CIENCIA DE DATOS: Algunos conceptos introductorios

DATA SCIENCE: Some introductory concepts

Fabricio O. Penna

fabricio.penna@gmail.com

Especialista en Investigación en Ciencias Sociales y Humanas. Profesor Asociado de Investigación Educativa I – Lic. y Prof. en Cs. De la Educación – Facultad de Ciencias Humanas e Investigador (PROICO 12-0220 "Estrés y salud: Una Propuesta Integradora para la Promoción, Prevención e Intervención de Problemas Psicosociales, Orientado a un Desarrollo Saludable de la Persona y la Comunidad", FaPsi, UNSL y PROICO 12-0718 "La investigación en psicología y su incidencia en la formación del psicólogo", FaPsi, UNSL).

O. Hernán Cobos

ohcobos@gmail.com

Especialista en Metodología de la Investigación Científica. Profesor Adjunto de Investigación Educativa I – Lic. y Prof. en Cs. de la Educación – Facultad de Ciencias Humanas e Investigador (PROICO 12-0220 "Estrés y salud: Una Propuesta Integradora para la Promoción, Prevención e Intervención de Problemas Psicosociales, Orientado a un Desarrollo Saludable de la Persona y la Comunidad" – FaPsi, UNSL).

Sebastián M. Vázquez Ferrero

lasagnad@gmail.com

Universidad Nacional de San Luis – Argentina. Dr. en Psicología. Jefe de Trabajos Prácticos de Investigación Educativa I – Lic. y Prof. en Cs. de la Educación – Facultad de Ciencias Humanas e Investigador (PROICO 12-0718 "La investigación en psicología y su incidencia en la formación del psicólogo", FaPsi, UNSL).

Adriano Penna

apenna99@gmail.com

Estudiante de la carrera Licenciatura en Análisis y Gestión de Datos. Facultad de Ciencias Físico Matemáticas y Naturales – Facultad de Ciencias Económicas, Jurídicas y Sociales.

Resumen

Desde tiempos inmemoriales, particularmente en lo referente a la historia del conocimiento, se ha dado la dualidad entre aquello que vemos y aquello que no vemos; entre la diversidad de lo visible y la necesidad de la unidad formal. A partir de Tales de Mileto, cuando él se planteaba lo que había detrás de todo suceso y lo que había detrás de la diversidad de formas de presentación de ciertos hechos -que, básicamente, buscaba un concepto indeterminado que unificara la infinidad de enumeraciones determinadas que visualizaba-, está presente dicha dualidad.

Por ende, la ciencia de datos -que trataremos en el presente artículo, siguiendo a Sosa Escudero (2022, p.162), como sinonimia de Estadística- es la herramienta crucial, desde una visión pospositivista, en la investigación científica y en la toma de decisiones, desde la medicina hasta la economía, y, su filosofía, busca comprender cómo funciona y qué significado tiene en el contexto del conocimiento humano. El presente artículo pretende dar una

introducción a ciertos conceptos al respecto de esta heredera de la estadística.

Palabras clave: ciencia de datos, pospositivismo, investigación científica, toma de decisiones

Abstract

Since time immemorial, particularly in the history of knowledge, there has been a duality between what we see and what we do not see; between the diversity of the visible and the need for formal unity. Starting with Thales of Miletus, when he wondered what was behind every event and what was behind the diversity of forms of presentation of certain facts - basically, he sought an indeterminate concept that would unify the infinity of determined enumerations that he visualized - this duality is present.

Therefore, data science - which we will address in this article, following Sosa Escudero (2022, p. 162), as a synonym for Statistics - is the crucial tool, from a post-positivist perspective, in scientific research and decision-making, from medicine to economics, and its philosophy seeks to understand how it works and what its meaning is in the context of human knowledge. This article aims to provide an introduction to certain concepts regarding this heir to statistics.

Keywords: Data science; Post-positivism; Scientific research; Decision making

Introducción

La ciencia de datos, en relación a sus orígenes, se confunde -casi- con la historia de la humanidad misma. Dicha ciencia, según lo expresado por Rao (1989, citado en Gutiérrez Cabría, 1994), "(…) posee gran antigüedad, pero escasa historia." (p.21). Así, si nos remitimos a tiempos pretéritos, pensando en civilizaciones como el antiguo Egipto, China, Grecia o el Imperio Romano

(sólo por mencionar algunas de ellas), ya realizaban censos, no sólo para determinar el número de habitantes y sus posesiones sino (y como era de esperar) para el cobro de impuestos. En ese sentido, las estadísticas son anteriores a los Estados nacionales como los conocemos y concebimos a partir de la revolución francesa.

A su vez, a su heredera, la ciencia de datos, podemos definirla como un método particular de abordar 'fenómenos colectivos', entendiendo por colectivos a aquellos hechos susceptibles de variar sin una regla asignable determinada. Otra manera de definirla es, también, como un 'fenómeno de masas'; en otras palabras, características que demuestran una regularidad sólo para grandes masas de observaciones/datos. En ese sentido, cabe apoyarse sobre la posición de Tukey (1962, p.6), quien acuñó el mismo término de ciencia de datos, planteándola como una evolución de la estadística. Más allá de ser una disquisición filosófico-epistemológica relevante, cabe plantear entonces que la ciencia de datos es al mismo tiempo una disciplina válida en sí misma y una herramienta muy relevante para la toma de decisiones, que permite elaborar conclusiones rigurosas y claramente debatibles a partir de la información obtenida. Estas conclusiones reproducibles y falibles facilitan la discusión pública. Por otro lado, cabe destacar que la ciencia de datos no está exenta, como toda producción humana, de una historia y un desarrollo propios, que marcan sesgos y perspectivas, abordados a continuación.

Desarollo

Una de las primeras y más elementales tareas de la ciencia de datos se refería a la observación objetiva de fenómenos, que debía llevarse a cabo mediante encuestas a todas las unidades de análisis y, los datos obtenidos, eran registrados y catalogados de forma sistemática. Al mismo tiempo, los y las científicos/as de datos han tratado de superar los clichés, sesgos y prejuicios resultantes de la recopilación no sistemática de información.

Al respecto, expresa un antiguo refrán español: "año bisiesto, año siniestro". ¿Y cómo se verificaría, desde la ciencia de datos, la verdad de este refrán? Una forma, por ejemplo, podría ser centrar la atención en un cultivo agrícola, establecer un criterio para poder definirlo como bueno o malo y proceder a la doble clasificación de un mismo cultivo, según el año bisiesto y sus bondades. Desde un punto de vista determinista, la cosecha debería ser mala cuando y sólo cuando sea año bisiesto. Sin embargo, podría haber una tendencia, más o menos pronunciada, a que las cosechas sean malas en los años bisiestos. Pero ¿hasta qué punto esta tendencia debería considerarse estable en el tiempo?

Este es otro aspecto de la ciencia de datos: el de estar preparada para la variabilidad de los resultados, superando el determinismo inherente a nuestra concepción tradicional de la naturaleza y la arraigada concepción causal que forma parte de la estructura cultural de nuestra civilización.

A lo expresado anteriormente, podemos agregar que uno de los debates fundamentales en la 'filosofía' de la ciencia de datos se relaciona con la objetividad y la subjetividad en la toma de decisiones. Por un lado, se arguye que proporciona un enfoque objetivo para comprender el mundo, ya que se basa en datos empíricos. Pero, por el otro, es dable reconocer que la elección de qué datos recopilar, cuáles métodos aplicar y cómo interpretar los resultados obtenidos, que bien pueden estar influenciados por sesgos personales y, a veces, por prejuicios, pueden resultar subjetivos. Dicho de otro modo, la filosofía se preocupa por analizar cómo minimizar los sesgos y garantizar tomas de decisiones más objetiva. Es decir, se preocupa por identificar falacias y/o errores comunes en el uso de la información que nos suministran los datos. Esto incluye entender la importancia de conceptos tales como: ¿cuáles son los tamaños de muestra adecuados?, ¿qué prueba de hipótesis utilizar acorde al tipo de variable y al nivel de medición?, ¿bajo qué supuestos aplicar un modelo de regresión y cuándo un modelo de correlación?, entre otros.

RevID. Revista de Investigación y Disciplinas. Número 11, 2024.
revid.unsl.edu.ar
ISSN: 2683-9040

Por otra parte, en los últimos años, ha habido una progresiva preocupación en la comunidad científica sobre la replicabilidad[17] de los resultados de algunas investigaciones, que cuestionan la potencia y la robustez de los modelos utilizados en la investigación y cómo los resultados (y por ende, la toma de decisiones) pueden verse afectados por problemas como el p-hacking[18], o bien, en la selección de informes.

Es menester recordar que Gauss[19], creador de la 'teoría de los errores de medición', puede considerarse como una de las raíces históricas más importantes de la ciencia de datos. En particular, el mencionado autor, es responsable (afortunadamente y de manera invaluable) de la 'curva de error accidental' además del 'método de mínimos cuadrados', que ha demostrado ser fundamental en el estudio de las relaciones entre dos o más variables.

Por su parte Neyman (1957), hace más de medio siglo, describe a la ciencia de datos diciendo "(…) que el propósito de toda investigación seria es proporcionar las bases para la selección de uno de varios cursos de acción contemplados (…)" (p.16), agregando además que "(…) el reconocimiento de que la conveniencia de tal o cual curso de acción depende de las circunstancias y, por supuesto, de las preferencias y creencias subjetivas del concepto individual." (p.16). Lo atrayente -por no decir sublime- de esta cita está en que la teoría de la contrastación debiera considerarse como una prudente toma de decisión, según que una determinada hipótesis sea, o no, rechazada.

[17] Término acuñado por Karl Raimund Popper (1902-1994) en su libro "Conjeturas y Refutaciones" donde, de manera muy sucinta, destaca que el único requisito para que una investigación sea científica es que sea *replicable*, es decir, que el estudio pueda repetirse.

[18] Expresión utilizada cuando los datos son manipulados -de manera consciente o no- para llegar a obtener resultados que favorezcan a quienes investigan

[19] Carl Friedrich Gauss (1777-1855). Matemático, Astrónomo y Físico alemán que contribuyó, entre otros temas, a la teoría de números, al análisis matemático y a la estadística, concibiendo, además, la llamada distribución normal, distribución de Laplace-Gauss o, simplemente, gaussiana.

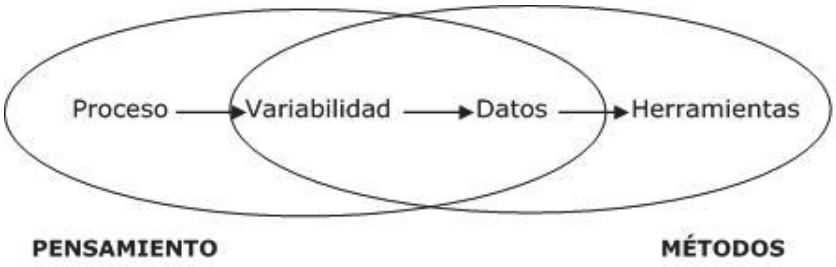
De acuerdo a lo mencionado por Snee (1993), "El interés personal en el tema en el que se aplica la metodología estadística y la participación personal en los procesos de recopilación de datos y resolución de problemas son esenciales para desarrollar el valor del pensamiento estadístico." (p.151). Cabe señalar, entonces, que la comunicación entre los y las científicos/as de datos e investigadores/as de distintas disciplinas, dependerá de los problemas básicos que le sean formulados a la ciencia de datos por las diversas ramas de la ciencia.

A lo que Gutiérrez Cabría (1994) añade que la utilización -y validez actual- de la ciencia de datos (obviamente, con sus modelos incluidos) es, sin lugar a dudas, "(…) una de las varias interpretaciones del método científico, la preferida hoy por la mayor parte de los científicos." (p.67). Siendo, sus métodos, los más aplicados por investigadoras e investigadores que permiten, en cierta forma, disminuir cuestiones subjetivas. Luego, el autor señalado supra, finaliza diciendo que la ciencia de datos es transversal y que está "(…) debido a su carácter instrumental al servicio de las demás disciplinas." (p.67).

Snee (1999) expresa que los elementos del pensamiento de la ciencia de datos son, sin lugar a duda, el proceso, la variabilidad y la necesidad de datos. Agrega, además, que estos elementos están conectados ya que, "(…) toda actividad, sea trabajo o no, es un proceso. Un proceso se define como cualquier actividad que convierte insumos en productos. Los problemas de la investigación empírica están asociados con uno o más procesos." (p.256). Por ende, sean cuales fueren, el o los procesos involucrados, nos dan los argumentos para el trabajo que lleva a cabo la ciencia de datos.

Lo expresado anteriormente, queda plasmado en la siguiente figura:

Figura 1: Relación entre el pensamiento y los métodos de la ciencia de datos.



Fuente: Snee (1999, p.256)

A todo esto, y más recientemente, según lo expresado por López Lozada (2004) que la ciencia de datos "(…) es la forma en que la información se ve, se procesa y se convierte en pasos de acción. Es una filosofía de pensamiento, no una forma de realizar cálculos matemáticos." (p.2), agregando, además, que:

El pensamiento estadístico utiliza el concepto de que toda actividad consiste en un conjunto de pasos interconectados que deben complementarse y completarse para lograr una meta planteada, donde se debe investigar cada paso para identificar áreas de oportunidad y mejora a fin de lograr el éxito personal o profesional. (p.2)

Ya, adentrándonos en los últimos lustros, hemos notado con gran estupor que hay investigadores e investigadoras que consideran que, con el auge del big data, la ciencia de datos está en plena decadencia y cometen un gran error conceptual. Creemos que esto puede deberse, tal vez, a que no llegan a entender el concepto de big data.

Pues, cuan hidalgos caballeros calzados con yelmo y armadura, salgamos en defensa de ella: la Ciencia de Datos…

Según Leonelli (2018) "Big Data son datos de diferentes tipos y orígenes que están conectados entre sí, a menudo en forma digital y de maneras que se

prestan al aprendizaje automático[20], para producir nuevas formas de análisis y conocimiento." (p.23), y luego, la autora mencionada, agrega que "Como lo discutieron dos destacados sociólogos de datos, Boyd y Crawford, la expresión Big Data hace referencia a "la capacidad de explorar, agregar y relacionar grandes conjuntos de datos"." (p.23).

Por su parte Sosa Escudero (2022) afirma que -y de ello estamos en un todo de acuerdo-:

Desde cierto punto de vista, el combo big data-machine learning es una "versión 2.0" de la estadística, tiene que ver con intentar aprender con datos, como lo venimos haciendo desde que tenemos memoria histórica. Pero desde otra perspectiva, el fenómeno refiere a una forma evolutiva de convivir con los datos, y es en este último sentido que la "ciencia de datos" tal vez exista más allá de la estadística, pero sin ser ajena a su herencia. (p.162)

Agregando, el autor mencionado supra, que si nos detenemos a pensar en big data, "(…) es sólo un fenómeno de "más datos" (…)" (p.174) y, por el cual, la ciencia de datos debería estar sumamente feliz y satisfecha ya que se deshace (¡muchas veces!) de ese aciago problema que representa "(…) liberarse de la escasez de materia prima con la cual trabajar." (p.174) y, dicho autor manifiesta luego que, más cantidad de datos es sólo un sinónimo de "(…) una abundancia de datos de naturaleza muy diferente." (p.174). O bien como indica Ojeda Ramírez (2017), que la ciencia de datos "(…) continúa su desarrollo, ahora más vertiginoso que nunca, en la era del Big Data." (p.35).

Es claro que, por otro lado, el big data no da las soluciones mágicas esperadas por la ciencia de datos (y, por ende, a investigadoras e investigadores que trabajan con ella) si no que, más bien, aumenta los problemas. Siendo, por ejemplo, un problema interesante para la ciencia de datos (sólo por

[20] Conocido, en la lengua sajona, como "machine learning".

mencionar uno) el planteado por Vasconcelos, Ramos y Coutinho (2023) donde aseguran que "Los desafíos involucran problemas típicos de las aplicaciones de big data dentro del alcance del procesamiento analítico de datos urbanos masivos y heterogéneos, como tratar con fuentes heterogéneas, procesar datos espaciales y proporcionar un almacenamiento eficiente." (p.51). Pero, de todas maneras, la ciencia de datos (Sosa Escudero, 2022) "(…) no ha perdido ni un centímetro de su relevancia, tal vez todo lo contrario." (p.175).

Para finalizar, Sosa Escudero (2022) asegura, además, que la ciencia de datos "(…) y la ciencia formal y empírica serán una componente clave para que el big data se transforme en una auténtica revolución." (p.177).

Como siempre, y no está de más mencionarlo, es factible pensar que detrás de los datos recopilados por big data estén presenten determinados sesgos que podrían impactar en el aprendizaje automático y que los errores, al momento de la toma de decisión, crezcan de forma -casi- exponencial.

Conclusión

De acuerdo a lo presentado en el presente artículo, la ciencia de datos está representada por situaciones de incertidumbre, tanto en la enseñanza (tema que trataremos, con mayor profundidad, en un próximo artículo) brillantemente definida por Gutiérrez Cabría (1994) cuando afirma que "Esta tendencia a la matematización de la estadística condujo a que en la enseñanza fueran descuidados los aspectos prácticos del análisis de datos." (p.299); como en la investigación empírica, siendo sustituida a veces por la variabilidad o más bien por la variación individual de un modelo o un esquema interpretativo de realidad. Es nuestra opinión que el énfasis debería darse en el método de investigación, recuperando todas sus fases: desde la definición de los hechos, el estudio de las técnicas de recolección de datos, su análisis,

la dócima para uno o varios parámetros poblacionales, hasta llegar a la toma de decisión.

Por otra parte, y conforme a lo dicho por Scardovi (1980), "(…) el descubrimiento de la variabilidad natural, de su papel y de su génesis es la clave metodológica del pensamiento científico que vino con la crisis de la Weltanschauung[21] determinista (…)" (p.5) y agrega, haciendo alusión tanto a la experimentación como a la ciencia de datos, que más que distinguir diferentes metodologías, son utilizadas para indicar fases y métodos de investigación, criterios y técnicas para contrastar hipótesis.

Llegado a este punto tenemos que preguntarnos: ¿cuál es el objeto de estudio de la ciencia de datos? ¿una situación de incertidumbre o de variabilidad? La respuesta más interesante cabría pensarse como que no es ni lo uno, ni lo otro. Consideramos preferible volver a situar los fenómenos colectivos (no deterministas) como base de toda investigación.

Debido a ello, podemos pensar a la ciencia de datos como un 'método de estudio de los fenómenos examinados desde un punto de vista no determinista'. Alternativamente, y de acuerdo a lo expresado por Scardovi (1980), se la puede concebir como "(…) el método del conocimiento científico (...)" (p.9).

Por último, y para finalizar, debiéramos plantearnos: ¿cómo se documenta la realidad a investigar?, ¿qué datos hemos tomado y por qué los estamos analizando?, ¿cómo se actualizan la información y los datos disponibles?, ¿cómo se construyen los modelos teóricos? Éstos son, sólo, algunos de los interrogantes que la ciencia de datos deberá, de manera perentoria, responder.

[21] Del alemán: cosmovisión.

Referencias bibliográficas

Gutiérrez Cabría, S. (1994). Filosofía de la estadística. Servei de Publicacions Universitat de València.

Leonelli, S. (2018). La ricerca scientifica nell'era dei Big Data: Cinque modi in cui i Big Data danneggiano la scienza, e come salvarla. Meltemi editore: Milano.

López Lozada, L. (2004). "Pensamiento estadístico: directivos con nuevas tecnologías de información y comunicación". Espacios, Vol. 25 (3).

Neyman, J. (1957). "'Inductive behavior' as a basic concept of philosophy of science". Rev. Inst. Internat. Statist. 25: 1/3, 7-22. https://doi.org/10.2307/1401671

Ojeda Ramírez, M.M. (2017). "La estadística en la Era del Big Data". Selección Gaceta Politécnica. Volumen 9, 31-35.

Scardovi, I. (1980). Appunti di statistica. Pàtron editore. Bologna.

Snee, R.D. (1993); "What's Missing in Statistical Education". The American Statistician, 47 (2), 149-154.

Snee, R.D. (1999); "Discussion: Development and use of statistical thinking: A new era". International Statistical Review, 67 (3), 255-258.

Sosa Escudero, W. (2022). ¿Qué es (y que no es) la estadística? 2ª ed. ampliada. Siglo XXI Editores: CABA.

Tukey, J.W. (1962). "The Future of Data Analysis". The Annals of Mathematical Statistics. 33(1), 1-67. ISSN 0003-4851. Doi:10.1214/aoms/1177704711.

Vasconcelos, F.F.; Ramos, V.T. y Coutinho, F.J. (2023). "Os desafios e soluções para a implementação de Big Data Analytics em cidades inteligentes". Companion Proceedings of the 38th Brazilian Symposium on Databases, pp.50-56.

Recibido: 05/08/2024
Aceptado: 27/11/2024

Cómo citar este artículo:
Pena, F. O, Cobos, H.O, Vázquez Ferrero, S.M y Pena, A. (2024).
CIENCIA DE DATOS: Algunos conceptos introductorios. RevID, Revista de Investigación y Disciplinas, Número 11, San Luis, p 151-163