un comportamiento raro en la mayoría de las personas, pero muy usual en unas pocas, como conductas compulsivas, adicciones comportamentales o a sustancias y parafilias (Rajeev, 2022).

Cabe señalar que se puede aplicar a los mismos datos de variables continuas que la distribución lognormal, como datos epidemiológicos (Beare y Toda, 2020), por lo que hay que determinar qué modelo de probabilidad tiene mejor ajuste (Charpentier y Flachaire, 2022; Feng et al., 2020). Cuando los datos son discretos en un conjunto acotado de números naturales, la opción es la distribución de Zipf, que es una variante de la distribución zeta. En cuanto al trasfondo social de lo expuesto, el buen ajuste de los datos a un modelo de probabilidad paretiano revela reglas de embudo y desigualdad muy definidas en reparto de recursos. A su vez, puede motivar políticas de segregación de un sector social responsable del disparo del gasto o la deuda, como ocurre hacia los adultos mayores con los seguros de salud (Granados-Martínez y Nava-Bolaños, 2019).

8. Bibliografía

- Ahmad, H. A. H. y Almetwally, E. M. (2020). Marshall-Olkin generalized Pareto distribution: Bayesian and non-Bayesian estimation. *Pakistan Journal of Statistics y Operation Research*, 16(1), 21–33. <http://dx.doi.org/10.18187/pjsor.v16i1.2935>
- Anderson, T. W. y Darling, D. A. (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23(2), 193–212. <http://dx.doi.org/10.1214/aoms/1177729437>
- Andria, J. (2022). A computational proposal for a robust estimation of the Pareto tail index: An application to emerging markets. *Applied Soft Computing*, 114, article 108048. <https://doi.org/10.1016/j.asoc.2021.108048>
- Aránguez-Sánchez, T. (2019). Luces y sombras del populismo. *Astrolabio: Revista Internacional de Filosofía*, 23, 39–53.
- Arnold, B. C. (2015). *Pareto distribution* (2nd ed.). New York: John Wiley y Sons, Ltd. <https://doi.org/10.1201/b18141>
- Barczy, M., Nedényi, F. K. y Sütő, L. (2023). Probability equivalent level of Value at Risk and higher-order Expected Shortfalls. *Insurance: Mathematics and Economics*, 108(1), 107–128. <https://doi.org/10.1016/j.insmatheco.2022.11.004>
- Barnoy, A. y Reich, Z. (2022). Trusting others: A pareto distribution of source and message credibility among news reporters. *Communication Research*, 49(2), 196–220. <https://doi.org/10.1177/0093650206298071>
- Beare, B. K. y Toda, A. A. (2020). On the emergence of a power law in the distribution of COVID-19 cases. *Physica D: Nonlinear Phenomena*, 412, article 132649. <https://doi.org/10.1016/j.physd.2020.132649>
- Bhoj, D. S. y Chandra, G. (2021). Ranked set sampling with lowest order statistics for Pareto distribution. *Communications in Statistics-Simulation and Computation*, 50, article 1904143. <https://doi.org/10.1080/03610918.2021.1904143>
- Campbell, M. R. y Brauer, M. (2021). Is discrimination widespread? Testing assumptions about bias on a university campus. *Journal of Experimental Psychology: General*, 150(4), 756–777. <https://doi.org/10.1037/xge0000983>
- Charpentier, A. y Flachaire, E. (2022). Pareto models for top incomes and wealth. *The Journal of Economic Inequality*, 20(1), 1–25. <https://doi.org/10.1007/s10888-021-09514-6>
- Chattamvelli, R. y Shanmugam, R. (2021). Pareto Distribution. In *Continuous Distributions in Engineering and the Applied Sciences-Part II* (pp. 179-188). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-02435-1_3
- Chen, B. Y., Zhang, K. Y., Wang, L. P., Jiang, S. y Liu, G. L. (2019). Generalized extreme value-pareto distribution function and its applications in ocean engineering. *China Ocean Engineering*, 33(2), 127–136. <https://doi.org/10.1007/s13344-019-0013-9>
- Chu, J., Dickin, O. y Nadarajah, S. (2019). A review of goodness of fit tests for Pareto distributions. *Journal of Computational and Applied Mathematics*, 361(17), 13–41. <https://doi.org/10.1016/j.cam.2019.04.018>
- Diawara, D., Kane, L., Dembele, S. y Lo, G. S. (2021). Applying of the extreme value theory for determining extreme claims in the automobile insurance sector: Case of a China car insurance. *Afrika Statistika*, 16(3), 2883–2909. <https://doi.org/10.16929/as/2021.2883.188>
- Fedotenkov, I. (2020). A review of more than one hundred Pareto-tail index estimators. *Statistica*, 80(3), 245–299. <https://doi.org/10.6092/issn.1973-2201/9533>
- Feng, M., Deng, L. J., Chen, F., Perc, M. y Kurths, J. (2020). The accumulative law and its probability model: an extension of the Pareto distribution and the log-normal distribution. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2237), article 2020.0019. <https://doi.org/10.1098/rspa.2020.0019>

- Gass, S. I. y Fu, M. C. (2016). Inverse transform method. In *Encyclopedia of operations research and management science* (pp. 815–815). Boston, MA: Springer. https://doi.org/10.1007/978-1-4419-1153-7_200343
- Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208, 73–79.
- Granados-Martínez, A. y Nava-Bolaños, I. (2019). Gastos catastróficos por motivos de salud y hogares con personas mayores en México. *Papeles de Población*, 25(99), 113–141. <https://doi.org/10.22185/24487147.2019.99.05>
- Landoni, J. S. y Villegas, L. (2022). *Pagan los pobres: Consecuencias negativas de políticas públicas con buenas (y malas) intenciones*. Buenos Aires: Editorial Galerna.
- Le Gall, P., Favre, A. C., Naveau, P. y Prieur, C. (2022). Improved regional frequency analysis of rainfall data. *Weather and Climate Extremes*, 36, article 100456. <https://doi.org/10.1016/j.wace.2022.100456>
- Lomax, K. S. (1954). Business failures. Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49(268), 847–852. <http://dx.doi.org/10.1080/01621459.1954.10501239>
- Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70), 209–219. <https://doi.org/10.2307/2276207>
- Mateus, A. y Caeiro, F. (2022). Confidence intervals for the shape parameter of a Pareto distribution. *AIP Conference Proceedings*, 2425, 320003. <https://doi.org/10.1063/5.0081541>
- McCarthy, D. M. y Winer, R. S. (2019). The Pareto rule in marketing revisited: is it 80/20 or 70/20?. *Marketing Letters*, 30(1), 139–150. <https://doi.org/10.1007/s11002-019-09490-y>
- Mojiri, A. y Ahmadi, K. (2022). Inequality in the distribution of resources in health care system by using the Gini coefficient and Lorenz curve (A case study of Sistan and Baluchestan province over a five-year period). *Payesh*, 21(3), 227–236. <https://doi.org/10.52547/payesh.21.3.227>
- Pareto, V. F. D. (1896). *Cours d'economie politique* (Vol. 1). Lausanne, Francia: F. Rouge éditeur.
- Pareto, V. F. D. (1897). *Cours d'economie politique* (Vol. 2). Lausanne, Francia: F. Rouge éditeur.
- Qian, W., Chen, W. y He, X. (2021). Parameter estimation for the Pareto distribution based on ranked set sampling. *Statistical Papers*, 62, 395–417. <https://doi.org/10.1007/s00362-019-01102-1>
- Rácz, E., Spasibko, K., Manceau, M., Ruppert, L., Chekhova, M. V. y Filip, R. (2023). Quantifying optical rogue waves. *arXiv*, article 2303.04615. <https://doi.org/10.48550/arXiv.2303.04615>
- Rajeev, C. D. S. (2022). Pareto principle and compulsive buying disorder - an analysis. *Journal of Educational and Social Research*, 8(2), 44–59.
- Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: John Wiley and Sons. <https://doi.org/10.1002/9780470316436>
- Rytgaard, M. (1990). Estimation in the Pareto distribution. *ASTIN Bulletin: The Journal of the IAA*, 20(2), 201–216. <https://doi.org/10.2143/AST.20.2.2005443>
- Safari, M. A. M., Masseran, N., Ibrahim, K. y Hussain, S. I. (2019). A robust and efficient estimator for the tail index of inverse Pareto distribution. *Physica A: Statistical Mechanics and its Applications*, 517(3), 431–439. <https://doi.org/10.1016/j.physa.2018.11.029>
- Sarabia, J. M., Jorda, V. y Prieto, F. (2019). On a new Pareto-type distribution with applications in the study of income inequality and risk analysis. *Physica A: Statistical Mechanics and its Applications*, 527, article 121277. <https://doi.org/10.1016/j.physa.2019.121277>
- Sinclair, C. D., Spurr, B. D. y Ahmad, M. I. (1990). Modified Anderson-Darling test. *Communications in Statistics - Theory and Methods*, 19(10), 3677–3686. <https://doi.org/10.1080/03610929008830405>

- Sitthiyot, T. y Holasut, K. (2021). A simple method for estimating the Lorenz curve. *Humanities and Social Sciences Communications*, 8(1), 1-9. <https://doi.org/10.1057/s41599-021-00948-x>
- Siudem, G., Nowak, P. y Gagolewski, M. (2022). Luck, reason, and the price-Pareto type-2 distribution. *Physica A: Statistical Mechanics and its Applications*, 606, article 128059. <https://doi.org/10.1016/j.physa.2022.128059>
- Song, I., Ryoung-Park, S. y Yoon, S. (2022). *Probability and random variables: theory and applications*. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-97679-8_3
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347), 730–737. <https://doi.org/10.2307/2286009>
- Stephens, M. A. (1986). Test based on EDF statistics. En R. B. D'Agostino y M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 97–193) New York: Marcel Dekker, Inc. <https://doi.org/10.1201/9780203753064-4>
- Sudharson, D. y Prabha, D. (2019). A novel machine learning approach for software reliability growth modelling with pareto distribution function. *Soft Computing*, 23(18), 8379–8387. <https://doi.org/10.1007/s00500-019-04047-7>
- Tokhirov, A. (2021). Remittances and inequality: the post-communist region. *Prague Economic Papers*, 30(4), 426–448. <https://doi.org/10.18267/j.pep.776>
- Warsono, Gustavia, E., Kurniasari, D., Amanto y Antonio, Y. (2019). On the comparison of the methods of parameter estimation for Pareto distribution. *Journal of Physics: Conference Series*, 1338(1) article 012042. <https://dx.doi.org/10.1088/1742-6596/1338/1/012042>
- World Bank (2022). *Gini index*. Recuperado de: <https://data.worldbank.org/indicator/>
- Xiong, B., Fu, H., Xu, F. y Jin, Y. (2019). Data-driven discovery of partial differential equations for multiple-physics electromagnetic problem. *arXiv*, article 1910.13531. <https://doi.org/10.48550/arXiv.1910.13531>
- Xu, T., Sedory, S. A. y Singh, S. (2022). Lowering the Cramer-Rao lower bounds of variance in randomized response sampling. *Communications in Statistics-Simulation and Computation*, 51(7), 4112–4126. <https://doi.org/10.1080/03610918.2020.1737874>
- Yang, X. y Zhou, P. (2022). Wealth inequality and social mobility: A simulation-based modelling approach. *Journal of Economic Behavior y Organization*, 196, 307–329. <https://doi.org/10.1016/j.jebo.2022.02.012>
- Zhang, Y., Wu, Y. y Yao, H. (2022). Optimal health insurance with constraints under utility of health, wealth and income. *Journal of Industrial and Management Optimization*, 18(3), 1519–1540. <https://doi.org/10.3934/jimo.2021031>

Autor.

José Moral de la Rubia

Facultad de Psicología de la Universidad Autónoma de Nuevo León, México.

Doctor en Filosofía y Ciencias de la Educación. Sección de Psicología por la Universidad de Alcalá Henares, Madrid. Psicólogo Especialista en Psicología Clínica vía PIR, Madrid. Licenciatura en Filosofía y Ciencias de la Educación. Sección de Psicología por la Universidad Pontificia Comillas, Madrid.

E-mail: jose.morald@uanl.edu.mx

Citado.

MORAL DE LA RUBIA, José (2025). Distribución de Pareto: un modelo de probabilidad en investigación social. *Revista Latinoamericana de Metodología de la Investigación Social – ReLMIS*, N°29, Año 15, pp. 22-50.

Plazos.

Recibido: 01/04/2023. Aceptado: 07/11/2023.



Uso de Procesamiento de Lenguaje Natural para procesar respuestas abiertas de una encuesta de Opinión Pública

Use of Natural Language Processing
to process open responses of a Public Opinion survey

Esteban Martínez Porras, Adrián Ramírez Fernández, Laura Solís Bastos y
José André Díaz-González

Resumen

Este artículo revisa la utilidad de utilizar Procesamiento de Lenguaje Natural (PLN) para el análisis de las respuestas abiertas brindadas a una encuesta de opinión pública. Se utilizan datos de una encuesta sobre percepción de la población costarricense respecto a diversos grupos migrantes y, a partir de ellos, se prueban diversos algoritmos, con el propósito de identificar cuál de ellos realiza una mejor clasificación de los datos. Se concluye que el algoritmo *random forest* es el que realiza una mejor clasificación automática de las respuestas, ayudando a disminuir la ambigüedad de la clasificación. El trabajo busca resaltar la utilidad que tiene el PLN para las investigaciones en Ciencias Sociales y, en especial, para el análisis de datos de preguntas abiertas aplicadas en estudios de opinión pública, ya que facilita la clasificación y análisis de gran cantidad de información no estructurada.

Palabras clave: Procesamiento de Lenguaje Natural; Minería de Texto; Opinión Pública; métodos de encuestas; *Machine Learning*.

Abstract

This article reviews the usefulness of using Natural Language Processing (NLP) for the analysis of open responses given to a public opinion survey. Data from a survey about the perception of the Costa Rican population regarding various migrant groups are used to test algorithms to identify which performs a better classification of the data. It is concluded that the random forest algorithm performs a better automatic classification of the answers, helping to reduce the ambiguity of the classification. The work seeks to highlight the usefulness of the PNL for research in Social Sciences and, especially, for the analysis of data from open questions applied in studies of public opinion, since it facilitates the classification and analysis of a large amount of unstructured information.

Keywords: Natural Language Processing; Text Mining; Public Opinion; survey methods, *Machine Learning*.

Introducción

Este artículo explora la pertinencia de utilizar técnicas de Procesamiento de Lenguaje Natural (PLN) para analizar las respuestas a preguntas abiertas obtenidas mediante la aplicación de una encuesta de opinión pública. Dado esto, el trabajo aquí presentado es en cierta forma exploratorio, ya que busca determinar la pertinencia de aplicar un método de análisis de uso común en el *Big Data* en una encuesta de opinión pública, con lo que se pretende generar un aporte metodológico que facilite el procesamiento y análisis de información recolectada en preguntas abiertas.

Para ello, el presente artículo se divide en seis secciones: en la primera se realiza una revisión de literatura académica sobre el uso de algoritmos en las ciencias sociales; en la segunda sección se realiza una presentación de la encuesta de la cual se toman los datos para realizar el presente análisis; en la tercera sección se explican los alcances y limitaciones del procesamiento de lenguaje natural para el análisis de textos; en la cuarta sección se expone el procedimiento metodológico seguido para el análisis de los datos; en la quinta sección se indican los resultados del análisis y, en la última sección, se encuentran las conclusiones.

Revisión de literatura académica sobre Algoritmos y Ciencias Sociales

Desde las Ciencias Sociales se han realizado diversos acercamientos hacia el uso de los algoritmos, sin embargo, prioritariamente la literatura académica se ha concentrado en los algoritmos y sus usos como fenómeno social, más que en identificar sus aportes en el desarrollo de procesos de investigación y análisis.

En el área de la comunicación social es una en la que se han desarrollado más investigaciones relacionadas al impacto de los algoritmos en el acceso y consumo de la información, especialmente en las redes sociales. Estos trabajos han mostrado que el uso de algoritmos que ayudan a promover información a partir de las reacciones emotivas de las personas usuarias tiende a facilitar la polarización de las opiniones (Fondevila Gascón, 2017). Otras investigaciones establecen que dada la implementación de algoritmos para seleccionar la información que difunden entre los usuarios, las redes sociales se han alejado de ser las plataformas de difusión o distribución de información que pretendían ser, y más bien han asumido un rol de editor de información, por las que decisiones que adoptan en el diseño de sus algoritmos deben ser discutidas a partir de la responsabilidad ética de dicho papel (Cetina Presuel & Martínez Sierra, 2019).

Otro grupo de trabajos se ha concentrado en estudiar la utilidad de los algoritmos para la toma de decisiones. Así, se ha determinado que es posible diseñar algoritmos que faciliten tomar decisiones en temas sociales complejos, o bien, que simplemente ayudan a simplificar las tareas cotidianas, lo cual permita a las personas concentrarse en tareas y problemas prioritarios (Zoran, 2016). También se ha investigado la utilidad de los algoritmos en los procesos de toma de decisiones públicas, así como en el diseño de políticas públicas; estos trabajos han mostrado que, si bien el uso de algoritmos puede ser útil para generar modelos de toma de decisiones ante situaciones complejas, al mismo tiempo reconocen que un diseño inadecuado de estos puede generar que ciertos grupos sean excluidos o no beneficiados de las decisiones adoptadas (Beer, 2017; Maxwell & Tomlinson, 2020).

Un tercer grupo de trabajos se han concentrado en identificar el valor de los algoritmos para el desarrollo de simulaciones, y han recalcado cómo estos han contribuido en el Modelamiento y Simulación Basado en Agentes (ABMS) para la comprensión y explicación de los fenómenos sociales, su emergencia, evolución y adaptación más que la predicción típica de los demás tipos de simulación (Díaz & Domínguez, 2013). También se ha destacado la utilidad que tiene el *Big Data* y el uso de algoritmos en los estudios de opinión pública, destacándose su aporte tanto en el proceso de recolección de información en redes o plataformas digitales, hasta sus aportes en los análisis de información y en la identificación de patrones de opinión (Cabrera-Álvarez, 2022; González, 2019; Gualda, 2022; Mamaqi et al., 2021; Porter et al., 2020; Tu et al., 2021).

Así, el uso de algoritmos en los análisis sociales y, especialmente, en los estudios de opinión pública dista de ser campo nuevo o emergente; sin embargo, posiblemente no puede considerarse aún que sea una herramienta de análisis consolidado en las disciplinas sociales. Desde las ciencias de la información se ha demostrado la utilidad de los algoritmos para clasificar y procesar información, y esto ha sido aprovechado por las ciencias sociales para generar modelos para la toma de decisiones y procesar datos de alta complejidad; no obstante, aún es necesario profundizar en los alcances y limitaciones que tienen los algoritmos como herramienta de investigación y, principalmente, promocionar y fomentar su uso entre las personas profesionales del área de ciencias sociales.

Debido a lo anterior, este trabajo busca realizar un aporte respecto al uso de algoritmos de aprendizaje automático (*Machine Learnig*) para procesar y analizar los datos recopilados en las preguntas abiertas aplicadas en una encuesta de opinión pública. Con esto, se espera mostrar como este tipo de algoritmo puede ser tanto una herramienta metodológica que facilita el trabajo de procesamiento de datos, como una herramienta analítica que facilite a las personas investigadoras la labor de comprender e interpretar la información recopilada.

Revisión de literatura académica referente a estudios de percepción sobre la población migrante en Costa Rica

El Instituto de Estudios Sociales en Población (IDESPO), en sus más de cuatro décadas de existencia ha abordado el fenómeno migratorio en el país, como un tema de relevancia para conocer el acontecer nacional y las percepciones de la población al respecto.

Cabe destacar que la labor que realiza este instituto en el marco del análisis de las percepciones de la población costarricense tiene un amplio reconocimiento a lo largo de su trayectoria, este tipo de estudios por lo regular se han desarrollado en la mayoría de los casos a través de encuestas telefónicas dirigidas a la población nacional y residente en el país.

A continuación, se detallarán algunos de los principales resultados de estudios sobre percepción acerca de la población migrante en Costa Rica, los que ejemplifican el desarrollo de material académico en esta línea.

- **Agosto, 2000.** La población costarricense de la gran área metropolitana frente a percepción hacia la población migrante, la política y los impuestos. Esta encuesta telefónica fue realizada a 400 personas, en este caso se aborda la percepción hacia la población migrante y en particular sobre la población nicaragüense.
- **Agosto, 2005.** Percepciones de la población costarricense sobre la inmigración de nicaragüenses. En esta encuesta telefónica el tamaño de la muestra fue de 600 y objetivo fue aportar a la opinión pública información referida a las percepciones de los costarricenses acerca de los inmigrantes.
- **Junio, 2006.** Se publica el informe de encuesta Identidades nacionales, integración y ciudadanía: percepciones hacia la inmigración. Presenta resultados de una encuesta nacional telefónica a hogares, la cual obtuvo una muestra de 600 personas mayores de edad. En esta oportunidad se abordan temas como los efectos de la migración sobre las instituciones sociales, políticas y económicas de nuestro país. Se plantea en específico como la inmigración es percibida como un “problema” de orden público, el cual necesita ser controlado mediante la restricción del acceso de inmigrantes a nuestro país.
- **Octubre, 2008.** Se publica el informe de encuesta Percepciones y actitudes de la población costarricense hacia la inmigración nicaragüense y la emigración de costarricenses al exterior. Corresponde a datos de una encuesta nacional telefónica a hogares, la cual obtuvo una muestra de 600 personas mayores de edad. En esta oportunidad se abordan temas como las percepciones acerca de Costa Rica como país receptor de población migrante (nicaragüenses, colombianos, indígenas), así como la construcción de esa visión de las otredades.

- **Agosto, 2014.** Se publica el informe de encuesta Construcción de opiniones públicas sobre la migración en Costa Rica. En este estudio, que empleó como medio de contacto la telefonía fija, se encuestaron 1000 personas y se valoraron las percepciones respecto a nacionalidades específicas de la población migrante residente en el país o en tránsito.
- **Abril, 2016.** Se publica el informe de encuesta Percepciones acerca de las relaciones entre Costa Rica y Nicaragua. Se realizó vía telefónica, a una muestra de 800 personas, mayores de edad y costarricenses. Tuvo por objetivo determinar las percepciones y opiniones de la población costarricense respecto a las relaciones entre Costa Rica y Nicaragua, a partir de su opinión sobre coyunturas o acontecimientos acaecidos en los meses anteriores en los que intervienen o se ven afectados ambos países.
- **Setiembre, 2019.** El informe percepciones de la población nacional sobre las migraciones, convivencia e integración en Costa Rica, muestra los resultados de una encuesta telefónica que empleó como método de contacto líneas celulares activas en el país, dirigida a personas costarricenses por nacimiento, de 18 años o más, la cual tuvo un alcance de 1002 personas encuestadas. Esta encuesta tuvo como objetivo determinar las percepciones de la población nacional acerca de las dinámicas de integración y convivencia de la población inmigrante en Costa Rica.

Es así como resulta interesante comprender de qué forma los estudios de percepción mediante encuestas son un insumo de gran relevancia para la investigación social, lo cual ha sido visible a través del tiempo para abordar temáticas específicas tales como las migraciones en Costa Rica.

Los estudios de percepción acerca de las migraciones en Costa Rica han permitido evidenciar como se construyen las opiniones en torno a la población migrante, en este sentido se ha observado que hay múltiples factores que pueden influir en la manera en la que forman estas percepciones, tales como la experiencia personal, la información publicada en medios de comunicación, los discursos políticos y políticas migratorias, estereotipos culturales, entre otros.

En el caso de la experiencia personal se observa que las interacciones individuales con personas migrantes podrían llegar a influir en la forma en que se percibe a esta población. Si alguien ha tenido experiencias positivas con migrantes, es más probable que tenga una percepción favorable hacia ellos. Por el contrario, las experiencias negativas pueden generar percepciones negativas.

Por otra parte, los medios de comunicación desempeñan un papel importante en la construcción de las percepciones sobre la migración. La forma en que se presentan las noticias y los reportajes sobre migrantes podría llegar a influir en cómo se percibe a esta población. Si los medios enfatizan los aspectos negativos o sensacionalistas exaltando el fenómeno migratorio o la nacionalidad de la persona migrante, esto puede contribuir a la formación de percepciones negativas.

Además, en el caso de discursos políticos y las políticas migratorias pueden tener un impacto significativo en la formación de percepciones sobre la población migrante. Dependiendo de cómo los líderes políticos aborden el tema de la migración, se podría generar percepciones positivas o negativas. Si se utiliza un lenguaje que estigmatiza a los migrantes o se promueven políticas restrictivas, esto puede influir en la forma en que se percibe a esta población.

Finalmente, los estereotipos culturales también podrían influir en la forma en que se percibe a los migrantes. Los estereotipos negativos o la falta de conocimiento sobre las culturas de origen de los migrantes pueden llevar a construir percepciones negativas. Por otro lado, si existe una mayor familiaridad y comprensión de las diferentes culturas, esto podría contribuir a la integración y la convivencia.

Es importante tener en cuenta que las percepciones sobre la población migrante pueden ser diversas. Además, estas percepciones pueden llegar a cambiar con el tiempo a medida que se

desarrollan condiciones sociopolíticas que favorezcan nuevas visiones en referencia a este fenómeno o la población migrante.

Las percepciones de la población costarricense sobre la población migrante

En el año 2012, desde el Programa Migraciones, Cambio Social e Identidades adscrito al IDESPO-UNA, se llevó a cabo la encuesta “Construcción de opiniones públicas sobre la migración en Costa Rica”, los datos fueron recolectados en el mes de mayo. Esta encuesta fue realizada a través de telefonía fija a 1000 personas costarricenses, mayores de 18 años y residentes del hogar contactado, contó con un error de muestreo del 3,1% y un nivel de confianza del 95%. Estos datos han sido retomados en la actualidad con el propósito de analizarlos desde una lectura distinta, empleando algoritmos de aprendizaje automático para clasificar y procesar información, de forma tal que pueda compararse los resultados obtenidos en el año 2012, y así valorar como el uso de un algoritmo puede llegar a facilitar su procesamiento.

En el marco de esta encuesta se aplicó un módulo de preguntas referentes a la percepción sobre distintas poblaciones migrantes en Costa Rica; primero se consultó: A continuación, me gustaría que me indique qué piensa de algunos grupos y de forma posterior se preguntaba ¿Por qué creé eso? Dentro de los grupos en consulta se encontraban: haitianos, indígenas panameños, colombianos, nicaragüenses, dominicanos, españoles, africanos, estadounidenses y chinos. El objetivo de esta pregunta se asociaba a valorar las posibles caracterizaciones acerca de la población migrante en Costa Rica, a partir de los imaginarios sociales que se construyen vinculados a las nacionalidades de las personas migrantes.

La migración es un fenómeno de gran relevancia para Costa Rica debido a la complejidad de su contexto migratorio. El país se caracteriza principalmente por ser receptor de migrantes, con la comunidad nicaragüense como el grupo mayoritario dentro de su territorio. En años recientes, además, ha adquirido un papel significativo como país de tránsito para personas migrantes que se desplazan, en su mayoría, hacia los Estados Unidos. Finalmente, aunque en menor medida, Costa Rica también es un país de origen de migrantes, siendo precisamente los Estados Unidos el principal destino elegido por los costarricenses que deciden emigrar.

Comprender las percepciones sobre las migraciones es un reto de complejidades, esto debido a que el fenómeno se encuentra presente dentro de la cotidianidad de la población, enmarcado en las condiciones en cuales se da esta movilidad poblacional, tanto regulares como irregulares. Además, esta construcción de las percepciones puede tender en algunos casos a estar marcada por mitos y estereotipos en torno a estos grupos poblacionales, tal como lo manifiesta Delgado Montaldo (2008) en referencia a Goffman:

De acuerdo con el estigma se manifiesta como una actitud negativa, discriminatoria, que se dirige siempre contra un individuo o un grupo al que se le considera inferior. En este sentido, aunque en ocasiones la víctima pueda ser una persona aislada, la principal razón por la cual se le estigmatiza, se le discrimina o se le evalúa negativamente, es porque pertenece a un determinado grupo (los 'nicas' o nicaragüenses, los 'paisas' o colombianos, los 'nochis' o chinos, en el caso de poblaciones inmigrantes, o bien, los afrocaribeños y los indígenas, quienes también son estigmatizados) (Delgado Montaldo, 2008: 86-87).

Dentro de las características que se mencionan de forma más frecuente, es común identificar algunas asociadas a elementos positivos, otras a elementos negativos y algunas poseen carácter ambivalente. También se observa que, según la nacionalidad, algunas surgen únicamente para grupos específicos, mientras que otras características se asocian a condiciones que fueron relacionadas al país de origen, incluso a la percepción sobre la construcción del imaginario de la nacionalidad, lo que plasma una diversidad de respuestas sobre la población migrante y muestra la complejidad que se requiere para el procesamiento de este tipo de datos (ver Figura 1).

Figura 1. Nubes de palabras asociadas a las caracterizaciones de población migrante en Costa Rica. 2012.



Fuente: Construcción propia, a partir de datos de IDESPO-UNA (2012).

De acuerdo con lo anterior, destaca esa visión diferenciadora que se refleja en el marco de las percepciones mencionadas por la población encuestada, y que trazan líneas para quién es denominado como migrante según su país de procedencia, tal como lo destaca Sandoval García (2004):

“Inmigración” se ha convertido en un concepto de “sentido común” que requiere ser discutido críticamente. Por ejemplo, europeos o norteamericanos que invierten en actividades turísticas en Costa Rica podrían ser considerados “inmigrantes”, pues han abandonado su país y residen en una nueva nación. Sin embargo, se les conoce como “inversionistas”, “pensionados” o “turistas”. Así, “inmigrante” es un término cuyo empleo es altamente selectivo, reservado para aquellos grupos considerados, en uno u otro sentido, como “conflictivos” (Sandoval García, 2004: 157).

Por otra parte, cabe destacar que las percepciones se construyen desde elementos subjetivos que corresponden a un momento determinado en el tiempo y que estas pueden ir variando según elementos que puedan incidir de forma significativa. Esto se refleja en las respuestas brindadas a la pregunta *¿Por qué creé eso?*, que procuran obtener esa explicación en torno a los calificativos otorgados a la población migrante según su nacionalidad.

Es así como se identifica que de las respuestas de mayor recurrencia destacan los medios de comunicación como una fuente importante de información sobre población migrante, asimismo, la experiencia personal basada en la convivencia con personas migrantes en diferentes espacios sociales, y rasgos estereotipados asociados por la población encuestada a la nacionalidad de la persona migrante o al país de origen.

Dado lo anterior, es que los datos recolectados en la pregunta abierta *¿Por qué cree eso?* Se someten a análisis utilizando PNL ya que, debido a la gran cantidad y diversidad de respuestas recibidas, aplicar esta metodología resulta valiosa para valorar tanto la existencia de

correspondencia entre las razones expresadas por la población encuestada y su percepción sobre los diversos grupos de población migrante sobre los cuales se consulta; asimismo, utilizar PNL aporta elementos analíticos adicionales que permitan a las personas investigadoras tener una mejor comprensión de las razones e imaginarios de la población costarricense que sustentan sus opiniones y percepciones en relación a estos grupos.

Análisis de texto usando el Procesamiento de Lenguaje Natural

Para el análisis de datos de estudios de opinión resulta apropiado utilizar un conjunto de técnicas que hacen uso de algoritmos de aprendizaje automático o *machine learning*; el cual trabaja sobre documentos de texto, y considera tanto su estructura interna y la distribución de las palabras para la codificación y categorización, las cuales se conoce como Procesamiento de Lenguaje Natural (PNL por sus siglas en inglés) (Bonaccorso, 2017). Así, el PNL permite analizar, comprender y entender el significado de las palabras a través de una computadora que realiza el proceso de manera automática. Como señalan Tintinago *et al.* (2018), el PNL es un campo de estudio que se enfoca en las interacciones entre el lenguaje humano y los ordenadores. Para esto utiliza como herramienta la inteligencia artificial.

Así, se denomina como NLP a un conjunto de técnicas de “*machine learning*” que trabaja sobre documentos de texto, considerando su estructura interna y la distribución de las palabras para la codificación y categorización (Bonaccorso, 2017). Es decir, el PNL permite analizar, comprender y entender el significado de las palabras a través de una computadora realizando el proceso de manera automática.

Una de las aplicaciones que tiene el NLP es la extracción de información de textos, la cual, según Moreira *et al.* (2021) “consiste en la obtención de partes que son importantes en el contenido para pasarlos a una base de datos llenos sobre un tema específico” (p. 130); lo cual facilita a grupo de investigadores en diversas disciplinas sistematizar y clasificar datos relevantes de información textual, que posteriormente puede servir para la interpretación, comprensión o búsqueda de significados para un mejor entendimiento del pensamiento humano.

Además, el NLP busca precisión y comprensión de los significados de acuerdo con consensos sociales de determinada lengua, por medio de algoritmos que hacen una valoración adecuada de las palabras. Según Fernández (2012: 3) el PNL “ocupa de la formulación e investigación de mecanismos eficaces para la comunicación entre personas, o entre personas y máquinas por medio de lenguajes de comunicación humana”.

Una de las ventajas de utilizar NLP es que minimiza las interpretaciones erróneas en el lenguaje. Es así como los modelos de inteligencia artificial aplicados a los lenguajes naturales, no solo se centran en la comprensión de los lenguajes sino en aspectos sobre el pensamiento humano y la organización de la información. En el caso de información presente en un documento de texto, o en un mensaje textual, el NLP sirve para facilitar la comprensión de este, dado que los datos en texto tienen mucha información, pero esta no está estructurada. Mediante un algoritmo de inteligencia artificial se puede preprocesar y convertir toda esa información en datos numéricos, además se puede extraer la semántica y significado del contenido.

Para el análisis de texto utilizando NLP, se aplica un método de *tokenización* de documentos, que consiste en separar palabra por palabra, en la cual se crean una serie de parámetros que se obtienen de un vocabulario común para crear su propia “bolsa de palabras”. Para ello se utilizan unidades semánticas que son palabras o grupos de palabras que se conocen como tokens. Con estos tokens se forman vectores y luego una matriz que finalmente se usa como entrada para los algoritmos de clasificación como por ejemplo *Naybe Bayes* el más utilizado para esto (Jansen, 2018).

De acuerdo con Fernández (2004), este modelo de disgregación del texto establece un procedimiento analítico que no se centra en unidades textuales aisladas, sino en el valor significativo social del texto en su totalidad; esto se logra a partir del establecimiento del corpus y los datos de entrenamiento y test que se ejecutan en el algoritmo.

La relevancia de la dimensión social del texto se origina desde el sentido práctico de las palabras, “ya que el procesamiento del significado dependerá de las condiciones no lingüísticas en las que se produce la instancia no comunicativa” (Fernández, 2004: 4). Es decir, el procesamiento del lenguaje a través del análisis textual deberá responder a la forma que se produce la convencionalización del significado del texto y esto se logra mediante la transformación del vocabulario en vectores que pueden ser fácilmente usados para clasificar y agrupar las palabras.

Como parte del proceso del algoritmo de PNL se remueven las *stopword* que son palabras de uso común en el habla como: artículos, conjunciones, preposiciones, adverbios, que no proveen una información semántica relevante para el análisis. Además, se realiza una simplificación de las palabras, al hacer una transformación de palabras a verbos o raíces, por ejemplo: El *ama* se transforma al verbo *amar*.

Este conjunto de reglas se aplica varias veces para transformar el texto etiquetado en oraciones que definen la asociación entre una palabra y una parte del habla con un sentimiento calificado. Para la implementación se usan herramientas para etiquetar y una base de datos con claves / frases con evaluaciones de polaridad de emociones” (Hernández y Gómez, 2013: 90).

Por consiguiente, el uso de algoritmos de PNL no es solo para clasificar palabras, sino que permite realizar un análisis más profundo del texto, como señalan Almela *et al.* (2012), la gramática en el lenguaje tiene una relevancia en la lingüística computacional sino también en la psicología, donde “las variaciones en su uso pueden aportar información valiosa sobre el estado mental de la persona, su edad, su sexo, su estatus social o la condición de verdad de su discurso” (p. 70). Es decir, un algoritmo de NLP puede facilitar a los investigadores en el área de lenguaje y psicología la clasificación e identificación de patrones de frases o palabras que permiten caracterizar a los individuos que las escriben.

Aunque actualmente la *minería de opiniones* (término que hace referencia al análisis de sentimientos por medio de textos de manera digital), está orientada más en áreas de mercadeo y manejo de imagen (Hernández y Gómez, 2013), estas técnicas de PNL pueden favorecer diversas áreas sociales para la comprensión del pensamiento humano.

De acuerdo con Bonaccorso (2017) el proceso NLP se puede resumir en las siguientes 3 fases:

- 1) Limpiar textos y prepararlos para aplicar los algoritmos de ML o solo quedarse con la palabra más relevante.
- 2) Crear el modelo de *Bag of words* (saco de palabras), que serán las palabras claves a clasificar por medio del algoritmo más apropiado.
- 3) Aplicar el modelo de ML, más apropiado al *Bag of words*.

Tras completar los pasos anteriores, se realiza el análisis del modelo aplicado usando la matriz de confusión, la cual muestra cómo es la precisión del algoritmo de clasificación que se utiliza. Esta matriz se compone de las siguientes entradas:

- Verdadero positivo (TP): Una muestra positiva correctamente clasificada.
- Falso positivo (FP): Una muestra negativa clasificada como positiva.
- Verdadero negativo (TN): Una muestra negativa correctamente clasificada.
- Falso negativo (FN): Una muestra positiva clasificada como negativa.

Los datos obtenidos de esta matriz se pueden mejorar realizando algunos cálculos o modificaciones. De acuerdo con Raschka y Mirjalili (2017) tenemos dos medidas generales que son el *error de predicción* (ERR) y la *exactitud* (ACC) que proporcionan información general acerca de cuántas muestras se clasifican erróneamente. Así que el ERR se puede entender como el ponderado de la suma de predicciones falsas con todas las predicciones. La fórmula es:

$$ERR = \frac{FP + FN}{TN + TP + FN + FP}$$

Por su parte, la exactitud es un cociente que se calcula como la suma de las predicciones correctas divididas por el número total de predicciones, como se muestra en la fórmula:

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} = 1 - ERR$$

La exactitud (o *accuracy*) muestra el porcentaje de predicciones correctas (verdaderos positivos y verdaderos negativos) frente al total de predicciones. Es útil para datos equilibrados. Además, se cuenta con otras métricas como son: la tasa de verdadero positivo (TPR) y la tasa de falsos positivos (FPR) métricas de rendimiento que son especialmente útiles para problemas de clase desequilibrado (Raschka y Mirjalili, 2017). Dichas métricas se calculan de la siguiente forma:

$$TRP = \frac{TP}{FN + TP}, \quad FPR = \frac{FP}{TN + FP}$$

La tasa de verdaderos positivos (TPR) se conoce también como *recall* (sensibilidad), esta métrica tiene la capacidad de detectar muestras positivas verdaderas entre todos los positivos potenciales. Por su parte el FPR se conoce como la especificidad, que de acuerdo con Bonaccorso (2017) es la proporción entre los casos negativos bien clasificados por el modelo, respecto al total de negativos. Esta métrica permite discriminar los casos negativos, es decir valorar que tan bueno el modelo para no obtener falsos positivos.

Otra de las métricas utilizadas es la precisión, la cual refiere a que tan cerca está el resultado de una predicción del valor verdadero; para lo cual se calcula el cociente entre los casos positivos bien clasificados por el modelo y el total de predicciones positivas:

$$Precision = \frac{TP}{FP + TP}$$

Se puede emplear una métrica más para ver el comportamiento de los valores, en este caso se aplica una media armónica ponderada entre la precisión y la tasa de verdaderos positivos (*recall*), la cual suele denominarse como *F1-score* y el valor de beta está entre 1 y 2, y se calcula de la siguiente manera:

$$F_{\beta} = (\beta^2 + 1) \frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

Esta métrica nos proporciona información sobre la precisión de la predicción, dado que la puntuación más alta se logra dando más importancia a la precisión (que es mayor), mientras que el menor corresponde a un predominio de *recall*. Por lo tanto, *F-Beta* (*F1-score*) es útil para tener una imagen compacta de la exactitud como compensación entre alta precisión y un número limitado de falsos negativos.

Método de análisis

Para el desarrollo del análisis, se aplicaron técnicas de procesamiento de lenguaje natural a datos de una encuesta de opinión sobre migraciones con el objetivo de hacer una clasificación de las opiniones en varias categorías. En síntesis, el NLP consiste en tomar un texto en lenguas naturales (español, inglés, etc) y dividirlo en palabras, analizarlas, buscar características comunes (patrones) que luego pueden ser usadas como entrada a diversos algoritmos de *machine learning* (regresión lineal, *Naive Bayes*, árboles de decisión, redes neuronales) y hacer clasificaciones. El proceso realizado se describe a continuación:

Fase 1: Limpieza de archivo. Se recibe un archivo Excel con datos preprocesados de una encuesta sobre migraciones de población. En el archivo vienen pares de columnas una con el texto de la opción y otra con la categoría asociada al tipo de población migrante. Se procesa el archivo y se unen todos los pares de columnas: Queda finalmente un archivo con 8919 registros con dos columnas que contiene todos los datos de los diferentes grupos de población (etnia) y sus diferentes categorías por grupo. Como cada grupo poblacional viene con diferentes categorías realiza un proceso de homologación del cual resultan 11 categorías como se muestra la figura 2.

Figura 2. Homologación de las categorías para su análisis

Grupo	HOMOLOGACIÓN								
	POBREZA (Cat 1)	BUSQ TRABA (Cat 2)	NECES APOYO (Cat 3)	REPRES OTRA CULTURA (Cat 4)	RASGOS NEGATIVOS (Cat 5)	NO CONOZCO (Cat 6)	RASGOS POSITIVO (Cat 7)	LUCHA DERE (Cat 8)	AMBIVALENTE (Cat 9)
ETNIA	1	2	3	4	5	6	7	8	9
HAITI	1	X	X	4	5	6	2,3	X	X
Indígenas panameños	1	X	X	7	4	6	5,2,3	8	X
Colombianos	X	X	X	X	1,3	X	2,5	X	4
Nicaragüenses	X	X	X	X	2,5	X	1,4	X	3
Dominicanos	X	X	X	X	1,3	X	2	X	4
Españoles	X	X	X	X	4	6	1,2,3,5	X	X
Africanos	X	X	X	X	1	3	2,4	X	X
Estadounidenses	X	X	X	X	3,6	X	1,2,4,5	X	X
Chinos	x	x	x	7	5,6		1,2,3,4	X	x

Fuente: Construcción propia.

En la Figura 2 se muestra la tabla de homologación de las categorías, la columna “Grupo” representa cada uno de los grupos de población analizados. En el resto de las columnas representan cada una de las categorías nuevas. El número de cada columna indica cuál o cuáles categorías de un grupo coinciden con la categoría nueva. Las “X” son las categorías que fueron movidas para formar parte de la nueva categorización. Por ejemplo, en Haití, categorías 2 (búsqueda de trabajo) y 3 (necesidad de apoyo para mejor calidad de vida) pasan ser parte de la categoría 7 (rasgos positivos).

Fase 2: Tokenizar. Se toma cada una de las filas del archivo que contienen una muestra del texto o párrafo ingresada por cada usuario. Cada uno de estos textos se divide en palabras

Fase 3: Eliminar caracteres. Se eliminan tildes, comas, caracteres especiales y se dejan solo caracteres de la entre “a” y “z”.

Fase 4: Solo minúsculas. Se pasa todo el texto a minúscula.

Fase 5: Quitar palabras irrelevantes. Dejar solo la raíz de las palabras. De cada verbo se elimina la conjugación presente, pasado y futuro y se deja solo la raíz. Por ejemplo: de la palabra “ayudarles” se obtiene “ayudar”

Fase 6: Crear la bolsa de palabras. Es una matriz donde se toman todas las palabras que quedaron del proceso anterior (Fase 5), y se coloca cada una en una columna. La última columna es el código de categoría. Las filas son cada una de las observaciones de los usuarios, y el número representa cuantas veces aparece la palabra en cada observación. La tabla 1 ilustra una parte de la bolsa de palabras.

Tabla 1. Ejemplo de matriz dispersa de la bolsa de palabras

	Problemas	Oportunidades	Bueno	Malo	Clasificación CÓDIGO
observación usuario 1	1	0	1	0	1
observación usuario 2	0	0	1	0	1
observación usuario 3	1	1	0	1	2
observación usuario n	0	0	1	0	7

Fuente: Construcción propia.

Fase 7: Seleccionar el algoritmo para el modelo de clasificación. En esta etapa, utilizando los datos de la etapa anterior, se probaron varios algoritmos y se determinó cuál de ellos daba la mejor clasificación. Los algoritmos que fueron probados fueron: algoritmo de *K-NN*, algoritmo de *SVM*, algoritmo de Bayes, algoritmo de *Random Forest* y algoritmo de Regresión Lineal

El modelo de clasificación consiste en tres etapas: 1) Entrenamiento y prueba, en el cual se toman una parte de los datos de encuesta para entrenar el modelo y otra parte para probar su funcionamiento. Estos datos corresponden con la bolsa de palabras del paso anterior. Los datos se dividieron de la siguiente manera: 80% datos de entrenamiento y 20% datos de prueba. 2) Crear el modelo, se usan los datos de entrenamiento para encontrar patrones aplicando los algoritmos de clasificación antes mencionados. 3) Probar el modelo, se comparan los datos de prueba con los datos de entrenamiento para medir la precisión de la clasificación realizada por el modelo. Para cada algoritmo se calcula la matriz de confusión y se obtienen las siguientes métricas: precisión, *recall*, *F1-score* y *Accuracy*. Finalmente, con el algoritmo con mejores resultados se aplican las técnicas de *K-fold cross* y *Grid Search* para optimizar los resultados.

Resultados del análisis

Dadas las características del proceso de recolección de datos, en las cuales las personas entrevistadas podían responder libremente sobre qué opinan respecto a las poblaciones migrantes en Costa Rica, las categorías 5 (rasgos negativos) y 7 (rasgos positivos) permiten recoger cómo dichas opiniones tienden a identificar rasgos positivos o negativos sobre dichos grupos; lo cual es un elemento valioso para comprender la construcción que la valoración costarricense realiza sobre las poblaciones migrantes y su disposición a aceptarlas o rechazarlas. Aunado a esto, en el proceso de homologación de categorías (ver figura 2), algunas categorías movieron todos o la mayoría de sus datos a otra, dejando estas sin datos (categorías 2 y 3) o con muy pocos datos (categorías 1,4,8 y 9). Por esta razón, los algoritmos de *Machine Learning* utilizados dan mejores clasificaciones con las categorías que tienen más datos, es el caso de las categorías 5 y 7 (relevantes para el estudio) y las categorías 88 y 99 (NSR y otros, no relevantes para el estudio).

Se puede observar en la figura 4 el análisis de todas las categorías para los datos de prueba (20%). Las categorías 2 y 3 ni siquiera son considerados para la prueba pues no cuentan con datos para ello. Las categorías 1,4 y 8 cuentan con muy pocos de entrenamiento y por ello las bajas predicciones. La categoría 9 cuenta con buenas predicciones, pero al tratarse de la categoría "Ambivalente" son sirve para el análisis.

Una vez obtenidas todas las matrices de confusión se calculó el *Accuracy* de cada algoritmo y la precisión, así como el *recall* y *f1-score* para las categorías más relevantes para la investigación que son la categoría 5 (rasgos negativos) y la categoría 7 (rasgos positivos) de cada algoritmo (ver tabla 2).

Tabla 2. Resultados de las métricas aplicadas

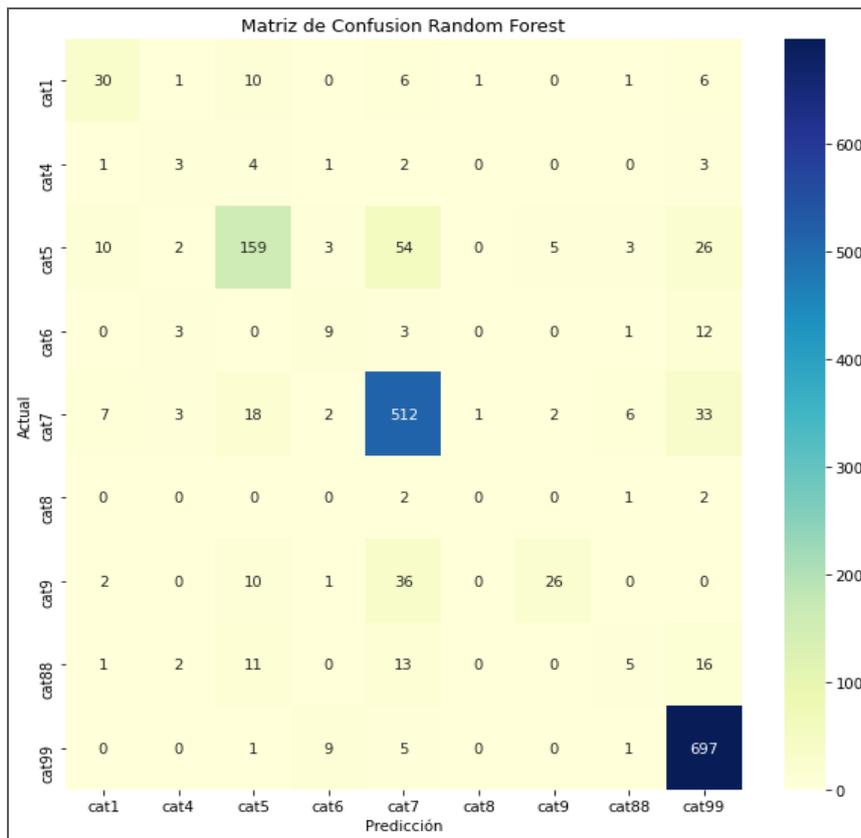
Algoritmo	Accuracy	Precisión (Cat 5)	Precisión (Cat 7)	Recall (Cat 5)	Recall (Cat 7)	F1 (Cat 5)	F1 (Cat 7)
algoritmo de K-NN	0.76	0.66	0.78	0.49	0.83	0.56	0.80
algoritmo de SVM	0.79	0.69	0.84	0.61	0.79	0.75	0.81
algoritmo de Bayes	0.56	0.34	0.75	0.31	0.29	0.33	0.42
algoritmo de Random Forest	0.81	0.74	0.81	1.58	0.88	0.65	0.84
algoritmo de Regresión Lineal	0.80	0.67	0.84	0.61	0.83	0.64	0.84

Fuente: Construcción propia.

Por los resultados obtenidos se determina que el algoritmo que realiza mejor la clasificación es el *Random Forest*. Donde se obtiene una precisión de un 74% y un 81% respectivamente. Y además todo el algoritmo tiene una exactitud de un 81%

Por otra parte, el gráfico 1 muestra el número de predicciones buenas (aciertos) y predicciones malas (desaciertos) por cada categoría. Así, el gráfico muestra que en las categorías 5 y 7 las predicciones buenas superan a las predicciones malas a diferencia de las otras categorías. La categoría 99 es un caso especial pues corresponde con la categoría "no se/no respondo".

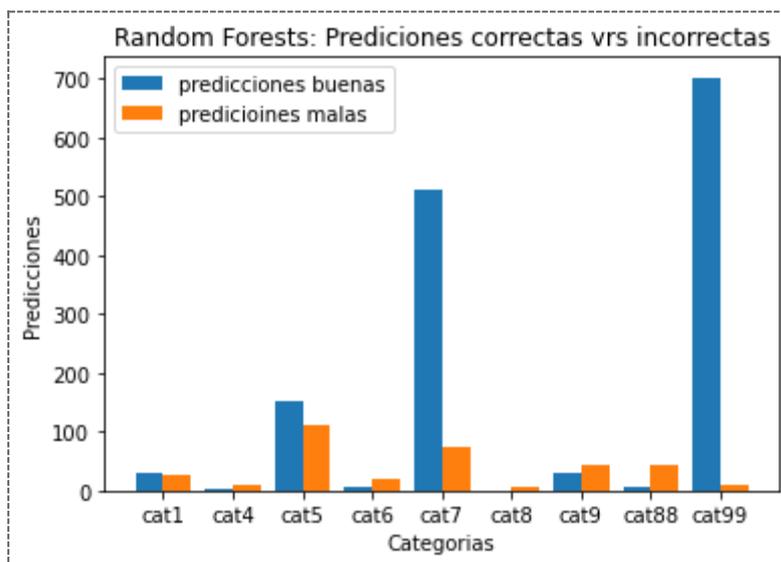
Figura 3. Matriz de confusión del random forest



Fuente: Construcción propia.

La figura 4 muestra la matriz de confusión, en la cual cada fila representa la predicción realizada por el modelo para cada categoría. La diagonal principal muestra los *verdaderos positivos*. Es decir, el dato estaba clasificado en una categoría y el modelo lo clasificó en esa misma categoría. Por ejemplo, de toda la fila de la categoría 5 (rasgos negativos), con 262 muestras, en 151 hubo acierto. El resto de los datos de la fila muestra en qué otra categoría cayó el dato en el que no hubo acierto. Además, la barra lateral corresponde con un mapa de calor, donde el color más intenso representa más aciertos.

Gráfico 1. Cantidad de predicciones buenas vs. malas



Fuente: Construcción propia.

El gráfico 1 muestra las predicciones correctas e incorrectas de cada categoría, se puede resaltar la predicción correcta de las categorías 5 (rasgos negativos) y categoría 7 (rasgos positivos). En el caso de la predicción de categoría 99 (no se responde) el modelo acierta con mayor precisión porque todas las observaciones son los caracteres NS, ns, N/S (no respondo).

Una vez determinado que el *random forest* es el algoritmo con mejores resultados, se aplicó el *Grid search* para comprobar que los parámetros usados por el algoritmo fueron los mejores. Luego se calculó la precisión del algoritmo usando la técnica de *K-Fold Cross* y se obtuvo una media global de 0.7751586188472583 que es un 77,6% de precisión y una desviación estándar 0.013976548078564405, es decir, 1.4%. Con de *K-Fold Cross* se puede determinar que la calidad de la precisión está balanceada y que se mantendrá en el rango de 72.2% a 78.8%. Por lo cual el modelo tiene una probabilidad baja de tener sobreajuste.

Conclusiones

El uso de NLP aplicado a encuestas con preguntas abiertas, demuestra que se puede automatizar el proceso clasificación con una precisión buena respecto al proceso que hace manualmente un grupo de clasificadores. Además, mejora el tiempo de clasificación, con lo cual los investigadores pueden enfocarse más rápido en el análisis de los datos.

Una vez construido el modelo este puede continuar entrenando y aplicando en diferentes encuestas que contengan preguntas similares. Entre más datos se tengan de entrenamiento el algoritmo mejora su precisión. Considerando que se debe ir analizando un posible sobreajuste a medida que se van aumentando los datos.

Se pudo observar que al aplicar el NLP con el algoritmo *random forest* a encuestas con preguntas abiertas permite que el sistema automáticamente clasifique las opiniones en diferentes categorías ayuda a disminuir la ambigüedad en la clasificación. Esto es un aporte importante cuando se está trabajando con datos de percepción sobre temas polémicos, o bien, donde el tipo, cantidad y calidad de la información recolectada es amplia, tal y como sucede con los datos analizados en el presente artículo. Asimismo, el proceso investigativo permitió crear una metodología para preprocesado de los datos. En este caso se pudo unificar datos de encuestas de diferentes grupos de población para generar un base más grande para entrenar el modelo.

Así, al aplicar el algoritmo *random forest* a datos sobre percepción de las poblaciones migrantes en Costa Rica, se pudo corroborar que este pudo clasificar de manera eficiente y adecuada los datos recolectados, tal y como otros trabajos han demostrado (Bonaccorso, 2017; Tintinago et al., 2018), demostrando que con suficiente entrenamiento puedo replicar los resultados de clasificación generados por las personas investigadoras, pero disminuyendo los problemas de ambigüedad y sesgos de clasificación, permitiendo tener una mejor calidad de datos para comprender el fenómeno de estudio.

Por último, lo expuesto en el artículo muestra lo importante que es ampliar el abordaje interdisciplinario en los estudios de opinión. Tradicionalmente los estudios de opinión han sido trabajado por disciplinas como la ciencia política, la sociología, la psicología, el marketing, entre otras; las cuales han permitido construir instrumentos y propuestas para mejorar la calidad y validez de dichos estudios; no obstante, el incorporar disciplinas no tradicionales, como en este caso la ciencia de datos, pueden generar nuevas herramientas y procedimientos para procesar y analizar la información recolectada en los estudios de opinión, permitiendo llegar a obtener un análisis más precisos de las actitudes, valoraciones y opiniones expresadas por las personas a través de las encuestas.

Referencias

- ALMELA SÁNCHEZ-LAFUENTE, Á., VALENCIA GARCÍA, R. & CANTOS GÓMEZ, P. (2012). Detectando la mentira en lenguaje escrito. *Procesamiento Del Lenguaje Natural*, (48), 65-72. Recuperado de <http://rua.ua.es/dspace/handle/10045/22032>
- BEER, D. (2017). The social power of algorithms. *Information, Communication & Society*, 20(1), 1-13. <https://doi.org/10.1080/1369118X.2016.1216147>
- BONACCORSO, G. (2017). *Machine learning Algorithms*. Birmingham, Mumbai: Packt Publishing Ltd.
- CABRERA-ÁLVAREZ, P. (2022). Survey Research in Times of Big Data: Investigación con encuestas en los tiempos del big data. *EMPIRIA: Revista de Metodología de Ciencias Sociales*, 53, 31-51. <https://doi.org/10.5944/empiria.53.2022.32611>
- CETINA PRESUEL, R. & MARTÍNEZ SIERRA, J. M. (2019). Algorithms and the News: Social Media Platforms as News Publishers and Distributors. *Revista de Comunicación*, 18(2), 261-285. <https://doi.org/10.26441/rc18.2-2019-a13>
- DELGADO MONTALDO, D. (2008). Percepciones de la inmigración e integración en Costa Rica. *Papeles de población*, 14(57), 65-91.
- DÍAZ, R. M. & DOMÍNGUEZ, Á. Z. (2013). Las ciencias sociales y los dispositivos de la complejidad. *Cuadernos de Administración*, 29(50), 123-131.
- FERNÁNDEZ, A. M. (2004). El procesamiento del texto como lenguaje natural. *Hermeneus: Revista de la Facultad de Traducción e Interpretación De Soria*, (6), 75-98.
- FERNÁNDEZ, G. M. (2012). Adquisición y representación del conocimiento mediante procesamiento del lenguaje natural. Universidad de la Coruña [Tesis Doctoral]. Recuperado de https://ruc.udc.es/dspace/bitstream/handle/2183/10057/FernandezGavilanes_Milagros_TD_2012.pdf?sequence=5&isAllowed=y
- FONDEVILA GASCÓN, J. F. (2017). Algoritmos sobre el impacto de los medios de comunicación en medios sociales: Estado de la cuestión. *Icono* 14, 15(1), 21-41. <https://doi.org/10.7195/ri14.v15i1.948>
- GONZÁLEZ, F. (2019). Big data, algoritmos y política: Las ciencias sociales en la era de las redes digitales. *Cinta de moebio*, 65, 267-280. <https://doi.org/10.4067/s0717-554x2019000200267>
- GUALDA, E. (2022). Social big data y sociología y ciencias sociales computacionales. *EMPIRIA: Revista de Metodología de Ciencias Sociales*, 53, 147-177. <https://doi.org/10.5944/empiria.53.2022.32631>
- HERNÁNDEZ, M., Y GÓMEZ, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica*, 32 (1), 87-96. <https://doi.org/10.33333/rp.vol32n0.32>
- IDESPO-UNA (2012). Encuesta: Construcción de opiniones públicas sobre la migración en Costa Rica.
- MAMAQI, X., BANDRES GOLDARAZ, E., & PEREZ CALLE, R. D. (2021). Analisis Big Data de la opinión pública sobre inmigración en redes sociales. En Pérez Calle, R.D., Trincado Aznar, E. y Gallego Abaroa, E. (Coord.) *Economía, empresa y justicia. Nuevos retos para el futuro* (pp. 1451-1467). Madrid : Dikson
- MAXWELL, J. & TOMLINSON, J. (2020). Proving algorithmic discrimination in government decision-making. *Oxford University Commonwealth Law Journal*, 20(2), 352-360. <https://doi.org/10.1080/14729342.2020.1833604>
- MOREIRA, D., CRUZ, I., GONZÁLEZ, K., QUIRUMBAY, A., MAGALLAN, C., GUARDA, T., ANDRADE, A. & CASTILLO, C. (2021). Análisis del Estado Actual de Procesamiento de Lenguaje Natural. *Revista Ibérica De Sistemas e Tecnologías De Informação*, 126-136.
- PORTER, N. D., VERDERY, A. M. & GADDIS, S. M. (2020). Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. *PLoS ONE*, 15(6), 1-21. <https://doi.org/10.1371/journal.pone.0233154>

RASCHKA, S. y MIRJALILI, V. (2017). *Python Machine learning. Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Mumbai: Packt Publishing.

SANDOVAL GARCÍA, C. (2004). El “otro” nicaraguense en el imaginario colectivo costarricense. Algunos retos analíticos y políticos. *Nómadas (Col)*, (20),152-159.

TINTINAGO, A., MIÑOZ, Y., URIBE, G. A. & ÁLVAREZ SÁNCHEZ, P. H. (2018). Etiquetado asistido de documentos de investigación mediante procesamiento de lenguaje natural y tecnologías de la web semántica. *Scientia et Technica*, 23(4), 528–537.

TU, S.-T., LU, L., HSIEH, C.-H. & WU, C.-Y. (2021). A New Internet Public Opinion Evaluation Model: A Case Study of Public Opinions on COVID-19 in Taiwan. *International Journal of Big Data and Analytics in Healthcare*, 6, 1-17. <https://doi.org/10.4018/IJBDAH.287603>

ZORAN, U. (2016). Las Ciencias Computacionales como recurso para la toma de decisiones: Los algoritmos. *Revista Antioqueña de las Ciencias Computacionales*, 6(2), 55-59.

Autores.**Esteban Martínez Porras**

Escuela de Informática, Universidad Nacional (Costa Rica) y Escuela de Matemáticas, Universidad de Costa Rica.

Máster en Planificación curricular y Licenciado en Enseñanza de la Matemática ambos grados por la Universidad de Costa Rica. Docente en la Sede Interuniversitaria de la Universidad Nacional. Investigador del proyecto Evaluación de las tecnologías digitales que refuerzan el desarrollo del pensamiento lógico matemático en estudiantes de ingeniería industrial: propuesta de indicadores de calidad de la Universidad de Costa Rica.

E-mail: esteban.martinez.porras@una.ac.cr

Adrián Ramírez Fernández

Escuela de informática, Universidad Nacional, Costa Rica.

Máster en Telemática por el Instituto Tecnológico de Costa Rica y Bachiller en Ciencias de la computación, Universidad de Costa Rica. Docente en la Sede Interuniversitaria de la Universidad Nacional.

E-mail: adrian.ramirez.fernandez@una.ac.cr

Laura Solís Bastos

Instituto de Estudios Sociales en Población (IDESPO), Universidad Nacional, Costa Rica.

Doctora en Demografía por la Universidad Nacional de Córdoba, Argentina, Máster en Estudios Latinoamericanos con énfasis en cultura y desarrollo, Licenciada en Sociología, ambos grados por la Universidad Nacional, Costa Rica. Investigadora del programa Umbral Político, adscrito al Idespo, Universidad Nacional, Costa Rica

E-mail: laura.solis.bastos@una.ac.cr

José André Díaz-González

Instituto de Estudios Sociales en Población (IDESPO), Universidad Nacional, Costa Rica.

Doctor en Gobierno y Políticas Públicas, Magíster en Historia y Licenciado en Ciencias Políticas por la Universidad de Costa Rica. Investigador del Programa Umbral Político del Instituto de Estudios Sociales en Población de la Universidad Nacional, y docente de la Escuela de Ciencias Políticas de la Universidad de Costa Rica.

E-mail: jose.diaz.gonzalez@una.ac.cr

Citado.

MARTÍNEZ PORRAS, Esteban; RAMÍREZ FERNÁNDEZ, Adrián; SOLÍS BASTOS, Laura y DÍAZ-GONZÁLEZ, José André (2025). Uso de Procesamiento de Lenguaje Natural para procesar respuestas abiertas de una encuesta de Opinión Pública. *Revista Latinoamericana de Metodología de la Investigación Social - ReLMIS*. N°29, Año 15, pp. 51-67.

Plazos.

Recibido: 28/10/2022. Aceptado: 15/06/2023.



¿Por qué soy de mi clase social? Un análisis cuantitativo y cualitativo de respuestas sobre autopercepción de clase en un estudio por encuesta del Área Metropolitana de Buenos Aires (2016)

Why am I of my social class? A
quantitative and qualitative analysis of responses on self-perception of class
in a survey study in the Buenos Aires Metropolitan Area (2016)

Rodolfo Elbert

Resumen

El artículo presenta una reflexión metodológica acerca de las siguientes categorías utilizadas para indagar la autopercepción de clase social en un estudio por encuesta (PI-Clases, 2016): “Alta”, “Media Alta”, “Media”, “Obrera” y “Baja”. En primer lugar, presentamos un análisis estadístico de nuestros resultados en comparación con otras encuestas recientes, mostrando cierta estabilidad en las preferencias subjetivas. En la segunda parte analizamos las respuestas a una pregunta abierta incluida en nuestra encuesta: “¿Por qué cree usted que pertenece a esa clase social?” Esto dio lugar a un corpus de 10.146 palabras, analizadas mediante una estrategia cuantitativa (diversas librerías de R) y cualitativa (codificación manual con el software Atlas.ti). Los resultados indican que los/as encuestados/as expresan justificaciones narrativas bien diferenciadas para la selección de diferentes categorías de clase. Aquellos/as que eligieron “Media” y “Media alta” se refirieron a la calidad de vida, educación e ingresos; los/as que seleccionaron clase “Obrera” hablaron, centralmente, de dignidad y esfuerzo; y, por último, los/as autopercebidos/as de clase “Baja” justificaron su elección por la falta de ingresos y diversas carencias en las condiciones materiales de vida.

Palabras clave: Clases sociales; encuestas; autopercepción de clase; cuestionario; métodos combinados.

Abstract

The article presents a methodological reflection on the following categories used to investigate the self-perception of social class in a survey study (PI-Clases, 2016): "Upper", "Upper Middle", "Middle", "Working", and "Lower". In the first section, we present a statistical analysis of our results compared to other recent surveys, showing some stability in subjective preferences. In the second part, we analyze the responses to an open question in our survey: "Why do you think you belong to that social class?" This resulted in a corpus of 10.146 words, which were processed using a mixed quantitative (with different R programming language packages) and qualitative (using Atlas.ti) methodological strategy. The results indicate that respondents expressed distinct narrative justifications for their selection of different class categories. Those who chose "Middle" and "Upper-Middle" referred to quality of life, education, and income; those who selected "Working" class primarily spoke of dignity and hard work; and finally, those who perceived themselves as "Low" class justified their choice by a lack of income and various deficiencies in material living conditions.

Keywords: Social class; survey; subjective class; questionnaire; mixed methods.

1. Introducción

El diseño de las preguntas del cuestionario en un estudio por encuesta es un proceso complejo que implica tomar una serie de decisiones teóricas y metodológicas de gran impacto en la calidad de los datos. Estas decisiones tienen que ver con diferentes aspectos como el tipo de pregunta, la operacionalización del concepto a ser medido y las palabras utilizadas en la descripción del tema a indagar, entre otras cuestiones (Schaeffer y Dykema, 2020). Para el caso de preguntas que requieren que los/as encuestados/as se auto-clasifiquen con respecto a su estatus o comportamientos pasados, el sistema de categorías provisto en el cuestionario se convierte en una cuestión de suma importancia para comprender los resultados. El artículo presenta una reflexión metodológica acerca de las categorías elegidas en el diseño de una pregunta sobre autopercepción de clase social incluida en una encuesta relevada por nuestro equipo de investigación (PI-Clases, 2016): *“Los diarios, la TV, los políticos, algunas personas, hablan de clases sociales. ¿Se considera usted a sí mismo como perteneciendo a una clase social?”* Las categorías de respuesta que ofrecimos fueron: “Alta”, “Media Alta”, “Media”, “Obrera” y “Baja”.

Abordamos dos interrogantes centrales: ¿Qué efecto tiene la variación en las categorías brindadas a los/as encuestados/as para medir la autopercepción de clase en un estudio por encuesta? ¿Qué significado tiene para las propias personas encuestadas la categoría de clase social seleccionada? Para responder estas preguntas, desplegamos una estrategia que combina el análisis estadístico de las pautas de autopercepción de clase en diversas encuestas recientes, con un análisis cuantitativo y cualitativo de las palabras que los/as encuestados/as utilizan para justificar sus elecciones. Los datos analizados provienen de la encuesta PI-Clases relevada por nuestro equipo de investigación entre los años 2015 y 2016. En la primera sección de análisis del trabajo ubicamos nuestros resultados cuantitativos en el contexto de otras encuestas recientes que utilizaron categorías diferentes en el diseño de la pregunta sobre autopercepción de clase (CEDOP-UBA, 2009-2010; ENES-Pisac, 2014; ESAyPP/PISAC-COVID-19, 2021). En la segunda parte del análisis nos enfocamos en las palabras y las narrativas que los propios encuestados utilizan cuando tienen que responder porqué eligieron su categoría de clase subjetiva.

Ubicamos este análisis en el marco de los debates metodológicos existentes sobre el desafío de medir con una o dos preguntas de un cuestionario un concepto complejo como la autopercepción de clase social (Hout, 2008). Esta complejidad radica en que el instrumento de recolección debe captar de manera fehaciente la imagen mental que construyen las personas sobre la estructura social y su ubicación en ella. Está claro que hay otras maneras de medir la autopercepción de clase —que puede contener, incluso, factores inconscientes o perspectivas macrosociales de formación de clase—, pero aquí nos restringimos a los diseños orientados a captar el proceso por el cual las personas se ubican, conscientemente, en una posición de clase al ser consultadas en el marco de un estudio por encuesta (Wright, 1997: 382). Al haber incluido en nuestra encuesta una pregunta cerrada con opciones de clase y, a la vez, una pregunta abierta que solicitaba la justificación de cada selección, incorporamos evidencia novedosa para la reflexión metodológica sobre la medición de este fenómeno.

2. Desafíos metodológicos en la medición de la autopercepción de clase

Según Wright (2015 [2005]), el estudio de las pautas de identidad de clase en una sociedad es una de las seis preguntas centrales que se debe plantear el análisis de clases sociales: ¿cómo explicamos la manera en la cual las personas se ubican subjetivamente, de manera individual y colectiva (a sí mismos y a otros), en una estructura de desigualdad?

El concepto central que guía las investigaciones empíricas alrededor de esta pregunta es el de identidad de clase, que podemos definir, operativamente, como la identificación subjetiva con una clase social. Esta identificación se construye en base a las experiencias biográficas en la estructura socioeconómica y la ubicación subjetiva en esa estructura a partir de la propia interpretación de esas experiencias.

Si bien desde esta perspectiva la expresión de la identidad de clase es de nivel individual y microsocioal, no deja de ser un proceso que se realiza de manera intersubjetiva y relacional, en

colaboración con otros/as significativos/as de nuestro entorno social inmediato (Sautu, 2004; Vryan Adler y Adler, 2003). A nivel un poco más general, se puede comprender la identidad de clase como una de las dimensiones de un concepto microsocioal de conciencia de clase, que incluye tanto este aspecto subjetivo de identificación con una clase social como una dimensión vinculada a la percepción subjetiva de los intereses de clases (Wright, 1997; Perez, 2014).

A su vez, la identidad de clase es un concepto complejo que incluye al menos dos dimensiones empíricas relevantes: la autopercepción de clase (definición de la propia clase social y de aquellos que comparten esta situación de clase) y la identificación de aquellos grupos que se diferencian u oponen a esta definición de clase (Thompson, 1966; Wright, 1997). Los estudios por encuesta, como el que analizamos en el presente artículo, se enfocan, por lo general, en la dimensión de *autopercepción* o *autoafiliación* de clase, definida como “uno de los posibles indicadores de la identificación del sujeto con una clase dada, lo que implica a la vez la percepción que él tiene del sistema de estratificación y de su propia posición en el mismo” (Germani, 2010 [1963]: 171). Tal como afirma Hout (2008), el desafío de los estudios por encuesta para la medición de la autopercepción de clase es el diseño de una o varias preguntas que permitan: i. que las personas encuestadas perciban la desigualdad social y la asocien con la existencia de clases sociales; ii. presentar una imagen de la desigualdad de clases que capture la percepción que tienen las personas (cuántas clases hay y cómo se llaman estas clases); iii., que los/as encuestados/as se ubiquen subjetivamente a sí mismos/as en el marco de esta estructura de desigualdad a partir de la selección de una de las categorías otorgadas. Un aspecto clave en el proceso de medición es el sistema de categorías que se propone al encuestado para que elija la que mejor representa su posición de clase.

Una modalidad muy utilizada en estudios de opinión pública es la de solicitar a los/as encuestados/as que se ubiquen a sí mismos/as en una escala numérica, de 1 a 10, que representa un espacio continuo desde las personas más pobres a las más ricas. En un estudio reciente con datos de Latinobarómetro, Marqués Perales y Rodríguez de la Fuente (2023) muestran el potencial de este tipo de mediciones para los estudios comparativos internacionales. Otra modalidad, que es la que utilizamos en nuestro estudio, es brindar opciones categóricas de clase social, para que los propios encuestados se ubiquen en ellas según su sentido de pertenencia. Esta modalidad se propone reflejar, en una tipología cualitativa, las categorías de clase teóricamente relevantes para la investigación, pero plantea el desafío operacional acerca de qué sistema de categorías puede representar mejor la imagen mental que los/as propios/as encuestados/as tienen del sistema de estratificación en clases sociales.

Dentro de las investigaciones que proporcionan opciones categóricas, una escala muy utilizada es la que brinda a los encuestados las siguientes opciones ordinales: “Alta”, “Media alta”, “Media”, “Media baja” y “Baja”. Se podría suponer que esta escala “gradacional pura” es la más utilizada debido a su coherencia metodológica interna. Sin embargo, es práctica común en las investigaciones sociológicas intentar “romper” la gradación para que el/la encuestado/a no confunda las etiquetas con categorías de ingresos y se permita, a su vez, una mayor diferenciación entre las clases o estratos propuestos. En palabras de Germani, cierto grado de discontinuidad es necesario para que se destaquen los diferentes estratos en la propia representación de los encuestados (2010 [1963]). En otras palabras, la “gradación pura” se rompe al incluir una o dos categorías que son puramente “de clase”, mediante lo cual se espera que el/la encuestado/a genere un “mapa mental” de la estructura de clases y no una gradación de ingresos. Si bien el estudio de la autopercepción de ingresos y su contraste con los ingresos reales de las personas es un tema importante, históricamente estuvo más asociado a las investigaciones en economía que en sociología (Cruces, Pérez Truglia y Tetaz, 2011).

Finalmente, es importante señalar que, dentro de los estudios que ofrecen opciones categóricas “discontinuas”, existen diferentes maneras de “romper” la gradación pura. Una alternativa muy utilizada en las investigaciones anglosajonas es reemplazar las categorías “Media baja” y “Baja” por la categoría “*Working class*”. En algunos países de habla hispana esta opción fue traducida como “clase trabajadora”, brindando las opciones “Alta”, “Media alta”, “Media”, “Trabajadora” y “Baja”. Como bien señala Jorrot (2008), en la Argentina la categoría “*working class*” puede también ser traducida como “clase obrera”. Cada una de estas decisiones operacionales tiene consecuencias en el plano de la medición, por lo cual el objetivo del presente

artículo es reflexionar acerca de los resultados obtenidos a partir de las opciones elegidas en nuestra investigación, ubicadas en el contexto de otros estudios realizados recientemente.

3. Datos y métodos

Esta investigación está basada en una estrategia de análisis multimétodo que combina el análisis estadístico de pautas de autopercepción de clase con un análisis cuantitativo y cualitativo de las palabras que estos/as encuestados/as utilizan para justificar sus elecciones. Los datos analizados provienen de la encuesta PI-Clases (2016). Se trata de una encuesta a una muestra probabilística de 1065 casos, de diseño estratificado y multietápico, con selección aleatoria en todas sus etapas. El universo de estudio es la población de 25 a 65 años al momento de la encuesta. La selección de este grupo etario se debe a que fue una encuesta sobre trayectorias ocupacionales, para lo cual se diseñó una muestra de la población en edad económicamente activa (Sautu et al., 2020).

En la sección del cuestionario dedicada al estudio de la percepción subjetiva de clase social se incluyeron las preguntas cerradas y abiertas que analizamos en el presente artículo. En primer lugar, a modo de apertura del tema, el cuestionario incluyó la siguiente pregunta: “*Los diarios, la TV, los políticos, algunas personas, hablan de clases sociales. ¿Se considera usted a sí mismo como perteneciendo a una clase social?*”. Tal como afirma Hout (2008), esta pregunta permite distinguir entre aquellos/as encuestados/as que tienen una identificación de tipo “no forzada” con una clase social y aquellos/as que rechazan en primera instancia este tipo de pertenencia. En nuestra encuesta, el 84% de los/as encuestados/as respondió afirmativamente a esta pregunta. El 16% de los/as encuestados/as que negó tener una posición de clase o que no quiso contestar la pregunta, tuvo que responder luego la siguiente pregunta: “*Muchas personas dicen que pertenecen a una clase determinada. Si tuviera que elegir, diría que pertenece a...*” y se les mostró una tarjeta con las opciones de clase social. Un total de 159 personas que respondieron negativamente la primera pregunta, pasaron a elegir una etiqueta de clase ante la repregunta, lo cual puede ser definido como una autopercepción de clase “forzada”.

Cuadro 1. Autopercepción de clase no forzada/forzada en PI-Clases 2016.

	N	%
Autopercepción no forzada	897	84,2
Autopercepción forzada	159	14,9
No tiene autopercepción de clase	9	0,8
Total	1065	100,0

Fuente: elaboración propia en base a encuesta PI-Clases (2016)

Luego de este “filtro”, los/as encuestados/as que aceptaron posicionarse fueron confrontados con un listado de cinco posiciones de clase posibles para elegir la propia: “Alta”, “Media alta”, “Media”, “Obrera” y “Baja”. Sumando los/as encuestados/as “no forzados/as” y los/as “forzados/as”, más del 99% de los/as encuestados/as eligió una posición de clase (solo 9 encuestados/as rechazaron la pertenencia de clase en ambas instancias).

La primera sección de análisis del artículo se focaliza en la distribución cuantitativa de los/as encuestados/as en la pregunta de autopercepción de clase. Esta sección analiza el porcentaje de encuestados/as que seleccionaron cada etiqueta de clase en nuestra encuesta y lo ubica en una perspectiva comparada con encuestas relevadas en el Área Metropolitana de Buenos Aires en los años 2009-2010 (CEDOP-UBA, 2009-2010), 2014 (ENES-Pisac, 2014) y 2021 (ESAYPP/PISAC-COVID-19, 2021). Estas encuestas fueron realizadas a una muestra probabilística nacional,¹ pero, para este artículo, restringimos el análisis a los casos ubicados en el Área

¹ El detalle de los diseños metodológicos de estas encuestas se puede encontrar en Jorrat (2012), Maceira (2015) y Dalle y Di Virgilio (2022), respectivamente.

Metropolitana de Buenos Aires. Lo interesante de estas encuestas es que, en algunos casos, los cuestionarios incluyeron etiquetas de clase diferentes al de nuestra encuesta, lo cual nos permitirá comprender las variaciones en las distribuciones de clase subjetiva según diferentes categorías propuestas en el cuestionario. Es importante aclarar que, si bien cada encuesta corresponde a un momento particular, el ejercicio comparativo presentado tiene un carácter metodológico y no se propone relacionar los resultados con cambios en el contexto y/o en la estructura de clases de cada momento histórico.²

Luego de ubicar los resultados cuantitativos de nuestro estudio en el contexto de otras encuestas, pasaremos a estudiar en profundidad el contenido subjetivo de las preferencias de clase de nuestros/as encuestados/as. Para ello, el cuestionario de la encuesta PI-Clases (2016) incluía la siguiente pregunta, formulada luego de que el/la encuestado/a eligiera su etiqueta de clase: *¿Por qué cree usted que pertenece a esa clase social?* Se trata de una pregunta totalmente abierta que permitió a los/as encuestado/as justificar su elección, y dar respuestas que tuvieron una extensión de, desde una palabra (las más breves) hasta de un párrafo de cuatro oraciones (las más extensas). Las respuestas a esta pregunta abierta constituyen un corpus de texto constituido por 10.146 palabras, que fueron procesadas mediante una estrategia metodológica mixta cuantitativa (con diferentes librerías de R) y cualitativa (mediante Atlas.ti).

4. Análisis comparativo de la relevancia de diferentes etiquetas de clase subjetiva en cuatro estudios por encuesta

El primer dato relevante para cualquier análisis de una pregunta incluida en un cuestionario, es una distribución de frecuencias simple que nos indique la relevancia de cada categoría en el total de la distribución. En el Cuadro 2 podemos observar la distribución total de los/as encuestados/as según la etiqueta de autopercepción de clase seleccionada:

Cuadro 2. Autopercepción de clase de encuestados/as. Encuesta PI-Clases 2016

Alta	1	0,1
Media alta	34	3,2
Media	500	47,3
Obrera	439	41,6
Baja	82	7,8
Total	1.056	100

Fuente: elaboración propia en base a encuesta PI-Clases (2016)

Como podemos observar, la autoubicación de los/as encuestados/as en la escala propuesta se concentra en dos categorías principales: la clase media (47,3%) y la clase obrera (41,6%). Un primer dato interesante es que, utilizando las opciones provistas, la distribución de la muestra se parte exactamente en dos mitades si tenemos en cuenta aquellos/as que eligieron las categorías de la clase media “hacia arriba” (“Alta”, “Media Alta” y “Media”) y aquellos/as que eligieron las categorías “hacia abajo” de la clase media (“Obrera” y “Baja”). En este punto, es interesante reflexionar sobre las opciones brindadas en nuestro cuestionario. A primera vista, parecería contraintuitivo incluir la opción “Obrera” en un listado de etiquetas gradacionales que van desde la clase “Alta” hasta la clase “Baja”. Se podría pensar que lo más lógico sería completar la escala sin romper la gradación, con un listado de opciones que podría ser el siguiente: “Alta”, “Media Alta”, “Media”, “Media Baja” y “Baja”. Sin embargo, como ya hemos señalado previamente,

² La encuesta del año 2021, sobre todo, presenta la particularidad de haber sido relevada luego de una prolongada cuarentena en el marco de la pandemia de COVID-19. Un análisis situado de sus resultados se puede encontrar en Dalle (2022). Para un análisis detallado de los resultados de la encuesta ENES-Pisac (2014), con respecto a la variable de clase subjetiva, ver Maceira (2018) y Assusa y Mansilla (2019).

en las investigaciones sociológicas se tiende a “romper” la escala gradacional para que el/la encuestado/a no confunda las etiquetas con categorías de ingresos.

Para poder evaluar las opciones brindadas por nuestro cuestionario presentamos a continuación la interpretación de nuestros resultados en perspectiva comparada con resultados de las diferentes encuestas recientes mencionadas con anterioridad (CEDOP-UBA 2009-2010; ENES-Pisac 2014, ESAyPP/PISAC-COVID-19). Como podemos observar en el Cuadro 3, un dato interesante es que las variaciones en las categorías propuestas se dan “hacia abajo” de la clase media, nunca “hacia arriba”. Como podemos ver, todas las encuestas comparten las categorías “Alta”, “Media Alta” y “Media”; y también todas empiezan a variar una vez que descendemos en la escala social. Es decir, cada equipo de investigación respondió de manera diferente a la pregunta sobre qué categorías proponer a las personas cuya identificación se da “hacia abajo” de la clase media.

Cuadro 3. Autopercepción de clase en diversas encuestas (2010-2021)

Autopercepción de clase	CEDOP-UBA (2009-2010)	ENES-Pisac (2014)	PI-Clases (2016)	ESAYPP/PISAC-COVID (2021)
Alta	0,0	0,3	0,1	0,1
Media Alta	2,3	3,8	3,2	1,7
Media	47,8	40,7	46,9	19,1
Media Baja	18,6	24,3	XXXXXXXX	15,5
Obrera	25,8	23,4	41,2	XXXXXXXX
Trabajadora	XXXXXXXX	XXXXXXXX	XXXXXXXX	50,7
Baja	5,5	7,6	7,7	12,8
Total	768	1057	1065	1029
% de personas que eligieron una clase	98%	98,1%	99,4%	99,3%

Fuente: elaboración propia en base a encuestas citadas.

En primer lugar, podemos señalar que en todas las encuestas hay un porcentaje similar de autopercepción de clase (ya sea “forzada” o “no forzada”). Con respecto a la distribución de los/as encuestados/as, encontramos, principalmente, que la categoría clase “Alta” es prácticamente inexistente en el Área Metropolitana de Buenos Aires. Aclarado este punto, podemos afirmar que en tres de las cuatro encuestas (2009-2010, 2014 y 2015) se da una pauta similar de autopercepción de clase “Media” y “Media alta”, categorías que incluyen alrededor del 50% de los casos en cada uno de estos estudios. Al interior de este grupo, claramente la categoría más importante es la de clase media, que abarca el 47,8% en 2009-2010, el 40,7% en 2014 y el 46,9% en 2015. La encuesta de 2021 (ESAYPP/PISAC-COVID-19) tiene un porcentaje bastante menor, con el 20,9% de los casos ubicados en la clase “Media” o “Media alta”.

Antes de continuar con el análisis, es importante señalar que la sumatoria mencionada en el párrafo anterior presenta el desafío de dónde ubicar la categoría “Media baja” cuando construimos categorías dicotómicas de “media o +” por un lado y “debajo de la media” por el otro. ¿Es una categoría que se debe sumar a la “Media” o a la “Obrera” y a la “Baja” para representar lo que está “abajo” de la “Media”? Nuestra decisión en este análisis comparativo es ubicar a la “Media-baja” “por debajo” de la “Media”. De esta manera, obtenemos una medición más “pura” de la relevancia de las autopercepciones de clase “Media” en nuestra sociedad.

La única excepción se hizo para el estudio ESAyPP/PISAC-COVID-19, ya que, en este caso, el porcentaje de personas que seleccionaron la clase “Media” fue del 19,1%, un porcentaje significativamente más bajo que en el resto de las investigaciones. La razón de esta diferencia, como explicaremos más abajo, es la inclusión de la opción de clase “Trabajadora” en lugar de clase “Obrera”, llevando a un crecimiento importante de las preferencias “por debajo” de la clase “Media”.

Ahora bien, para comprender las pautas de distribución de las categorías “por debajo de la clase media” realizaremos una doble comparación, tomando nuestra encuesta de 2016 como eje

del análisis. En primer lugar, podemos comparar nuestros resultados con aquellos de las encuestas de 2009-2010 (CEDOP-UBA) y 2014 (ENES-Pisac). Mientras nuestra encuesta reemplazó la categoría “Media baja” por la categoría “Obrera”, en las otras encuestas se incluyen tres categorías “por debajo de la Media”: “Media baja”, “Obrera” y “Baja”. Los resultados muestran que nuestra decisión no afectó la distribución de los casos. En las encuestas de 2009-2010 y 2014, la sumatoria de casos entre clase “Media baja” y clase “Obrera” es del 44,4% y 47,7% respectivamente, mientras que nuestra encuesta contiene un 41,6% de los casos en la categoría clase “Obrera”. Pareciera que aquellos casos que eligieron “Media baja” en los estudios previos, de ser confrontados con nuestro sistema de categorías, hubieran elegido clase “Obrera”. Finalmente, en los tres estudios se observa una similar relevancia en la categoría clase “Baja”, seleccionada por entre el 5% y el 7% de los/as encuestados/as.

Con respecto a la comparación de nuestra encuesta PI-Clases (2016) con la encuesta ESAyPP/PISAC-COVID-19 (2021), es interesante el contraste entre la categoría clase “Obrera” y la categoría clase “Trabajadora” utilizada en este último estudio. Los resultados muestran que la inclusión de la categoría clase “Trabajadora” afecta los resultados, siendo la categoría más elegida en la encuesta de 2021, con un 50,7% de los casos. Tal como afirma Jorrat (2008), la categoría “clase trabajadora” tiene una connotación positiva en la Argentina, lo cual tiene la consecuencia de que, al ser incluida entre las opciones, se convierte en la categoría más elegida. Esto es lo que ocurre, claramente, en la encuesta de 2021, generando, a su vez, una drástica reducción en el porcentaje de encuestados que seleccionan la clase “Media” (19,1%). Sumando las opciones de clase “Trabajadora” y clase “Baja”, en la encuesta de 2021 el 63,6% de los encuestados se autopercibe “debajo de la clase Media”, cifra que asciende al 79% si sumamos a los que seleccionaron “Media baja”. Estos números reflejan un fuerte contraste con nuestra encuesta, en la cual un 48,9% de los casos se ubican “debajo de la clase Media” si sumamos a los que eligieron “Obrera” y “Baja”. Es importante aclarar que no existen mediciones “mejores”; sino que cada medición debe ser evaluada con respecto a lo que se propone captar.

5. El porqué de una clase: análisis de las justificaciones de los/as encuestados/as para su autopercepción de clase

Tomando como punto de partida la estabilidad de las pautas en la autopercepción de clase a través de las diferentes encuestas, en este apartado analizaremos las palabras que las propias personas utilizan para justificar la selección de las diferentes categorías. Para ello, nuestra encuesta incluyó una pregunta abierta que fue respondida por 983 personas que, en primer lugar, habían elegido una etiqueta de clase. El objetivo de este análisis es poder determinar si existen diferencias relevantes en las justificaciones subjetivas de las diferentes categorías de clase. El supuesto que guía el análisis es que la decisión de elegir una categoría por sobre otra está basada en la construcción de significado del porqué esa categoría es apropiada para describir la propia situación de clase: ¿qué palabras están asociadas a la identidad de clase “Media”? ¿Por qué una persona considera pertenecer a la clase “Obrera” y no a la clase “Baja”? ¿Qué lugar tienen las justificaciones económicas para explicar la pertenencia a la clase “Media alta”? El Cuadro 4 presenta un resumen con la información analizada para cada una de las categorías de clase:

Etiqueta de clase	Cantidad de casos	Cantidad de palabras (con stopwords)	Cantidad de palabras (sin stopwords)
Media alta	32	372	125
Media	484	5615	1877
Obrera	393	3417	992
Baja	74	742	218
Total	983	10146	3212

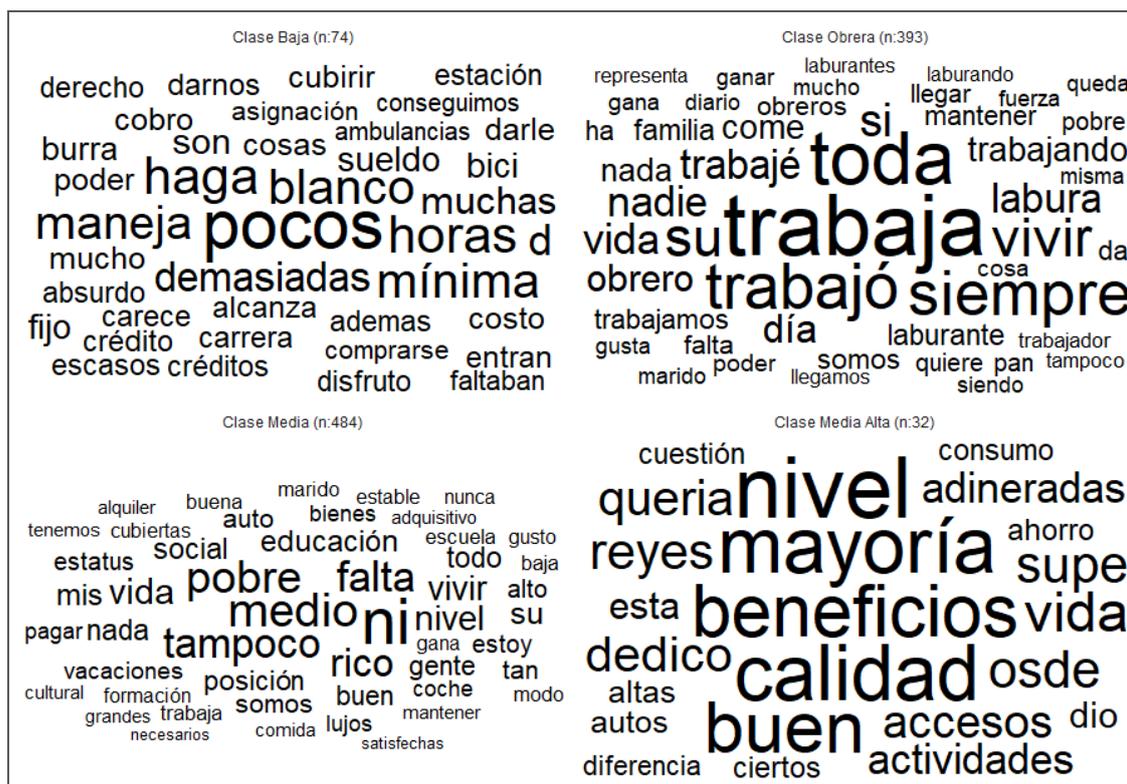
Fuente: elaboración propia en base a encuesta PI-Clases (2016)

El punto de partida del análisis son las 10.146 palabras incluidas en las transcripciones textuales de las respuestas a nuestra pregunta abierta.

Como podemos ver, las categorías subjetivas con mayor cantidad de palabras son las de clase “Media” (5615) y clase “Obrera” (3417), lo cual tiene sentido ya que fueron seleccionadas por un mayor número de personas en nuestra encuesta. Es importante recordar que, al ser confrontadas con la pregunta por el porqué de su clase social, las personas brindaron respuestas que van de un mínimo de una palabra a un máximo de tres o cuatro oraciones (entre 20 y 25 palabras aproximadamente). El promedio de palabras utilizadas por nuestros/as encuestados/as es de diez palabras por respuesta. Tal como podemos ver en el Cuadro 4, el panorama cambia un poco cuando excluimos del corpus textual a las palabras “vacías” o “stop words” (pronombres, artículos y preposiciones, entre otras). Nuestro análisis se realizó luego de esta exclusión, por lo cual el corpus textual, que es la base de esta sección, incluye 3212 palabras totales, distribuidas según clase subjetiva de acuerdo a lo presentado en el Cuadro 4.

A partir de esta fuente de datos, realizamos, en primer lugar, un análisis cuantitativo descriptivo de las palabras más utilizadas para justificar cada categoría de clase social. Presentamos a continuación una nube de palabras para cada categoría, elaboradas a partir de la técnica “frecuencia inversa de documentos (*Tf_idf*, por sus siglas en inglés), que consiste en dividir la cantidad de apariciones de una palabra en un documento por un logaritmo de la cantidad de documentos en los cuales aparece esa palabra³. Desde este puntaje se construyeron los rankings de las palabras que más aparecen en cada una de las clases subjetivas, ponderando a favor de las palabras que tienen mayor capacidad de distinguir a esa clase de las otras clases sociales:

Figura 1. Nubes de palabras para justificación de clase subjetiva en base a medida “*Tf_idf*”



Fuente: Elaboración propia en base a PI-Clases (2016).

³ Agradezco a Joaquín Gueler, estudiante avanzado de la Carrera de Sociología, UBA, que me brindó su asistencia para el análisis cuantitativo de palabras con R que se presenta en este gráfico.

Tal como vemos, existen diferencias importantes en las palabras numéricamente más relevantes para cada una de las clases subjetivas. Las personas que eligieron la clase “Media” utilizaron, principalmente, palabras como “medio”, “tampoco”, “ni”, “pobre”, “rico”. Mediante el uso frecuente de estas palabras, los/as encuestados/as expresan una autoubicación equidistante en la estructura social, tanto en relación a “los ricos” como a “los pobres” (palabras que también tienen relevancia en esta nube). En segundo lugar, las personas con identificación obrera les dan mayor relevancia a palabras como “trabaja”, “trabajó”, “siempre”, “vivir”, que están asociadas a la noción de “una vida de trabajo”. Para estos/as encuestados/as, el autopoicionamiento de clase se construye, centralmente, como una identidad basada en la necesidad de trabajar para vivir. Las identidades de clase “Baja”, por otra parte, se expresan en palabras que se refieren a una carencia o una falta, como “pocos”, “demasiadas” y “mínimas”. Como veremos a continuación, esas expresiones de falta aparecen mayoritariamente asociadas a los ingresos, al trabajo y a las condiciones materiales de vida en general. Finalmente, la nube de palabras de la clase “Media alta” resalta palabras como “beneficios”, “calidad”, “mayoría”, “nivel”, que expresan la distinción de pertenecer a un nivel de vida privilegiado.

En este punto es importante profundizar el análisis de las respuestas, pasando de las palabras individuales más mencionadas al tipo de argumentos que prevalecen en las justificaciones de cada categoría de clase. Si bien las nubes de palabras nos permiten identificar con precisión el primer anillo del sentido de cada clase (observando a simple vista cuáles son las palabras con más presencia), es importante combinarlas con un análisis temático o de tópicos que permita profundizar en las justificaciones narrativas para cada categoría.⁴ Este análisis permite, por un lado, ubicar a las palabras en sus contextos de sentido y, por otro lado, comprender el sentido de las palabras más alejadas del núcleo principal, que pierden relevancia en el análisis puramente cuantitativo. En el siguiente cuadro podemos observar los temas más relevantes que identificamos mediante una codificación manual (no automatizada) realizada a través del software Atlas.ti:

Cuadro 5. Narrativas de clase identificadas en cada clase subjetiva (análisis temático)

Media Alta	Media	Obrera	Baja
Nivel de vida	Ni-Ni (Ni pobre ni rico) Nivel de vida (Ingresos)	Trabajo Somos laburantes (no clase media)	Ingresos (bajos) Trabajo (carencias)
Consumos	Cultura y Educación Trabajo y supervivencia	Sobrevivo Dignidad y esfuerzo	No tengo/No soy Humilde
Trabajo			
Educación			

Fuente: Elaboración propia en base a PI-Clases (2016).

Con el objetivo de representar de la manera más precisa posible las narrativas asociadas a cada una de las clases subjetivas, presentamos a continuación un análisis que combina los hallazgos cuantitativos de las nubes de palabras con el análisis cualitativo realizado mediante la codificación temática. Es importante aclarar que se realizó una articulación flexible de ambos criterios con el objetivo de representar, de la manera más precisa posible, las diferentes dimensiones y subdimensiones presentes en cada narrativa de clase. El orden de presentación de las narraciones respetará la relevancia relativa de cada una de las clases en la muestra analizada, comenzando por la clase “Media” (elegida por 484 personas), y continuado con las clases “Obrera” (393 personas), “Baja” (74 personas) y, por último, “Media alta” (32 personas).

⁴ En esta etapa del análisis se podría haber optado por una detección automática de tópicos a partir de alguna técnica basada en el procesamiento de lenguaje natural (Rosati et al., 2021). Sin embargo, al tratarse de un número limitado de textos, consideramos que la codificación manual podía responder adecuadamente a las necesidades de la investigación. A futuro se podrían comparar los temas emergentes de la codificación manual con una detección automática de tópicos, o, incluso, el desarrollo de una codificación que articule etapas manuales y automatizadas (Nelson, 2017).

Como ya vimos, la nube de palabras representativa de la clase media tiene en el centro a tres palabras que resumen muy bien la pauta central de estas interpretaciones: “medio”, “tampoco” y “ni”. Estas palabras forman parte de una perspectiva relacional de la posición de clase “Media” que apareció con frecuencia en las respuestas analizadas. Para estas personas, ser de clase “Media” es ubicarse en un lugar intermedio de la estructura social, que se diferencia “hacia abajo” de los pobres y “hacia arriba” de “los ricos” o la clase “Alta”: *“Ni pobre ni rico, una persona que está en el medio”, “No somos ni muy pobres, ni tampoco de tener mucha plata”, “No soy ni muy pobre ni muy bien, estoy en el medio”, “No soy ni pobre ni millonario”*. Un buen resumen de este tipo de interpretaciones lo brindó una encuestada que afirmó ser de clase media *“Porque no vivo mal. No me falta nada, pero tampoco me sobra. Vivo bien”*.

El segundo anillo de palabras surgidas de las justificaciones de la identidad de clase media se refiere al nivel de vida de los/as encuestados/as y su capacidad de acceso a determinados bienes y servicios. En este nivel, aparecen con relevancia palabras como “posición”, “nivel”, “vida” o “estatus”, conceptos que expresan cierta evaluación positiva del propio lugar en la estructura social. Para los/as encuestados/as que utilizan estas palabras, la pertenencia a la clase media se explica *“Por el nivel de vida que llevo”, “Por el nivel de vida que tengo”, “Por mi estatus, por mi forma de vida”*. El rasgo positivo de estas respuestas se puede resumir en la idea de “vivir bien”: *“Me alcanza lo que gano y vivo bien”, “Vivo tranquilo”, “Vivo bien”*. En particular, cuando tienen que justificar esta idea, los/as encuestados/as se refirieron a una dimensión económica, con la mención de bienes y servicios a los que pueden acceder (“vacaciones”, “auto”, “poder adquisitivo”) o a un aspecto educativo/cultural, con referencias a “educación”, “formación” o “escuela”, entre otras palabras. Finalmente, hay un grupo de palabras que tuvieron menor importancia relativa, pero que refieren a una interpretación central de la identidad de clase “Media”: son palabras como “trabaja”, “alquiler”, “necesidades satisfechas”; que expresan una mirada referida a una “clase media trabajadora”, que tiene las necesidades básicas cubiertas, pero tiene que trabajar para mantener ese nivel de vida.

La segunda categoría de clase más elegida por los/as encuestados/as fue la de clase “Obrera”, seleccionada por 393 encuestados/as (con justificación) que dieron lugar a un corpus textual de 992 palabras (sin “stop words”). Las palabras más relevantes de esta nube son “trabaja”, “trabajó”, “vivir”, “vida” y “siempre”, que expresan la presencia continua del trabajo a lo largo de la propia vida, con encuestados/as que afirmaron que se sentían obreros/as porque *“durante toda la vida trabajé”, “toda la vida laburé”, “porque lo mío fue siempre laburando”,* o, simplemente, porque *“trabajo, trabajo y trabajo”*. Ampliando estas justificaciones, aparecen en la nube palabras como “familia”, “comer”, “día”, “pan” y “mantener”, que vinculan el trabajo con la necesidad de la subsistencia y de mantener una familia. Estos/as encuestados/as ponen el énfasis en que el trabajo les permite lograr un piso mínimo de subsistencia y escaparle al hambre: *“considero que lo que tengo lo tengo trabajando, no tengo el lujo de no trabajar”, “hoy pertenezco a la clase obrera porque día a día tengo que ganar para mantener a mi familia”, “trabajo para mantener a mi familia”*. En resumen, para los/as encuestados/as la importancia del trabajo reside en poder *“cubrir mis necesidades básicas a mucho esfuerzo”, “me importa llegar a la quincena y que no les falte nada a mi hija o mi nieto”*.

Ahora bien, más allá del aspecto material vinculado a la “vida de trabajo”, en un segundo lugar de relevancia, la nube muestra palabras como “obrero”, “obrerros”, “laburante” “laburantes” y “trabajador”, que aparecen en las respuestas expresando una afirmación de la propia identidad. Estas personas afirmaron que pertenecían a la clase “Obrera” *“Porque soy un trabajador”* o *“Porque soy una trabajadora”, “Porque me siento trabajador”, “Por laburante”,* o *“Porque soy simple y laburante”*. Continuando con esta afirmación identitaria, hubo respuestas que vincularon la propia identidad a una historia familiar (*“Porque mis dos viejos fueron laburantes”, “Porque vengo de familia trabajadora*) o incluso a una noción más amplia de clase social (*“Pertenezco a la clase trabajadora, formo parte del grupo que sale a buscar trabajo”, “La clase trabajadora, a la cual pertenezco, es la que levanta al país”, “Porque soy de la clase trabajadora, la que trabaja”, “Porque soy pobre, me identifiqué con la clase obrera”*).

Un aspecto final que es importante resaltar del análisis de estas respuestas es que, más allá de la mención a numerosas carencias materiales, los/as encuestados/as que eligieron la opción clase “Obrera” tuvieron, por lo general, una valoración positiva de la propia identidad. El

aspecto valorativo no es evidente en la nube de palabras, pero surge del análisis temático de las respuestas asociado a esta nube. A través del mismo, encontramos que la noción de trabajo se valora positivamente por dos cuestiones principales. La primera, es que los/as encuestados/as relacionan la identidad obrera con las nociones de orgullo y dignidad: *“No bajo la cabeza, no me doy por vencido ni aún vencido”, “Primero por mis ideas, porque no puedo aparentar algo que no soy”*. En segundo lugar, la idea de esfuerzo y sacrificio con el orgullo de no depender de nadie para sostener la propia posición de clase: *“Porque toda mi vida, lo poco que logré, fue siempre con sacrificio, no dependiendo de nadie”, “Lo que tengo lo conseguí laburando, nadie me regaló nada”*.

En tercer lugar, hubo 74 encuestados/as que eligieron la categoría clase “Baja” cuando tuvieron que definir su autoubicación de clase y brindaron una explicación para esa selección. El análisis de la nube de palabras en conjunto con el análisis temático de las respuestas nos permite afirmar que la elección de la categoría de autoidentificación clase “Baja” tuvo, por lo general, una connotación negativa para los/as encuestados/as. Estas miradas expresan cierta insatisfacción o pesimismo por la propia posición de clase. Las palabras que más aparecen en esta nube son aquellas que se refieren a cantidades: “pocos”, “mínima”, “demasiadas”, “mucho” y “muchas”. Estas palabras aparecen en las comparaciones que realizan los encuestados para ubicarse por debajo de la escala social. En primer lugar, señalan algunas cuestiones materiales que le faltan a ellos/as para estar en la clase “Media” o “Alta”: *“La clase baja son los que menos tienen”, “Clase alta es cuando uno tiene algo y yo todavía ni una bici tengo”, “La clase alta y media son las que tienen dinero, la clase baja son los que menos tienen”, “Porque la clase media vive de otra manera”*. En particular, el tema de la vivienda aparece en estas respuestas, con foco en la falta de una vivienda y en la imposibilidad del acceso al crédito: *“No puedo acceder a una vivienda, ni a un crédito [...]”, “Por la situación laboral y no poder acceder a préstamos para poder comprar una casa”, “Porque no alcanza el dinero para comprarse una casa. Ya eso, para mí, es bajo”, “Porque no tengo terreno y no consigo trabajo”, “Porque todavía no podemos tener casa propia”*. En resumen, desde estas perspectivas, la falta de una vivienda está en el centro de la explicación para la pertenencia a la clase “Baja”.

Por otra parte, aparecen respuestas enfocadas en el trabajo. La palabra “trabajo” tuvo un bajo índice “if_df” por aparecer también en otras categorías de clase, pero en términos de frecuencias brutas tuvo mucha presencia en las respuestas de clase “Baja”. A diferencia de las respuestas de clase “Obrera”, en las respuestas de clase “Baja” las personas no se refieren de manera positiva al propio trabajo, sino que señalan su falta o su carácter precario. Aquellos que no tenían trabajo al momento de la encuesta afirmaron ser de clase “Baja”, *“Porque estoy sin laburo. Con dificultades económicas”, “Porque no tengo trabajo”, “Porque no tengo trabajo, y no me dan la posibilidad de trabajar en lo que se”, “Porque carezco de muchas cosas y no tengo un trabajo que me saque del pozo”*. Otras respuestas remarcaron la falta de un trabajo registrado o “en blanco”: *“Porque no tengo trabajo en blanco”, “[...] no tengo trabajo en blanco”*. Finalmente, aparecen respuestas que enfatizan que, a pesar de tener trabajo, el sueldo no les alcanza: *“Porque no alcanza el sueldo”, “Porque ya hoy no alcanza el sueldo para vivir, no llegás a fin de mes [...]”, “A veces no llego a fin de mes”, “Por el poder adquisitivo”*; y también respuestas enfocadas en la precariedad que implica el exceso de trabajo: *“Porque me la paso trabajando”, “Porque tengo que trabajar como burra para sobrevivir” “Si no trabajo no como, y vivo en la calle [trabajando]”, “Trabajo demasiadas horas. Me va bien, pero son demasiadas horas. Tengo un nene que casi no lo veo, no lo disfruto”*. Ya sea por falta de trabajo, por exceso de horas o por ausencia de derechos, en estas respuestas el trabajo aparece en el centro de la explicación para la pertenencia a la clase “Baja”.

El último grupo que analizaremos en este apartado son aquellos/as encuestados/as que se identificaron subjetivamente con la categoría clase “Media alta”. Esta categoría fue seleccionada por 34 personas, que representan el 3,3% de los/as encuestados/as. De ellos/as, hubo 2 que no justificaron esa pertenencia, por lo cual el análisis subsiguiente se basa en las respuestas de 32 encuestados/as que eligieron esta categoría y justificaron esa decisión. Si bien se trata de una categoría elegida por un número relativamente bajo de encuestados/as, podemos encontrar en sus respuestas un patrón interpretativo particular para la autoubicación de clase que los/as distingue de las otras posiciones. En el centro de la nube de palabras construida en base a estas respuestas aparecen palabras como “calidad”, “vida”, “beneficios”, “buen” y “nivel”, que expresan una noción de “distinción” en base a la “calidad de vida” que está en el corazón de estas

interpretaciones. Para estas personas, la pertenencia a la clase media alta se explica por *“educación, entorno, nivel de ingreso, calidad de vida”, “Por el estándar de vida”, “No vivo mal, vivo bien, tengo lo que quiero”, “Porque tengo más privilegios que la mayoría [...]”*. Son interpretaciones basadas en la idea de la “distinción”, que para ellos/as significa la pertenencia a un grupo social privilegiado (*“Por el conjunto de amigos que tengo, la educación, las actividades que hago, las oportunidades y beneficios que tengo”*) y en un lugar obtenido a partir del propio esfuerzo: *“Tengo los privilegios que me supe conseguir”, “Por el esfuerzo y el tipo de trabajo. La vida que hice, para progresar y tener un buen nivel de vida”*,

Cuando reflexionan sobre las características de su lugar privilegiado sobresalen palabras como “ahorro”, “auto”, “beneficios” y “consumo”, que ponen énfasis en las diferencias económicas a su favor: *“Por el nivel de ingresos”, “Porque [...] vivo en un barrio con servicios, tengo auto”; “Porque tengo una casa grande, dos autos”, “por el acceso al sistema de salud privado que tiene (OSDE) que esa es una gran diferencia respecto a los que no estudian, eso es lo que le di a mis hijos”, “No vivo mal, vivo bien, tengo lo que quiero”*. Otro factor en la construcción de esta idea de distinción es el nivel educativo, representado en la nube a palabras como “acceso” y “nivel”, y que aparece por lo general asociado a los ingresos económicos de la persona: *“Por el nivel de educación e ingresos”, “Por el nivel de ingresos, por el nivel de educación”, “Por los estudios que logré”*. En resumen, las personas de clase “Media alta” se consideran “privilegiadas” en la estructura social principalmente por sus ingresos, lo cual determina cierta capacidad de consumo y ahorro que no tienen otras clases sociales. Cuando tienen que explicar estos ingresos, aparece el concepto de acceso privilegiado a una educación que les permitió ubicarse en lo más alto de la escala social.

6. Conclusión

Las investigaciones sobre autopercepción de clase social enfrentan el problema de la falta de homogeneidad en las categorías que brindan a los/as encuestados/as para que se auto-ubiquen en la estructura de clases. Las opciones van desde las ordinales continuas hasta las ordinales discontinuas o incluso escalas cuantitativas que ubican a los/as encuestados/as en un rango numérico de 1 a 10. En este artículo intentamos utilizar a nuestro favor esta diversidad de enfoques con el objetivo de comparar los resultados obtenidos por los diferentes sistemas de categorías utilizados.

Una primera conclusión es que, más allá del tipo de pregunta y categorías utilizadas, la gran mayoría de las personas tiene una respuesta disponible cuando es consultada por su posición de clase. Este hallazgo desmiente opiniones bastante difundidas acerca de “la muerte de las clases” o la irrelevancia de las clases desde la perspectiva subjetiva de las personas. Tal como afirma Hout (2008), si las clases sociales fueran un “artefacto” teórico-metodológico de las ciencias sociales sin anclaje en dinámicas sociales reales, sería muy difícil observar que, de manera estable, 8 de cada 10 personas se autoidentifican de manera “no forzada” con una clase social y 9 de cada 10 lo hacen luego de la repregunta.

En segundo lugar, podemos afirmar que aproximadamente la mitad de los encuestados/as se ubica en la clase “Media” o categorías superiores, desmintiendo la noción muy difundida en los medios de comunicación acerca de que “la gran mayoría de los argentinos se autopercebe en la clase media”. De ninguna manera esto es así, y la comparación de las diversas encuestas presentadas en este artículo confirma que, al menos la mitad de los/as encuestados/as decide ubicarse “por debajo” de la clase “Media”, en alguna de las siguientes categorías: “Media baja”, “Trabajadora”, “Obrera” o “Baja”. Los resultados también muestran algunas variaciones en los resultados dependiendo de las categorías que incluye el cuestionario. El hallazgo más relevante al respecto es que la inclusión de la categoría “Trabajadora” genera una atracción muy superior al resto de las categorías, llevando a una reducción de la importancia relativa de la clase “Media”.

El análisis de las palabras utilizadas por los/as encuestados/as nos mostró que las categorías subjetivas de clase están sostenidas por narrativas bien diferenciadas. Las personas autopercebidas de clase “Media” centraron sus respuestas en la idea de “ni-ni”, que sostiene la pertenencia de clase por no ser ni rico ni pobre. A su vez, enfatizan su nivel educativo y de ingresos, lo que consideran que les permite ubicarse en esa clase social. La siguiente categoría de clase en

nuestro estudio fue la “Obrera”, que también tiene connotaciones positivas para los/as encuestados/as. En este caso, las respuestas estuvieron centradas en el orgullo y la dignidad de pertenecer a esa clase, y en el hecho de “trabajar toda la vida” como una marca de clase. Por último, podemos mencionar que las categorías menos relevantes también tuvieron justificaciones que las distinguen. Las personas que eligieron clase “Baja”, tuvieron, por lo general, una mirada pesimista y negativa de su clase social, centrada en aquello que les falta para ascender en la escala social (ingresos, vivienda, educación, etc.). Las personas que eligieron clase “Media alta” hicieron énfasis en su posición privilegiada y en la distinción entre su clase y las de “abajo”, centrada en sus ingresos, relaciones, bienes y nivel educativo.

Tal como podemos ver, las personas eligen activamente una clase social y, a la vez, tienen argumentos para justificar esa selección. Nuestros datos muestran cierta coherencia en las pautas de identidad de clase entre diferentes encuestas, y también detectan diferencias significativas en las maneras en las cuales las personas justifican sus elecciones. En términos metodológicos, esta investigación brinda argumentos a favor de la combinación de diferentes estrategias de análisis (cuanti y cuali) para analizar el fenómeno de la autopercepción de clase. Si bien los estudios cuantitativos clásicos sobre el tema siguen siendo relevantes, la incorporación de nuevas preguntas y tipos de datos para explorar esta temática puede aportar resultados novedosos. Por un lado, la comparación de encuestas que brindan opciones de clases diferentes es interesante, ya que permite captar diferencias sutiles en la comprensión legada de las categorías de clase. Esta línea de investigación se podría profundizar a partir del estudio de cómo los diferentes sistemas de categorías afectan la correspondencia entre clase objetiva y clase subjetiva, por ejemplo. Por otro lado, el análisis combinado cualitativo y cuantitativo permite ampliar nuestra comprensión del sentido de las clases para las personas, y se podría profundizar con la incorporación de nuevas técnicas computacionales como el análisis de frecuencia inversa de documentos (utilizado en este artículo), la detección automática de tópicos y el análisis de sentimientos asociados a cada categoría de clase, entre otros.

Bibliografía

- ASSUSA, G. y MANSILLA, H. O. (2019). La clase social como posición y representación. Un análisis sociológico de la autoafiliación en la estructura social Argentina (2014-2015). *Laboratorio*, 29.
- CRUCES, G.; PÉREZ TRUGLIA, R. y TETAZ, M. (2013). Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment. *Journal of Public Economics*, 98, 100-112.
- DALLE, P. (Comp.) (2022). *Estructura social de Argentina en tiempos de pandemia. Efectos de la doble crisis y recomposición social en disputa*. Buenos Aires: IIGG-Agencia I+D+i.
- DALLE, P. y DI VIRGILIO, M. M. (2022). Estructura social de Argentina y políticas públicas durante la pandemia de Covid-19: El diseño de una encuesta nacional comparativa interregional. *Revista Latinoamericana de Metodología de las Ciencias Sociales*, 12(2).
- ELBERT, R. (2018). Informality, Class Structure, and Class Identity in Contemporary Argentina. *Latin American Perspectives*, 45(1), 47-62.
- GERMANI, G. (2010[1963]). Clase social subjetiva e indicadores objetivos de estratificación. En C. Mera y J. Rebon (Comps.), *Gino Germani: La sociedad en cuestión*. Buenos Aires: CLACSO.
- HOUT, M. (2008). How Class Works: Objective and Subjective Aspects of Class Since the 1970s. En A. Lareau y C. Dalto (Eds.), *Social Class: How Does It Work?* Nueva York: Russell Sage Foundation.
- JORRAT, J. R. (2008). Percepciones de clase en la Argentina. *Revista Estudios del Trabajo*, 36, 49-83.
- JORRAT, J. R. (2012). Clase, identidad de clase y percepción de las sociedades desde elitistas hasta igualitarias: Argentina en un contexto comparativo internacional. *Desarrollo Económico. Revista de Ciencias Sociales*, 52(205), 63-93.
- KELLY, J. (1988). *Trade unions and socialist politics*. Londres: Verso.
- MACEIRA, V. (2015). Un abordaje teórico-metodológico para la investigación de la estructura, la movilidad social y las condiciones de vida: La propuesta ENES-PISAC. *Revista Latinoamericana de Metodología de las Ciencias Sociales*, 5(2).
- _____ (2018). Clases y diferenciación social. En J.I. Piovani y A. Salvia (Coords.), *La Argentina en el siglo XXI: Cómo somos, vivimos y convivimos en una sociedad desigual: Encuesta Nacional sobre la Estructura Social*. Buenos Aires: Siglo XXI.
- MARQUÉS PERALES, I. y RODRIGUEZ DE LA FUENTE, J. (2023). Posicionamiento subjetivo y condición socioeconómica en América Latina (2006-2020). Una aproximación desde el análisis multinivel. *DADOS - Revista de Ciências Sociais*, 67(4), 1-38.
- NELSON, L. K. (2020). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, 49 (1): 3-42.
- PEREZ, P. (2014). Cómo entender y estudiar la conciencia de clase en la sociedad capitalista contemporánea. Una propuesta. *Revista Theomai*, 29.
- ROSATI, G., CHAZARRETA, A., DOMENECH, L. y MAGUIRE, T. (2021). Una aproximación a los temas acerca de la COVID-19: Aplicación de técnicas de procesamiento de lenguaje natural sobre comentarios de lectores de noticias digitales. *Papeles de Trabajo*, 15 (28), 64-91.
- SAUTU, R. (2004). Estilos y prácticas de la investigación biográfica. En R. Sautu (Comp.), *El Método biográfico: La reconstrucción de la sociedad a partir del testimonio de los actores*. Buenos Aires: Ediciones Lumiere.
- SAUTU, R., BONIOLO, P., DALLE, P. y ELBERT, R. (Eds.). (2020). *El análisis de clases sociales: Pensando la movilidad social, la residencia, los lazos sociales, la identidad y la agencia*. Buenos Aires: IIGG-CLACSO.
- SCHAEFFER, N. C. y DYKEMA, J. (2020). Advances in the Science of Asking Questions. *Annual Review of Sociology*, 46(1), 37-60.

THOMPSON, E. P. (1966) *The Making of the English Working Class*. New York: Vintage Books.

VRVAN, K.y ADLER, P. (2003). Identity. En L. Reynolds y N. Herman-Kinney (Comps.), *Handbook of Symbolic Interactionism*. Oxford: Altamira Press.

WRIGHT, E. O. (1997). *Class counts: Comparative studies in class analysis*. Cambridge: Cambridge Univ. Press.

_____ (2015 [2005]). *Modelos de análisis de clases*. Valencia: Tirant Humanidades.

Autor.

Rodolfo Elbert

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Instituto de Investigaciones Gino Germani, Universidad de Buenos Aires (UBA), Argentina.

Doctor en Sociología por la Universidad de Wisconsin (Madison, Estados Unidos). Magister en Investigación en Ciencias Sociales y Licenciado en Sociología por la Universidad de Buenos Aires (UBA). Investigador Adjunto del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) en el Instituto Gino Germani (UBA). Profesor Adjunto en la Carrera de Sociología (UBA).

E-mail: elbert.rodolfo@gmail.com

Citado.

ELBERT, Rodolfo (2025). ¿Por qué soy de mi clase social? Un análisis cuantitativo y cualitativo de respuestas sobre autopercepción de clase en un estudio por encuesta del Área Metropolitana de Buenos Aires (2016). *Revista Latinoamericana de Metodología de la Investigación Social - ReLMIS*, N°29, Año 15, pp. 68-82.

Plazos.

Recibido: 22/08/2023. Aceptado: 23/02/2024.



Experiencia de uso de Chat GPT-3 como asistente de búsqueda en revisiones sistemáticas de literatura

Experience in using GPT-3 Chat as a search assistant
in systematic literature reviews

Claudia Cabrera Borges

Resumen

Este artículo tiene como objetivo revisar el uso de chatbots GPT como apoyo para investigadores que realizan revisiones de literatura bajo el protocolo SALSA en el ámbito de la investigación educativa, desde una perspectiva crítica sobre el uso de la inteligencia artificial generativa. Se explora la utilidad del chatbot GPT-3 en la primera fase del protocolo (búsqueda/*search*), articulando la interacción con la herramienta a partir de "autonomía de los alumnos" como palabra o término clave en el diseño de este proceso. La metodología implicó revisión de fuentes académicas que reportan procesos similares, y la implementación del uso del chatbot en la fase de diseño de ecuaciones de búsqueda para el caso. Los resultados indican que el chatbot GPT-3 mejoró la eficiencia al proporcionar resultados precisos, coincidiendo con estudios previos. En conclusión, esta herramienta puede ser valiosa en la realización de estas investigaciones, pero se destaca la importancia de abordar su uso con un enfoque crítico y considerar aspectos éticos en la investigación educativa. Se sugiere explorar más las potencialidades de estas herramientas generativas, así como sus limitaciones, en futuros estudios.

Palabras clave: Revisión sistemática de literatura; Protocolo SALSA; Inteligencia Artificial Generativa; Ecuaciones de búsqueda; Colaboración.

Abstract

This article aims to review the use of GPT chatbots as support for researchers who conduct literature reviews under the SALSA protocol in the field of educational research, from a critical perspective on the use of generative artificial intelligence. The usefulness of the GPT-3 chatbot in the first phase of the protocol (search) is explored, articulating the interaction with the tool based on "student autonomy" as a keyword or term in the design of this process. The methodology involved reviewing academic sources that report similar processes and implementing the use of the chatbot in the design phase of search equations for the case. The results indicate that the GPT-3 chatbot improved efficiency by providing accurate results, coinciding with previous studies. In conclusion, this tool can be valuable in conducting this research, but the importance of addressing its use with a critical approach and considering ethical aspects in educational research is highlighted. It is suggested to explore more the potentialities of these generative tools, as well as their limitations, in future studies.

Keywords: Systematic literature review; SALSA protocol; Generative Artificial Intelligence; Search equations; Collaboration.

Introducción¹

Las revisiones de literatura son un tipo de investigación que recopila, evalúa e interpreta de forma sistemática, rigurosa y transparente investigaciones previas sobre un tema o pregunta específicos (Grant & Booth, 2009). Se basa en una metodología clara y replicable que incluye las siguientes fases: definir el objetivo y el alcance de la revisión, identificar las fuentes y los criterios de búsqueda, ejecutar las consultas en las bases de datos determinadas, seleccionar los estudios relevantes mediante un proceso de cribado, extraer y analizar los datos de los estudios incluidos, y presentar y discutir los resultados (Onwuegbuzie & Frels, 2016).

El objetivo de una revisión sistemática de literatura es sintetizar el conocimiento disponible, identificar brechas y necesidades de investigación futura, y orientar la toma de decisiones basada en evidencia (Booth *et al.*, 2016; Hart, 2018). Se trata de una herramienta valiosa para la investigación científica, especialmente en el ámbito de la salud, la educación y las ciencias sociales. Los investigadores deben desarrollar una serie de competencias digitales y de investigación (Böttcher & Thiel, 2017; Questa-Tortero *et al.*, 2022) para implementarlas apropiadamente, en tanto se requiere la aplicación de una metodología específica y rigurosa, que es recomendable llevar a cabo en equipo para asegurar su fiabilidad y evitar errores (Codina, 2023a, Leite *et al.*, 2023).

Una forma de realizar este tipo de revisiones es la aplicación del marco metodológico SALSA (Grant & Booth, 2009). Esta metodología consiste en diseñar ecuaciones de búsqueda para consultar bases de datos y recuperar artículos relevantes sobre un tema o pregunta específicos (Codina, 2017). En el marco de esta metodología se plantean cuatro fases para el proceso de revisión: i) búsqueda, ii) evaluación, iii) síntesis y iv) análisis (en inglés Search, Appraisal, Synthesis, Analysis términos de los que deriva el nombre). A partir de este marco, se genera un protocolo que debe documentar cada fase para asegurar que la revisión es reproducible (García-Peñalvo, 2022).

En líneas generales, y siguiendo a García-Peñalvo (2022) la fase de búsqueda del protocolo indica, a partir de criterios definidos, cómo se realiza la exploración de fuentes primarias a revisar. En esta fase se definen las palabras clave, el diseño de la ecuación de búsqueda, los filtros (por ejemplo, temporalidad, idiomas, tipos de fuentes, entre otros) que determinan los criterios de inclusión y exclusión, y en qué bases de datos se realiza la búsqueda. Como resultado, se obtiene una serie de registros de fuentes primarias. Con estos insumos, se desarrollan las fases siguientes del protocolo SALSA. Se evalúan (Appraisal) las fuentes obtenidas en la búsqueda y se define si formarán parte del corpus final de la revisión, registrando los criterios pragmáticos o de calidad de las fuentes. Finalmente, durante las fases de síntesis y análisis (Synthesis y Analysis) se busca representar sintéticamente cada fuente primaria según sus características y en relación con las preguntas que originan la revisión de literatura, llegando a una descripción y valoración de los hallazgos sobre el estado del conocimiento en el ámbito de la revisión.

Por lo tanto, los artículos de revisión sistemática de literatura son valiosos en la medida en que sintetizan información para los lectores, y proporcionan claridad y conocimientos sobre un tema (Hart, 2018; Onwuegbuzie & Frels, 2016; Zawacki-Richter *et al.* 2020). En la misma línea, las revisiones sistemáticas son la base para proponer nuevos proyectos de investigación o trabajos académicos (Codina, 2020). A pesar de la importancia de las revisiones sistemáticas, un número significativo de publicaciones actuales aún refieren a enfoques tradicionales que pueden ser susceptibles de sesgo y, como resultado, pueden proporcionar conclusiones incorrectas. Por lo tanto, es necesario que los autores, editores y revisores sean estrictos y que los metodólogos puedan apoyar a toda la comunidad científica sobre este tema (Haddaway, 2020).

Por otro lado, el proceso de realizar una revisión sistemática puede ser largo y complejo, y requiere un alto nivel de experiencia tanto en el tema como en las bases de datos específicas que se utilizan. La complejidad se debe, en primera instancia, al volumen del material bibliográfico. La búsqueda, organización y selección de información requieren de mucho tiempo debido a la falta de sistematización apropiada de los documentos, la gestión deficiente de los metadatos, y las

¹ En el trabajo reportado en este artículo participaron como investigadoras la Dra. Andrea Tejera Techera, Dra. Mariela Questa-Tortero y Dra. Claudia Cabrera Borges.

diferencias de resultados que aportan los buscadores (Carbajal-Degante *et al.*, 2023). En segunda instancia, puede resultar complejo el diseño de la ecuación de búsqueda a utilizar, especialmente si se tienen en cuenta los criterios de validez para su formulación (Codina, 2023b). Estudios previos mencionan que un alto porcentaje de las revisiones sistemáticas de literatura no informan de manera adecuada los métodos de búsqueda y no realizan búsquedas exhaustivas de fuentes debido a debilidades en la formulación de las ecuaciones con operadores booleanos, lo que sucede tanto en el campo de la medicina como en el de la educación (de Kock *et al.*, 2021; Polanin *et al.*, 2017). Por lo tanto, definir y reportar las ecuaciones de búsqueda resulta un paso fundamental de la primera fase del protocolo SALSA.

Una ecuación de búsqueda combina palabras clave (keywords) con operadores booleanos (AND, OR, NOT) y es utilizada en una base de datos para recuperar fuentes de interés para la revisión (Codina, 2023b; Newman & Gough, 2020). Para la selección de las palabras clave está recomendado el uso de tesauros para partir de términos normalizados y vinculados de manera jerárquica o equivalente, en las bases de datos disponibles (Cajal *et al.*, 2020).

En la actualidad, las herramientas basadas en inteligencia artificial están en constante evolución y no solo cambian sus utilidades, sino que también cambian sus nombres comerciales (por ejemplo, Bard de Google actualmente se denomina Gemini, el chat de Bing pasó a denominarse Copilot, por mencionar algunos ejemplos). Más allá de la denominación, las distintas herramientas como los chatbots con inteligencia artificial conversacional, tal el caso del Chat GPT, motores de búsqueda académica, herramientas de detección de duplicados, de selección de estudios, gestores de referencias, entre otros, pueden facilitar los procesos de revisión sistemática de literatura contribuyendo al trabajo de los investigadores (Chubb *et al.*, 2022; Cárdenas, 2023). Específicamente, el uso de chatbots GPT ha demostrado gran utilidad para interactuar con información disponible, siendo esta una de las razones que llevó a seleccionar esta herramienta. Otros de los motivos que llevaron a su selección fueron la accesibilidad de la misma para cualquier usuario y el auge que tuvo en el ámbito académico y no académico al momento de realizar el presente estudio. Estas herramientas constituyen sistemas conversacionales basados en modelos generativos entrenados con extensas cantidades de texto, lo que les permite generar respuestas coherentes y pertinentes a partir de consultas o mensajes de los usuarios, llamados prompts (Brown *et al.*, 2020). Dichos sistemas tienen la capacidad de aprovechar el conocimiento implícito en los textos, para proporcionar información sobre diversas temáticas, así como de sugerir fuentes, criterios y estrategias de búsqueda.

En este contexto, se plantea que el uso de inteligencia artificial puede agilizar el proceso de las revisiones sistemáticas, ahorrar tiempo y recursos, mejorar la calidad y la rigurosidad de sus revisiones, y obtener resultados más completos y fiables. Además, su uso también podría tener beneficios para el desarrollo de competencias de investigación y para el trabajo colaborativo de los equipos de investigadores (Murillo & Martínez, 2019). Por lo tanto, este artículo propone reportar la experiencia de uso de chatbots GPT como apoyo para los investigadores que realizan revisiones de literatura siguiendo el protocolo SALSA en el ámbito de la investigación educativa, desde una perspectiva crítica para la utilización de la inteligencia artificial. Concretamente, se plantea su utilidad para el desarrollo de la primera fase del protocolo: búsqueda/search, debido a que los chatbots GPT, más específicamente el chatbot GPT-3, pueden facilitar esta fase al interactuar con los investigadores mediante un lenguaje natural y comprensible, ofreciendo orientación, feedback y recomendaciones personalizadas (Lopezosa *et al.*, 2023).

Estas ventajas podrían mejorar la calidad de las ecuaciones de búsqueda, y reducir los sesgos al seleccionar artículos a partir de dichas ecuaciones. Por ejemplo, estas herramientas de inteligencia artificial generativa pueden ayudar a los investigadores a identificar las palabras clave y los operadores booleanos más adecuados para construir las ecuaciones de búsqueda, y a seleccionar las bases de datos más relevantes y actualizadas para consultar según el ámbito y el tipo de literatura que se busca (Lopezosa *et al.*, 2023). No obstante, es necesario mantener estándares metodológicos de calidad durante todo el proceso (van Dijk *et al.*, 2023).

Dentro de los estándares para el diseño de las ecuaciones se indican los criterios: i) validez relacional, que implica consistencia de la ecuación diseñada con el tema del estudio, y ii) validez lógica, relacionada con la coherencia en el uso de los operadores booleanos (Alexander, 2020; Codina, 2023b). Para lograr esto, se recomienda que la ecuación sea diseñada de manera

colaborativa a la interna del equipo de investigación, o bien, que sea sometida a revisión por pares fuera del equipo. Adicionalmente, y en línea con los argumentos previos, la inteligencia artificial podría aportar tanto al diseño como a la validez de las ecuaciones de búsqueda en este contexto.

En general, y aunque se advierte que es necesario un mayor número de estudios al respecto, se considera que la inteligencia artificial generativa puede ser útil a los científicos sociales para desarrollar tareas de investigación menos calificadas o repetitivas (Bail, 2023; Carbajal-Degante *et al.*, 2023). Asimismo, se considera que los investigadores deberían desarrollar competencias en el uso de herramientas de inteligencia artificial para asegurar la calidad de las revisiones teniendo en cuenta el volumen y la dispersión de las fuentes existentes (Caviedes & Roco, 2022).

No obstante, en vista de la cantidad de herramientas desarrolladas a partir de la inteligencia artificial, su uso puede impactar en todo tipo de investigaciones, más allá de las revisiones sistemáticas (ver, por ejemplo, Lopezosa & Codina, 2023). En cualquier escenario de utilización de estas herramientas basadas en inteligencia artificial para la investigación, es necesario tener en cuenta algunas desventajas. Por ejemplo, se ha puntualizado la existencia de sesgo en las respuestas obtenidas, debido al sesgo existente en los datos que alimentan estos modelos (Müller, 2021). Esto puede generar resultados inexactos, y plantea cuestiones éticas sobre el uso de la inteligencia artificial para la investigación (van-Dis *et al.*, 2022).

Especialmente, cabe tener presentes criterios éticos del uso de este tipo de herramientas en la educación y la investigación (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], 2022), y si las investigaciones asistidas por inteligencia artificial son reproducibles (Bail, 2023). Por otra parte, es preciso asegurar la transparencia de los procesos, lo que puede lograrse a partir de i) hacer explícita la contextualización y criterios de la toma de decisiones llevadas a cabo para el diseño y la implementación de una investigación (interpretabilidad), y ii) explicitar las decisiones tomadas en ese contexto (explicación) (Miller, 2019).

Para aportar a lo antes expuesto, en este artículo exploramos el uso de la inteligencia artificial, específicamente del chatbot GPT-3 en la primera fase del protocolo SALSA para revisiones sistemáticas de literatura, con un enfoque en sus posibles contribuciones a la investigación educativa, a partir del uso del término “autonomía de los alumnos” que oficia como palabra clave de ejemplo para la demostración que se propone esta contribución. Esta aportación es parte de un proyecto más abarcativo que busca responder la siguiente pregunta: ¿Cómo contribuye el uso de inteligencia artificial en una revisión sistemática de literatura según el protocolo SALSA en el ámbito de la investigación educativa?

Atendiendo a la necesidad de aportar una mirada crítica sobre el uso de este tipo de herramientas, esta fase, y este artículo en concreto, busca dar respuesta a estas dos preguntas de investigación:

- ¿Qué efectos genera el chatbot GPT-3 sobre la fase de búsqueda del concepto de autonomía de los alumnos?
- ¿Cómo se caracteriza la modalidad de trabajo del equipo de investigación en este marco y en esta fase?

Dados los potenciales usos de la inteligencia artificial y la reciente irrupción de manera generalizada en diversos ámbitos del desarrollo humano, este aporte puede considerarse relevante e innovador en el área de la investigación educativa mediante revisiones sistemáticas de literatura (Gabashvili, 2023; Pávãloaia & Necula, 2023). Teniendo en cuenta el conocimiento existente al respecto, si bien se encuentran publicaciones que involucran la revisión sistemática de tópicos relacionados con la educación (Chiu *et al.*, 2023; Tahiru, 2021; Tate *et al.*, 2023; Yan *et al.*, 2023), en ningún caso aplican la inteligencia artificial generativa al proceso de revisión de literatura. Por otra parte, reportes sobre la aplicación del chatbot GPT para revisiones sistemáticas de literatura aún resultan escasos (Lopezosa *et al.*, 2023; Wang *et al.*, 2023) por lo que, a pesar del creciente interés en el uso de inteligencia artificial en revisiones sistemáticas de literatura, todavía existe una brecha significativa de conocimiento sobre su efectividad y beneficios potenciales, particularmente relacionada al uso del protocolo SALSA.

Método

El abordaje metodológico corresponde a un enfoque cualitativo que contó con la flexibilidad necesaria para que el diseño se construyera durante el proceso de investigación (Lopezosa *et al.*, 2023; Wang *et al.*, 2023). El trabajo supuso la consideración de diferentes tópicos que actuaron de guía durante el estudio: 1) selección del marco de referencia para la revisión sistemática, en este caso el protocolo SALSA (Grant & Booth, 2009; Codina, 2017; García-Peñalvo, 2022), 2) la inclusión de tesauros para validar palabras clave, 3) proceso colaborativo de exploración de alternativas que brinda el chatbot GPT-3 (<https://chat.openai.com/>), diseño de prompts y toma de decisiones en equipo (Cajal *et al.*, 2020). A continuación, se brindan detalles de cada tópico. Si bien se imbrican unos con otros, se los separa para una mejor descripción:

- 1) El protocolo SALSA ofició de guía en el inicio del proceso de revisión. En este caso el foco estuvo puesto en la primera fase, correspondiente a búsqueda/search, con el propósito de responder a la pregunta: "¿qué fuentes utilizamos?" (Lopezosa *et al.*, 2023: 11). Como primer paso en el proceso de esta fase, se procedió a la búsqueda de palabras clave. Una vez identificadas las palabras clave se las validó y jerarquizó recurriendo al uso de tesauros. El siguiente paso ha sido el diseño de la ecuación de búsqueda. El chatbot GPT-3 se utilizó como herramienta, tanto para la obtención de palabras clave, como para el diseño y reelaboración de la fórmula de búsqueda.
- 2) Para la validación se recurrió a tres tesauros: el de la UNESCO, el de ERIC, y el Tesoro Europeo de Educación en el entendido que son los más utilizados en la investigación social y educativa. Una vez que se obtuvieron las palabras relacionadas con la temática seleccionada proporcionadas por el chatbot GPT-3 se procedió a buscarlas en los tres tesauros. De un total de 32 palabras solo 13 fueron halladas en los tesauros según el siguiente detalle: 5 fueron identificadas en el tesoro de la UNESCO, 5 en ERIC y 3 en el Tesoro Europeo de Educación. Como puede apreciarse 19 palabras devueltas por el chatbot a partir de la consulta no aparecieron en ninguno de los tesauros seleccionados.
- 3) El proceso de trabajo en el grupo de investigación involucró diversos aspectos entre los que se destaca el de *exploración de alternativas del chatbot GPT-3* para lo que se llevaron a cabo instancias de prueba de manera sincrónica por parte de las tres investigadoras (códigos I1, I2, I3) con el correspondiente registro e intercambio que nutrió las decisiones tomadas. Corresponde mencionar, además, que el intercambio del que emergen las decisiones se llevó adelante de manera sincrónica y asincrónica, con el uso de herramientas colaborativas (documentos, videollamadas, pizarras, entre otros).

Otro aspecto por destacar es el *diseño de prompts*, para lo que se tomaron como referencia los aportes de Lopezosa *et al.* (2023): "Pedir a ChatGPT que aporte sugerencias de palabras clave para desarrollar la ecuación de búsqueda a partir de los temas de nuestra investigación" (Lopezosa *et al.*, 2023: 13) y "Pedir a ChatGPT que revise y eventualmente mejore una ecuación de búsqueda" (Lopezosa *et al.*, 2023: 14). En este sentido, se diseñaron prompts tanto para la identificación de palabras clave como para la construcción de ecuaciones de búsqueda. En ambos casos se realizaron exploraciones por parte de cada investigadora que socializaron en el equipo. Otra característica del proceso fue la aplicación sucesiva de prompts que permitieron dar órdenes secuenciadas con el propósito de mejorar las respuestas brindadas por el chatbot.

Finalmente, interesa puntualizar un aspecto relevante en el contexto de la investigación en general y en la valoración del uso de la inteligencia artificial generativa en la asistencia a los procesos de revisión sistemática. Se trata de la *toma de decisiones colaborativas* por parte del equipo de investigación en base a los productos y las orientaciones brindados por el chatbot GPT-3. En esta fase del proceso, los aportes de la herramienta se toman como insumos para la toma de decisiones sobre cómo avanzar en el desarrollo de la fase.

Una característica de este trabajo es que el diseño se va construyendo de manera participativa en la medida que se realizan los distintos avances, y las decisiones emergen de la valoración y discusión de las alternativas mediante la colaboración del equipo (Leite *et al.*, 2023;

Murillo & Martínez, 2019). Los productos brindados por el chatbot, en este caso: palabras clave y ecuaciones de búsqueda fueron valorados utilizando como aportes: i) las orientaciones del propio chatbot, ii) las sugerencias de autores referentes (Codina, 2023a; Lopezosa et al., 2023; Wang et al., 2023), iii) otras herramientas utilizadas en investigación como los tesauros, y iv) la experiencia de las investigadoras.

Resultados

Se presentan los hallazgos obtenidos de aplicar los procesos sugeridos por investigaciones y reportes previos (Codina, 2023a; Lopezosa et al., 2023; Wang et al., 2023), organizados a partir de los pasos que se siguieron por parte del equipo de investigación.

Identificación y jerarquización de palabras clave

El éxito en la búsqueda de la información está determinado por la estrategia que se implemente, para lo que es necesario seleccionar palabras clave acordes a las necesidades de información que determine la investigación (García-Peñalvo, 2022). Se detalla cómo se usó el chatbot GPT-3 en el proceso de selección y validación de las palabras clave para este estudio.

Prompts de búsqueda y resultados obtenidos

A partir de la selección del concepto de interés: “autonomía de los alumnos”, al cual se arriba a partir de fuentes teóricas de referencia (Instituto Nacional de Evaluación Educativa [INEEd], 2022), se solicita al chatbot GPT-3 que aporte palabras clave usando los prompts que aparecen en la tabla 1 (Lopezosa et al., 2023).

Tabla 1: Prompts utilizados por las investigadoras para interactuar con el chatbot GPT-3

Investigadora	Prompt
I1	Palabras clave relacionadas con autonomía de los alumnos que aparezcan en el tesoro ERIC, devolver un listado con palabras en inglés y su correspondencia en español y portugués
I2	Palabras clave de autonomía de los alumnos
I3	1: Podrías decirme una palabra clave para una investigación sobre el tema autonomía de los alumnos en el proceso de enseñanza 2: Qué otras palabras clave como sinónimos de "aprendizaje autónomo" puedes sugerir como foco de la investigación.

Nota: Se decidió redactar los prompts con mínimas variantes, pero conservando la misma idea. La codificación de la columna “Investigadora” corresponde a cada involucrada en el estudio y se mantiene a lo largo de este artículo.

Fuente: Elaboración propia.

Como resultados de la búsqueda realizada por las tres investigadoras se obtiene un total de 32 palabras clave, de las que 20 están en el idioma español (autodeterminación, autoeficacia, toma de decisiones, responsabilidad, empoderamiento, autogestión del aprendizaje, libertad académica, aprendizaje independiente, aprendizaje autosuficiente, aprendizaje autónomo guiado, aprendizaje autónomo asistido, aprendizaje autónomo, colaborativo, autoevaluación, motivación intrínseca, aprendizaje significativo, enseñanza personalizada, evaluación formativa, aprendizaje

autoadministrado, aprendizaje autoeficaz). Por otra parte, 12 palabras están en el idioma inglés (autonomous learning, self-efficacy, learner control, student-centered learning, student autonomy, learner autonomy, personalized learning, goal setting, self-directed learning, independent study, active learning, metacognition).

Uso de tesauros como estrategia de validación

La estrategia para valorar la pertinencia de las palabras ofrecidas por el chatbot GPT-3 fue buscar qué palabras estaban presentes en tres tesauros (Cajal *et al.*, 2020). Como fue mencionado en el apartado metodológico, en este caso los tesauros seleccionados fueron: el tesoro de la UNESCO, ERIC y el Tesoro Europeo de Educación (tabla 2).

Tabla 2: Términos ofrecidos por el chatbot GPT-3 identificados en cada tesoro

Términos	Tesauros		
	UNESCO	ERIC	Europeo
Self-directed learning			x
Aprendizaje autoadministrado			x
Aprendizaje autoeficaz			x
Independent study		x	
Active learning		x	
Metacognition		x	
Learner controlled instruction		x	
Personal autonomy		x	
Autoevaluación	x		
Motivación (intrínseca)	x		
Aprendizaje (significativo)	x		
Enseñanza personalizada	x		
Evaluación formativa	x		

Nota: El sombreado corresponde a las palabras seleccionadas posteriormente para solicitar al chatbot GPT-3 una fórmula de búsqueda.

Fuente: Elaboración propia.

Como puede apreciarse en la tabla 2, como resultado de la búsqueda, se obtuvo que 19 palabras no aparecen en ninguno de los tesauros, 5 en el Tesoro ERIC, 5 en el Tesoro de la UNESCO y 3 aparecen en el Tesoro Europeo de Educación.

Decisiones colaborativas para la jerarquización de palabras

Luego de verificar la validez de las palabras obtenidas con el chatbot GPT-3 sometiéndolas a la búsqueda en tesauros, se decide jerarquizar aquellas palabras que pertenecen al tesoro ERIC debido a que ofrece términos en inglés lo que podría ampliar el abanico de publicaciones seleccionadas, tal como lo sugiere Codina (2023a).

Obtención de fórmulas de búsqueda

De acuerdo con lo propuesto por Lopezosa *et al.* (2023) a partir de las palabras clave se solicita al chatbot GPT-3 que ofrezca fórmulas de búsqueda que permitan dar cumplimiento a la primera fase del protocolo SALSA. Estos autores sugieren, además, solicitar al chatbot GPT-3 una revisión y mejora de la fórmula obtenida inicialmente.

Las palabras clave como punto de partida

Con las palabras seleccionadas como orientadoras se construye el prompt (ver tabla 3) para proporcionar al chatbot GPT-3 con el propósito de obtener una fórmula de búsqueda que permita llevar a cabo una revisión sistemática vinculada al tema seleccionado: “autonomía de los alumnos” (Codina, 2023b).

Tabla 3: Prompt diseñado para solicitar al chatbot GPT-3 una ecuación de búsqueda

Prompt utilizado para interactuar con el chatbot	Proponer una ecuación de búsqueda a partir de las siguientes palabras clave: personal autonomy, Learner controlled instruction, Independent study, Active learning, Metacognition
--------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Fuente: Elaboración propia.

El prompt elaborado es ingresado por cada investigadora en el chatbot GPT-3 para obtener fórmulas de búsqueda que se analizaron a fin de tomar decisiones.

Abanico de ecuaciones para la toma de decisiones

Las solicitudes de ecuaciones de búsqueda basadas en los mismos términos devuelven por parte del chatbot propuestas que tienen en común el recurrir a fórmulas con operadores booleanos (Codina, 2023b; Newman & Gough, 2020). En lo que respecta a las diferencias, aparecen diversos criterios referidos, por ejemplo, al grado de inclusión es decir según integre operadores booleanos “OR” cuyo resultado de búsqueda resulta menos abarcativo por ser excluyente o “AND” que en lugar de excluir agrega alternativas. Las respuestas también difieren en la cantidad de información. La tabla 4 muestra los resultados obtenidos por las tres investigadoras que tomaron el mismo prompt como punto de partida, obteniendo diferentes ecuaciones en el chatbot GPT-3.

Tabla 4: Ecuaciones de búsqueda ofrecidas por el chatbot GPT-3 a cada investigadora

Prompt utilizado para interactuar con el chatbot	Respuesta generada por el chatbot a cada investigadora		
	I1	I2	I3
Proponer una ecuación de búsqueda a partir de las siguientes palabras clave: personal autonomy, learner controlled instruction, independent study, active learning, metacognition	"personal autonomy" OR "learner controlled instruction" OR "independent study" OR "active learning" OR "metacognition"	(personal autonomy OR learner autonomy) AND (learner controlled instruction OR independent study OR active learning OR metacognition)	("autonomía personal" O "instrucción controlada por el alumno" O "estudio independiente" O "aprendizaje activo" O "metacognición") Y (educación O aprendizaje O enseñanza O instrucción O pedagogía)

Nota: La investigadora tiene activado el traductor en el navegador durante la interacción con el chatbot GPT-3.

Fuente: Elaboración propia.

Lo mencionado en cuanto a las diferentes fórmulas obtenidas a partir de un mismo prompt, también ocurre con las búsquedas sucesivas por parte de un mismo usuario (tabla 5).

Tabla 5: Ecuaciones de búsqueda ofrecidas por el chatbot GPT-3 a un mismo usuario de Open AI en sucesivas búsquedas

Prompt utilizado para interactuar con el chatbot	Respuesta 1	Respuesta 2	Respuesta 3
Proponer una ecuación de búsqueda a partir de las siguientes palabras clave: personal autonomy, learner controlled instruction, independent study, active learning, metacognition	(personal autonomy OR learner autonomy) AND (learner controlled instruction OR independent study OR active learning OR metacognition	(personal autonomy OR learner autonomy OR student autonomy) AND (learner controlled instruction OR self-directed learning OR independent study OR active learning) AND (metacognition OR self-regulated learning)	("personal autonomy" OR "learner autonomy" OR "student autonomy") AND ("learner controlled instruction" OR "self-directed learning" OR "independent study" OR "problem-based learning" OR "active learning") AND ("metacognition" OR "self-regulated learning" OR "goal setting" OR "intrinsic motivation")

Nota: Esta tabla fue realizada a partir de las interacciones de la Investigadora I2 con la herramienta.
Fuente: Elaboración propia.

Si bien las respuestas varían, tanto si son realizadas más de una vez por el mismo investigador, como si la realizan investigadores diferentes, el análisis del abanico de propuestas para la toma de decisiones de cuál o cuáles utilizar, se enriquece si se trata de una decisión colaborativa tomada por el equipo.

Otro aspecto que interesa en relación con la búsqueda es la posibilidad de solicitar al chatbot GPT-3 que mejore la fórmula entregada, como proponen Lopezosa *et al.* (2023). Es posible afirmar que, al igual que ocurriera con la solicitud de fórmulas a partir de palabras clave, en la respuesta a la solicitud de mejorar una fórmula el chatbot GPT-3 también devuelve alternativas con características diferentes y con más información respecto a la fórmula inicial. La tabla 6 muestra las mejoras a las diferentes fórmulas de búsqueda ofrecidas por el chatbot GPT-3 a cada investigadora, según la versión inicial que cada una había obtenido.

Tabla 6: Revisiones de las fórmulas de búsqueda ofrecidas por chatbot GPT-3 a cada investigadora

Investigadora	Respuesta inicial del chatbot	Respuesta al prompt: Revisa la siguiente ecuación de búsqueda y propone mejoras
11	"personal autonomy" OR "learner controlled instruction" OR "independent study" OR "active learning" OR "metacognition"	("personal autonomy" OR "learner controlled instruction" OR "independent study" OR "active learning" OR "metacognition") AND ("educational psychology" OR "self-directed learning") AND (year:[2016 TO 2023]) AND (database: "Education Source"
12	(personal autonomy OR learner autonomy) AND (learner controlled instruction OR independent study OR active learning OR metacognition	personal autonomy OR learner autonomy) AND (learner controlled instruction OR independent study OR active learning OR metacognition
13	("autonomía personal" O "instrucción controlada por el alumno" O "estudio independiente" O "aprendizaje activo" O "metacognición") Y (educación O aprendizaje O enseñanza O instrucción O pedagogía)	("autonomía personal" O "aprendizaje autodirigido" O "aprendizaje autorregulado" O "instrucción controlada por el alumno" O "estudio independiente" O "aprendizaje activo" O "aprendizaje práctico" O "aprendizaje experiencial" O "metacognición ") Y (artículo O libro O video) Y fecha: >= 2019

Nota: La investigadora tiene activado el traductor en el navegador durante la interacción con el chatbot GPT-3.

Fuente: Elaboración propia.

Las fórmulas que ofrece el chatbot GPT-3 son acompañadas, además, con argumentos sobre cómo se compone la fórmula y en el caso de que se solicite mejoras, explica en qué consiste la mejora. En el siguiente apartado se detallan las características de los argumentos que ofrece el chatbot GPT-3 acompañando la fórmula ofrecida, el que oficia como orientador en la toma de decisiones.

Orientaciones complementarias

Las sugerencias ofrecidas por el chatbot GPT-3 refieren a aspectos coincidentes con los propuestos por autores como Alexander (2020); Codina (2023b) y Newman & Gough (2020), como son: el uso de palabras clave, operadores booleanos, comillas y acotación de fechas. Al momento de analizar el contenido de las orientaciones brindadas por el chatbot GPT-3 se sugiere precisar la búsqueda como, por ejemplo, utilizando determinados operadores booleanos, o la inclusión de rangos de años.

En otros casos, las sugerencias hacen referencia a la rigurosidad académica, por ejemplo, cuando indica que se recurra a bases de datos específicas. También recomienda incorporar variantes en el uso de palabras clave ya sea por el agregado de sinónimos, términos relacionados y/o términos específicos. Para cada caso ofrece algunos ejemplos a partir de los que es posible valorar su pertinencia. Este último aspecto supone el tipo de sugerencia que requiere más expertise por parte de los investigadores en el tema, ya que deberán valorar la pertinencia de los conceptos sugeridos. Como ocurre en las sugerencias de variantes en las palabras clave, todas las sugerencias están acompañadas con la ejemplificación correspondiente, brindando más insumos a considerar en la toma de decisiones informada.

La tabla 7 presenta las diferentes indicaciones realizadas por la herramienta, agrupándolas por tipo de sugerencia y especificando los ejemplos que brinda el chatbot GPT-3 para cada una.

Tabla 7: Tipos de sugerencias ofrecidas por el chatbot GPT-3 a cada investigadora

Tipo de sugerencia	Sugerencia	Respuesta del chatbot GPT-3
Precisar la búsqueda	Operadores booleanos	No olvides <u>agregar los operadores booleanos</u> apropiados, como "AND" y "OR", para ajustar la búsqueda según tus necesidades (I1)
	Comillas para frases exactas	<u>Utiliza comillas para frases exactas</u> : Si estás buscando una frase exacta, es recomendable utilizar comillas alrededor de la frase. <u>Por ejemplo</u> , si estás buscando información sobre "learner controlled instruction", deberías poner la frase entre comillas para asegurarte de que los resultados incluyan la frase exacta (I2)
	Rango de años	<u>Limita el rango de años</u> : Puedes limitar la búsqueda a un rango de años específico para asegurarte de que los resultados sean relevantes y actualizados (I1)
Rigurosidad académica	Bases de datos	<u>Especifica la base de datos</u> : Si estás buscando en una base de datos específica, es recomendable incluir el nombre de la base de datos en la búsqueda. <u>Por ejemplo</u> , si estás buscando en la base de datos de PubMed, podrías incluir "PubMed" como un término adicional en la búsqueda (I2)
Variantes o agregados en el uso de palabras clave	Agregar palabras clave	Añade palabras clave adicionales: Puedes considerar agregar palabras clave adicionales que sean relevantes para tu búsqueda, como "educational psychology" o "self-directed learning" (I1)
	Sinónimos	<u>Agregar sinónimos</u> : Puede ser útil agregar sinónimos de las palabras clave para ampliar la búsqueda. <u>Por ejemplo</u> , para "autonomía personal", también puede usar "aprendizaje autodirigido" o "aprendizaje autorregulado". Para "aprendizaje activo", puede agregar "aprendizaje práctico" o "aprendizaje experiencial". Al agregar los sinónimos, se pueden encontrar recursos relevantes que no usen exactamente las palabras clave especificadas (I3)
	Términos relacionados	<u>Incluye términos relacionados</u> : Para obtener una búsqueda más completa, es recomendable incluir términos relacionados con tu tema de investigación. <u>Por ejemplo</u> , podrías agregar términos como "self-regulated learning", "goal setting", "motivation", "intrinsic motivation", etc. (I2)
	Términos específicos	<u>Utiliza términos más específicos</u> : Los términos "personal autonomy" y "learner autonomy" son bastante amplios. Para obtener resultados más precisos, es posible que desees utilizar términos más específicos que estén relacionados con tu tema de investigación. <u>Por ejemplo</u> , podrías buscar "student autonomy" o "autonomous learning" en lugar de "learner autonomy" (I2)
Especificar el tipo de recurso		<u>Especificar el tipo de recurso</u> : Excesivamente de lo que esté buscando, puede especificar el tipo de recurso que desea encontrar. <u>Por ejemplo</u> , puede buscar solo artículos académicos o libros, o puede buscar recursos multimedia como videos o presentaciones. Esto también se puede hacer a través de filtros de búsqueda en los recursos en línea (I3)

Fuente: Elaboración propia.

Cuando se le solicita al chatbot GPT-3 que mejore una ecuación de búsqueda, las sugerencias incluyen más información si se lo compara con interacción con el primer prompt de solicitud de ecuación, por lo que se puede decir que ocurre una expansión. Asimismo, se destaca que el chatbot GPT-3 incluye los aspectos sugeridos en la fórmula mejorada.

Otro aspecto que interesa analizar es la recurrencia con la que aparecen las sugerencias en las búsquedas de las tres investigadoras (tabla 8).

Tabla 8: Palabras ofrecidas por el chatbot GPT-3 como sugerencias, según su recurrencia en las diferentes interacciones con la herramienta

Sugerencias indicadas por el chatbot	Investigadora			Presente en respuestas de dos de las investigadoras	Presente en respuestas de las tres investigadoras
	I1	I2	I3		
Operadores booleanos	X	X	X		
Bases de datos	X	X			
Rango de años	X		X		
Comillas		X			
Términos relacionados	X	X	X		
Sinónimos		X			
Términos específicos			X		
Recursos			X		

Fuente: Elaboración propia.

Mientras que todas reciben sugerencias de utilizar operadores booleanos y el ofrecimiento de términos relacionados o sinónimos, la sugerencia de recurrir a bases de datos y acotar las fechas aparecen en dos de las búsquedas, en tanto otros aspectos como el uso de comillas aparece solo en una de las búsquedas, tal como lo muestra la tabla 8.

Selección de la fórmula de búsqueda

Finalmente corresponde aludir a la selección de la fórmula de búsqueda que se va a utilizar. A continuación, se detallan los criterios utilizados en dicha selección:

- No incorporar comillas en las palabras clave que se compongan de un único término (Codina, 2023a).
- Usar todas las palabras en el idioma inglés (Codina, 2023a).
- Seleccionar el operador booleano “OR” entre todas las palabras clave.
- Valorar las sugerencias de palabras clave que realiza el chat que no fueron propuestas por el investigador ("educational psychology" OR "self-directed learning") y el operador booleano que sugiere para vincular las palabras que propone con las que le habían sido sugeridas (AND).
- Incorporar el operador booleano “AND” para:
- Demarcar temporalmente la búsqueda.
- Seleccionar artículos dado que la revisión sistemática apunta a trabajos empíricos (“journal articles”).
- Descartar la sugerencia de base de datos porque la búsqueda se realiza directamente en la base de datos deseada, por ejemplo, Scopus.

La fórmula final (figura 1) no se condice con una respuesta del chatbot sino que emerge de la construcción que realizan las investigadoras valorando los diferentes criterios estipulados.

Figura 1: Fórmula a la que se llega como resultado del proceso descripto

("personal autonomy" OR "learner controlled instruction" OR "independent study" OR "active learning" OR metacognition) AND (year >= 2019) AND (source: "journal articles")

Nota: Esta fórmula ha sido construida por las investigadoras, valorando las sugerencias realizadas por el chatbot GPT-3 en base a los criterios acordados de forma colaborativa.

Fuente: Elaboración propia.

Conclusiones

Este artículo presenta una experiencia de uso del chatbot GPT-3 para facilitar el proceso de búsqueda en una revisión sistemática de literatura sobre un concepto relacionado con la educación, siguiendo el protocolo SALSA. El propósito ha sido explorar las posibilidades que ofrece la herramienta como apoyo al investigador en la etapa. Con la asistencia del chatbot se diseñó una ecuación de búsqueda usando palabras clave, contrastadas con los tesauros, a ser aplicada en dos bases de datos.

Como conclusiones generales puede establecerse que, al combinar la experiencia de los investigadores con las capacidades de procesamiento del lenguaje natural del chatbot GPT-3, ha sido posible agilizar el proceso de diseño de ecuaciones de búsqueda para revisiones sistemáticas de literatura. A nivel de las competencias de investigación, la herramienta ha ayudado a las investigadoras a desarrollar estrategias a partir del uso de recursos especializados como tesauros. Esto mejoró la calidad de la ecuación de búsqueda, lo que se espera redunde en la realización de una revisión más precisa y completa sobre el concepto amplio de "autonomía de los alumnos" en el ámbito educativo.

A la vez, el uso de la inteligencia artificial ha reportado beneficios para el trabajo colaborativo del equipo de investigación. Las investigadoras han compartido y discutido sus estrategias de búsqueda, favoreciendo intercambios críticos y fomentando espacios de mayor aprendizaje y colaboración en el trabajo en equipo. En este contexto, puede afirmarse que el chatbot GPT-3 tiene el potencial de mejorar significativamente la capacidad de los investigadores para realizar revisiones sistemáticas de la literatura y otras formas de investigación, en la medida que fomenta desafíos para el diseño de estudios asistidos por inteligencia artificial generativa y propone la urgencia de procesos más críticos y la necesidad de colaboración (por ejemplo, ver Lopezosa et al., 2023).

Es importante tener en cuenta que la inteligencia artificial en el proceso reportado en este artículo no reemplazó la labor humana en la realización de revisiones sistemáticas, sino que la complementó. Es necesario contar con la experiencia y el juicio crítico de los investigadores para interpretar los resultados obtenidos mediante la herramienta y tomar decisiones informadas.

Como limitaciones y a modo de sugerencia para otras indagaciones, es conveniente realizar la búsqueda utilizando los prompts en el idioma inglés y sin activar el traductor automático a otros idiomas (como, por ejemplo, español), en el entendido que el chatbot GPT-3 resulta más preciso en el uso de ese idioma. Es necesario mencionar además que en el estudio solo se construyó una ecuación de búsqueda compuesta cuando autores como Codina (2023b) proponen la idea de que se pueden diseñar y aplicar varias ecuaciones simples para luego combinar resultados. Por otra parte, el chatbot GPT-3 ha tenido actualizaciones por lo que el momento en que se realizaron las búsquedas podría suponer diferencias con lo propuesto en otros estudios y dificultar su comparación. Si bien, en coincidencia con Wang et al. (2023), se propone que la herramienta puede actuar como facilitador del proceso de búsqueda, el trabajo realizado solo

aborda aspectos cualitativos por lo que no fue cuantificada su efectividad. Se coincide también con Wang et al. (2023), respecto a los problemas que conlleva el uso del chatbot respecto a la variabilidad y dificultades de reproducibilidad.

Con respecto a las proyecciones, el equipo de investigación seguirá desarrollando la revisión sistemática según los pasos del protocolo SALSA, con el uso de inteligencia artificial generativa como asistente, comenzando por la ejecución de la ecuación de búsqueda a la que se ha arribado mediante el procedimiento descrito en este artículo.

Al reportar los hallazgos de esta experiencia, se espera proporcionar nuevas perspectivas sobre el potencial de la inteligencia artificial generativa en la primera fase del protocolo SALSA para revisiones sistemáticas de literatura. En última instancia, el propósito es contribuir a la comunidad científica mediante aportes al desarrollo continuo de metodologías más efectivas y eficientes para diseñar e implementar la investigación.

Referencias

- ALEXANDER, P. (2020). Methodological Guidance Paper: The art and science of quality systematic reviews. *Review of Educational Research*, 90(1), 6–23. Recuperado de: <https://doi.org/10.3102/0034654319854352>
- BAIL, C. (2023, May 12). Can Generative AI Improve Social Science? *SocArXiv*. Recuperado de: <https://doi.org/10.31235/osf.io/rwtzs>
- BOOTH, A., SUTTON, A., & PAPAIOANNOU, D. (2016). *Systematic approaches to a successful literature review* (2° ed.). London: Sage.
- BÖTTCHER, F., & THIEL, F. (2017). Evaluating research-oriented teaching: a new instrument to assess university students' research competences. *Higher Education*, 75(1), 91-110. Recuperado de: <https://doi.org/10.1007/s10734-017-0128-y>
- BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., AGARWAL, S., HERBERT-VOSS, A., KRUEGER, G., HENIGHAN, T., CHILD, R., RAMESH, A., ZIEGLER, D., WU, J., WINTER, C., HESSE, C., CHEN, M., SIGLER, E., LITWIN, M., GRAY, S., CHESS, B., CLARK, J., BERNER, C., MCCANDLISH, S., RADFORD, A., SUTSKEVER, I., y AMODEI, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems* (Vol. 33, pp. 1877-1901). Curran Associates. Recuperado de: <https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- CAJAL, B., JIMENEZ, R., GERVILLA, E., & MONTANO, J. J. (2020). Doing a systematic review in health sciences. *Clínica y Salud*, 31(2), 77-83. Recuperado de: <https://doi.org/10.5093/clysa2020a15>
- CARBAJAL-DEGANTE, E., HERNÁNDEZ, M., SÁNCHEZ-MENDIOLA, M. (2023). Hacia revisiones de la literatura más eficientes potenciadas por inteligencia artificial. *Investigación en Educación Médica*, 12(47), 111-119. Recuperado de: <https://doi.org/10.22201/fm.20075057e.2023.47.23526>
- CÁRDENAS, J. (17 de marzo de 2023). Cómo hacer una revisión de literatura con una herramienta de IA y redes bibliométricas. *Networkianos. Blog de Sociología*. Recuperado de: <https://networkianos.com/como-hacer-una-revision-de-literatura-con-una-herramienta-de-ia-y-redes/>
- CAVIEDES, M., & ROCO, Á. (2022). Algoritmos de búsqueda e inteligencia artificial, una ayuda imprescindible en el desarrollo de revisiones sistematizadas. *Nutrición Hospitalaria*, 39(6), 1434-1435. Recuperado de: <https://dx.doi.org/10.20960/nh.04336>
- CHIU, T., XIA, Q., ZHOU, X., CHAI, C., & CHENG, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4(2023), 100118. Recuperado de: <https://doi.org/10.1016/j.caeai.2022.100118>
- CHUBB, J., COWLING, P., & REED, D. (2022). Speeding up to keep up: exploring the use of AI in the research process. *AI & Society*, 37, 1439–1457. Recuperado de: <https://doi.org/10.1007/s00146-021-01259-0>
- CODINA, L. (20 de abril de 2017). Revisiones de la literatura y cómo llevarlas a cabo con garantías: systematicreviews y SALSA framework. *Lluís Codina*. Recuperado de: <https://www.lluiscodina.com/revision-sistemica-salsa-framework/>
- _____ (2020). Revisiones bibliográficas sistematizadas en Ciencias Humanas y Sociales. 1: Fundamentos. En C. Lopezosa, J. Díaz-Noci & L. Codina, (Eds.). *Anuario de métodos de investigación en Comunicación Social* (n.1, pp. 50-60). Barcelona: Universitat Pompeu Fabra. Recuperado de: <https://doi.org/10.31009/metodos.2020.i01.05>
- _____ (26 de junio de 2023a). Web of Science y Scopus: cómo evitar los errores más frecuentes en revisiones de la literatura. *Lluís Codina*. <https://www.lluiscodina.com/web-of-science-scopus-errores/>
- _____ (24 de abril de 2023b). Validez de las ecuaciones de búsqueda en revisiones sistemáticas: validez relacional y lógica. *Lluís Codina*. Recuperado de: <https://www.lluiscodina.com/ecuaciones-de-busqueda-operadores-booleanos/>

DE KOCK, S., STIRK, L., ROSS, J., DUFFY, S., NOAKE, C., & MISSO, K. (2021). Systematic review search methods evaluated using the Preferred Reporting of Items for Systematic Reviews and Meta-Analyses and the risk of bias in systematic reviews tool. *International Journal of Technology Assessment in Health Care*, 37(1), e18. Recuperado de: <https://doi.org/10.1017/S0266462320002135>

GABASHVILI, I. (2023). The impact and applications of ChatGPT: a systematic review of literature reviews. *Aurametrix*. Recuperado de: <https://doi.org/10.48550/arXiv.2305.18086>

GARCÍA-PEÑALVO, F. (2022). Desarrollo de estados de la cuestión robustos: Revisiones Sistemáticas de Literatura. *Education in the Knowledge Society (EKS)*, 23, e28600. Recuperado de: <https://doi.org/10.14201/eks.28600>

GRANT, M. & BOOTH, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26, 91–108. Recuperado de: <https://doi.org/10.1111/j.1471-1842.2009.00848.x>

HADDAWAY, N. (19 de octubre de 2020). 8 common problems with literature reviews and how to fix them. *LSE Impact Blog*. Recuperado de: <https://blogs.lse.ac.uk/impactofsocialsciences/2020/10/19/8-common-problems-with-literature-reviews-and-how-to-fix-them/>

HART, C. (2018). *Doing a literature review* (2nd Ed.). London: Sage.

INSTITUTO NACIONAL DE EVALUACIÓN EDUCATIVA (INEED). (2022). *Reporte de Aristas 9. ¿Qué factores escolares contribuyen a que algunos centros tengan desempeños mayores a los esperados?* Recuperado de: <https://www.ineed.edu.uy/images/Aristas/Publicaciones/Reportes/Reporte-9-Factores-escolares-centros-desempenos-mayores-esperados.pdf>

LEITE, R., RIERA, R., MARCÍLIO, G., MARTINS, K., GOMES, L., RESENDE, G., RODRIGUES DE OLIVEIRA, F., & CABRERA, A. (2023). Many systematic reviews with a single author are indexed in PubMed. *Journal of Clinical Epidemiology*, 156(2023), 124-126. <https://doi.org/10.1016/j.jclinepi.2023.01.007>

LOPEZOSA, C., & CODINA, L. (2023). ChatGPT y software CAQDAS para el análisis cualitativo de entrevistas: pasos para combinar la inteligencia artificial de OpenAI con ATLAS.ti, Nvivo y MAXQDA. Serie Editorial DigiDoc. PCUV04/2023. Barcelona: Universitat Pompeu Fabra.

LOPEZOSA, C., CODINA, L., & FERRAN-FERRER, N. (2023). *ChatGPT como apoyo a las systematicscopingreviews: integrando la inteligencia artificial con el framework SALSA*. Colección del CRICC. Universitat de Barcelona. Recuperado de: <http://hdl.handle.net/2445/193691>

MILLER, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267(Feb. 2019), 1-38. Recuperado de: <https://doi.org/10.1016/j.artint.2018.07.007>

MÜLLER, V. (2021). Ethics of artificial intelligence. In A. Elliott (Ed.), *The Routledge social science handbook of AI* (pp. 122-137). London: Routledge.

MURILLO, F., & MARTÍNEZ, C. (2019). Una mirada a la investigación educativa en América Latina a partir de sus artículos. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, 17(2), 5-25. Recuperado de: <https://doi.org/10.15366/reice2019.17.2.001>

NEWMAN, M., & GOUGH, D. (2020). Systematic reviews in educational research: Methodology, perspectives and application. In O. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond & K. Buntins (Eds.), *Systematic reviews in educational research. Methodology, perspectives and application* (pp. 3-22). Wiesbaden: Springer.

ONWUEGBUZIE, A., & FRELS, R. (2016). *Seven steps to a comprehensive literature review: a multimodal and cultural approach*. London: Sage.

ORGANIZACIÓN DE LAS NACIONES UNIDAS PARA LA EDUCACIÓN, LA CIENCIA Y LA CULTURA [UNESCO]. (2022). *Recomendación sobre la ética de la inteligencia artificial*. (Documento de programa/reunión). Francia. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

PĂVĂLOAIA, V., & NECULA, S. (2023). Artificial intelligence as a disruptive technology - A systematic literature review. *Electronics*, 12, 1102. Recuperado de: <https://doi.org/10.3390/electronics12051102>

POLANIN, J., MAYNARD, B., & DELL, N. (2017). Overviews in education research: A systematic review and analysis. *Review of Educational Research*, 87(1), 172–203. Recuperado de: <https://doi.org/10.3102/0034654316631117>

QUESTA-TORTEROLO, M., CABRERA BORGES, C., & TEJERA TECHERA, A. (2022). ¿Qué es la investigación? En A. Tejera Techera & M. Questa-Tortero (Coords.), *Competencias y herramientas de investigación aplicada con foco en la gestión educativa* (pp. 9-37). Uruguay: Universidad ORT.

TAHIRU, F. (2021). AI in education: A systematic literature review. *Journal of Cases on Information Technology*, 23(1), 1-20. Recuperado de: <https://doi.org/10.4018/JCIT.2021010101>

TATE, T., DOROUDI, S., RITCHIE, D., XU, Y., & WARSCHAUER, M. (2023). Educational research and ai-generated writing: confronting the coming tsunami. *EdArXiv*. Recuperado de: <https://doi.org/10.35542/osf.io/4mec3>

VAN DIJK, S., BRUSSE-KEIZER, M., BUCSÁN, C., VAN DER PALEN, J., DOGGEN, C., & LENFERINK, A. (2023). Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open*, 13, e072254. Recuperado de: <https://doi.org/10.1136/bmjopen-2023-072254>

VAN-DIS, E., BOLLEN, J., ZUIDEMA, W., VAN ROOIJ, R., & BOCKTING, C. (2023). ChatGPT: five priorities for research. *Springer Nature*, 614, 224-226. Recuperado de: <https://www.nature.com/articles/d41586-023-00288-7>

WANG, S., SCELLS, H., KOOPMAN, B., & ZUCCON, G. (2023). Can ChatGPT write a good boolean query for systematic review literature search? *ArXiv*. Recuperado de: <https://doi.org/10.48550/arXiv.2302.03495>

YAN, L., SHA, L., ZHAO, L., LI, Y., MARTINEZ-MALDONADO, R., CHEN, G., LI, X., JIN, Y., & GAŠEVIĆ, D. (2023). Practical and ethical challenges of large language models in education: a systematic literature review. *ArXiv*. Recuperado de: <https://doi.org/10.48550/arXiv.2303.13379>

ZAWACKI-RICHTER, O., KERRES, M., BEDENLIER, S., BOND, M., & BUNTINS, K. (2020). *Systematic reviews in educational research: Methodology, perspectives and application*. Wiesbaden: Springer Nature.

Autora.

Claudia Cabrera Borges

Consejo de Formación en Educación, Administración Nacional de Educación Pública (ANEP-CFE). Instituto de educación de la Universidad ORT, Uruguay.

Doctora en Educación por la Universidad ORT. Docente de Didáctica del profesorado de Biología (ANEP-CFE). Docente y tutora de Maestría en la Universidad ORT.

E-mail: claudiaanahi@gmail.com

Citado.

CABRERA BORGES, Claudia (2025). Experiencia de uso de Chat GPT-3 como asistente de búsqueda en revisiones sistemáticas de literatura. *Revista Latinoamericana de Metodología de la Investigación Social – ReLMIS*, N°29, Año 15, pp. 83-99.

Plazos.

Recibido: 13/08/2023. Aceptado: 20/02/2024.

Reseña bibliográfica:

Repensando la Metodología en Ciencias Sociales: desafíos de la creación y validación de indicadores

Reseña del libro:

DE SENA, A. (2018). *Políticas sociales y microemprendimientos socio-productivos: Indicadores de sustentabilidad*
Buenos Aires: Editorial Ciccus.

Adriana Elisabet Romiti

Es conocido que trabajar en el campo de las Ciencias Sociales no es tarea sencilla dado que el objeto a analizar es el comportamiento humano y su desarrollo. Es por esto que el libro "*Políticas sociales y microemprendimientos socio-productivos: Indicadores de sustentabilidad*", de Angélica De Sena, es el resultado de dos décadas de investigación cuyo objetivo central es "construir y validar indicadores de sustentabilidad de los microemprendimientos que permitieran establecer su grado de desarrollo y sus posibilidades de reproducción en el tiempo" (De Sena, 2018:15). La autora examina cómo las relaciones sociales, económicas y políticas de los actores involucrados influyen tanto en el desarrollo como en la permanencia de estos emprendimientos a lo largo de los últimos treinta años en Argentina.

El libro inicia su recorrido con un breve prefacio donde la autora expone los motivos que la llevaron a la realización del libro, el camino recorrido, las problemáticas surgidas y los desafíos encontrados al tratar de establecer la sustentabilidad de los microemprendimientos. A continuación, en la introducción, De Sena presenta el marco teórico que fundamenta su estudio, así como los aspectos metodológicos adoptados para abordar la temática central del libro. Se describe el enfoque de investigación utilizado, incluyendo las herramientas y técnicas utilizadas para recolectar y analizar los datos. Asimismo, la autora ofrece una visión preliminar de los resultados obtenidos durante el proceso investigativo, los cuales serán desarrollados más profundamente a lo largo de la obra. El libro está estructurado en siete capítulos, cuyo propósito es contribuir a la formación de nuevos indicadores, y su validación, para determinar la sustentabilidad de microemprendimientos en Latinoamérica.

El primer capítulo presenta diversas definiciones del término 'políticas sociales' y analiza su relación con la sociedad, la política y la economía. Asimismo, examina el impacto de dichas políticas en el desarrollo de microemprendimientos, los cuales pueden proporcionar a los individuos mecanismos de subsistencia y/o desarrollo. De este modo, se exploran alternativas de inserción de los microemprendimientos en la Población Económicamente Activa desde una perspectiva diferente a la establecida por el modelo económico hegemónico.

El segundo capítulo analiza las dificultades metodológicas que debieron enfrentarse durante la investigación para definir un marco de estudio para los microemprendimientos, dado que éstos presentan una gran heterogeneidad en términos de composición, origen, localización,

nivel de producción y acceso al crédito, entre otras dimensiones. Asimismo, examina la evolución del microemprendimiento en Argentina, su rol dentro de la estructura económica del país, la participación de la mujer en este ámbito y la importancia del respaldo estatal para la sostenibilidad y el desarrollo de las economías locales.

El tercer capítulo desarrolla el diseño metodológico de la investigación y el proceso de validación de los datos recopilados, que luego se utilizarán para la construcción de indicadores cualitativos y cuantitativos que posibiliten analizar la sostenibilidad de los microemprendimientos a lo largo del tiempo. Dada la complejidad del objeto de estudio, se adopta un enfoque multi-método que ajusta los objetivos de la investigación para garantizar la confiabilidad de los resultados. El propósito principal del capítulo es exponer el proceso metodológico involucrado en el armado, construcción y validación de indicadores para medir los distintos grados de desarrollo de los microemprendimientos con capacidad de producción y reproducción, ya sean de propiedad individual o asociativa, con o sin participación del trabajo familiar, y con distintos niveles de apoyo estatal a través de políticas sociales sin poseer datos previos. El diseño metodológico y validación se estructuró en cinco fases secuenciales. En la primera, la autora aplica el método autoetnográfico con el propósito de delimitar el objeto de estudio y determinar los indicadores más adecuados para evaluar la sustentabilidad de un microemprendimiento. En la segunda, se establece el marco muestral sobre el cual se llevará a cabo el proceso de indagación, con el fin de obtener los indicadores de sustentabilidad más relevantes. En la tercera fase, se busca constituir cualitativamente los parámetros de alcance de los indicadores definidos en las etapas anteriores o, en su caso, corregirlos para asegurar su utilidad en la obtención de información primaria pertinente. En la cuarta, se procede a la validación de los indicadores propuestos mediante su aplicación práctica. Finalmente, en la quinta fase, se obtienen datos secundarios a través del análisis de los resultados obtenidos en la etapa anterior, con el objetivo de generar información secundaria que respalde o refute los datos primarios recogidos.

El cuarto capítulo examina el uso de la autoetnografía como estrategia metodológica para articular la trayectoria profesional de la autora con su experiencia personal en el estudio de los microemprendimientos. Este enfoque permite definir los parámetros necesarios para la delimitación de las prácticas de intervención, así como para la recolección y operacionalización de los datos. Desde esta perspectiva, De Sena identifica un elemento central en la evolución de los microemprendimientos a lo largo del tiempo: la transformación del sujeto emprendedor y los factores que inciden en su motivación para emprender. En este sentido, el comportamiento del emprendedor se vincula con tres dimensiones clave que influyen en el éxito o fracaso del microemprendimiento: el contexto territorial en el que se desarrolla la actividad económica, el tipo de actividad productiva y la configuración interna y externa del emprendimiento. Estos elementos constituyen los indicadores fundamentales para analizar la sustentabilidad.

El quinto capítulo está dedicado al análisis de los datos cuantitativos obtenidos en las fases previas, al tiempo que desarrolla una nueva metodología para su interpretación. Dicha metodología se fundamenta en tres parámetros específicos que permiten una mejor valoración del peso de cada indicador seleccionado. Estos parámetros incluyen: (1) la ubicación de los microemprendimientos (a nivel Municipal), (2) la perdurabilidad en el tiempo según evaluaciones previas de equipos técnicos, y (3) las variables y dimensiones asociadas a la superficie de inclusión de los indicadores base de la investigación. Para garantizar una interpretación clara y precisa de los datos, éstos fueron categorizados en correspondencia con los indicadores obtenidos en la fase anterior. El criterio central de selección considera exclusivamente aquellos microemprendimientos que han recibido financiamiento del Estado Nacional a través del Municipio. El análisis sigue una estructura progresiva que permite la comparación de datos en distintas unidades de observación. Se inicia con el estudio de las características generales de cada Municipio, seguido por un análisis específico de cada microemprendimiento y, finalmente, una descomposición detallada de la composición de los integrantes de cada unidad productiva. Los hallazgos obtenidos se presentan tanto de manera detallada como sintética. Para ello, se incorpora un cuadro sinóptico que resume los principales resultados y se complementa con gráficos que evidencian la relevancia de cada indicador para la sustentabilidad de los microemprendimientos analizados. Esta combinación de representaciones permite una mejor comprensión y facilita la comparación de los datos en relación con la metodología aplicada.

El sexto capítulo, desarrolla el proceso de validación de los indicadores de sustentabilidad definidos en etapas previas buscando transformar los datos cualitativos obtenidos en cuantitativos, estructurando su análisis en torno a tres ejes fundamentales: territorio, características del microemprendimiento y conformación interna. A partir de un enfoque metodológico riguroso, se busca garantizar que estos indicadores sean aplicables y efectivos en la evaluación de distintos microemprendimientos. El proceso de validación se organizó de manera secuencial. En primer lugar, se realizaron entrevistas semiestructuradas con expertos técnicos y profesionales de programas socio-productivos en la Provincia de Buenos Aires. Luego, se implementó un cuestionario estructurado en línea, dirigido a especialistas en promoción de microemprendimientos a nivel nacional y provincial. En una etapa posterior se produjeron datos cuantitativos que permitieron validar el proceso generado durante la etapa de indagación respecto al establecimiento del marco muestral y de los indicadores de sustentabilidad. Para eso, se realizaron encuestas *on-line* a técnicos y profesionales que se desempeñaban en el ámbito nacional y/o de la Provincia de Buenos Aires en la promoción de microemprendimientos: el cuestionario estructurado elaborado permitió obtener datos cuantitativos para validar el proceso. Por último, se sistematizaron las entrevistas realizadas a individuos que forman parte de algún microemprendimiento y cuya permanencia se logra evidenciar. A medida que se avanza en la lectura del capítulo, la autora va exponiendo los resultados y los compara con los obtenidos en etapas anteriores, con el fin de establecer la solidez de los indicadores hallados.

En el último capítulo, De Sena presenta una síntesis de los principales resultados obtenidos durante el proceso de validación de los indicadores, los cuales fueron desarrollados y evaluados al concluir cada fase del proceso investigativo. El propósito de esta síntesis es invitar al lector a reflexionar sobre los hallazgos y considerar la posibilidad de revisar ciertos aspectos fundamentales de los indicadores o sus dimensiones, con el objetivo de que, en sus aplicaciones futuras, éstos ofrezcan información precisa y de mayor calidad.

Como cierre del libro, la autora plantea una reflexión sobre la dificultad que enfrentan las Ciencias Sociales para articular enfoques cualitativos y cuantitativos en la obtención de datos. En este sentido, sugiere que dicha dificultad podría ser superada mediante el uso del diseño multi-método, el análisis de datos secundarios y la inclusión de los sujetos bajo estudio en el proceso investigativo. Además, destaca la importancia de evaluar qué supuestos teóricos se ajustan mejor a la realidad que se desea estudiar, con el fin de establecer un vínculo necesario entre la teoría y la realidad observada.

En conclusión, este libro nos invita a reflexionar sobre la importancia de comprender que, en las Ciencias Sociales, los sujetos bajo estudio no pueden analizarse de manera aislada, sino en relación con sus contextos familiares, sociales, económicos y culturales. Desde esta perspectiva, se enfatiza la necesidad de superar los marcos teóricos y metodológicos convencionales, evitando una adhesión acrítica a los supuestos disciplinares. En lugar de ello, se propone una aproximación que favorezca la creatividad sociológica y la constante interrogación del investigador sobre sus propias herramientas analíticas. Como señalan De Sena y Cervio (2023), producir conocimiento científico sobre el mundo social implica formular preguntas y explorar caminos que obligan al investigador a involucrarse. Este planteo evidencia que ningún método por sí solo resulta suficiente para abordar la complejidad de los fenómenos sociales, lo que exige una apertura hacia nuevas estrategias de indagación. Así, la obra sugiere la exploración de enfoques innovadores, la construcción de marcos teórico-metodológicos propios y la reelaboración de procedimientos existentes, en un esfuerzo continuo por lograr una articulación sólida entre teoría, metodología y epistemología. La investigación social, en este sentido, se concibe como un ejercicio dinámico que no solo analiza la realidad, sino que también se transforma y la transforma en el proceso, desafiando al investigador a construir herramientas metodológicas más flexibles y situadas, capaces de captar la complejidad del mundo social.

Bibliografía

DE SENA, A. (2018). *Políticas sociales y microemprendimientos socio-productivos: Indicadores de sustentabilidad*. Buenos Aires: Ciccus.

DE SENA, A. y CERVIO, A.L. (2023). Presentación: Revisando algunas huellas del quehacer metodológico. *Revista Latinoamericana de Metodología de la Investigación Social*, 13 (26), 4-8.

Autora.

Adriana Elisabet Romiti

Universidad Nacional de la Matanza (UNLaM), Argentina.

Estudiante de Licenciatura en Economía (UNLaM).

E-mail: aeromiti75@gmail.com

Citado.

ROMITI, Adriana Elisabet (2025). Repensando la Metodología en Ciencias Sociales: desafíos de la creación y validación de indicadores. *Revista Latinoamericana de Metodología de la Investigación Social - ReLMIS*, N°29, Año 15, pp. 100-103.

Plazos.

Recibido: 28/12/2024. Aceptado: 10/03/2025.