PAIDEIA XXI

Vol. 12, N° 1, Lima, enero-junio 2022, pp. 161-167
ISSN Versión Impresa: 2221-7770; ISSN Versión Electrónica: 2519-5700

http://revistas.urp.edu.pe/index.php/Paideia

COMMENTARY / COMENTARIO

TRAINING DECISION AS AN ELEMENT OF SCIENTIFIC RESEARCH FROM THE CONCEPTUALIZATION OF STATISTICS AND DATA SCIENCE: THE OBVIOUS, NOT SO OBVIOUS

DECISIÓN FORMATIVA COMO ELEMENTO DE LA INVESTIGACIÓN CIENTÍFICA DESDE LA CONCEPTUALIZACIÓN ESTADÍSTICA Y CIENCIA DE DATOS: LO OBVIO, NO TAN OBVIO

George Argota-Pérez[1]*; Yadira Argota-Pérez[2]; Rina María Álvarez-Becerra[3] & María Gilda Reyes-Diaz[4]

[1] Centro de Investigaciones Avanzadas y Formación Superior en Educación, Salud y Medio Ambiente ¨AMTAWI¨. Perú. george.argota@gmail.com
[3] Casa Consultora DISAIC. La Habana, Cuba. solyap87@gmail.com
[3] Facultad de Ciencias de la Salud. Universidad Nacional Jorge Basadre Grohmann (UNJBG). Tacna, Perú. rinaalvarezb@gmail.com
[4] Facultad de Farmacia y Bioquímica. Universidad Nacional "San Luis Gonzaga" (UNICA). Ica, Perú. maria.reyes@unica.edu.pe

* Corresponding author: george.argota@gmail.com

George Argota-Pérez: https://orcid.org/0000-0003-2560-6749
Yadira Argota-Pérez: https://orcid.org/0000-0002-0880-4394
Rina María Álvarez-Becerra: https://orcid.org/0000-0002-5455-6632
María Gilda Reyes-Diaz: https://orcid.org/0000-0002-6607-9247

ABSTRACT

The purpose of the study was to describe the need for decision-making from the conceptualized training between Statistics and Data Science. Four elements are key in science: theory, data, methodology, and problem, because if the data is part of science then it seems wrong that there is a DataScience since no methodology from Data Science can decide, the "ideal or correct" pattern since

doi:10.31381/paideia.v12i1.4841

there are multiple patterns to be understood. On the other hand, if the statistical programs are incapable of analyzing hundreds of thousands of data (it makes no sense when decisions are recognized from a random probabilistic sample and, on the contrary, not considered makes it impossible to make inferences), then the possibility of representing diversity from Statistics is limited since there is centralization in minimizing the sums of the deviations to the mean square and not understanding the diversity that Data Science performs. It is concluded that Statistics adds to reliability and validity, while Data Science allows the development of methodologies that condition the incorporation of technologies where it is difficult to unmark the barrier between Statistics and Data Science because on some occasions they are indistinct from each other and in other cases an association is shared. Therefore, mastery of data processing from Statistics and machine learning facilitated by Data Science is required for the decision, but there must be training in both fields of study.

Keywords: data science – decisions – professional competence – statistics

RESUMEN

El propósito del estudio fue describir la necesidad en la toma de decisiones desde la formación conceptualizada entre la Estadística y Ciencia de Datos. Cuatro elementos son claves en la ciencia: teoría, datos, metodología y problema, por cuanto, si los datos forman parte de la ciencia entonces, parece erróneo que exista una Ciencia de Datos, pues ninguna metodología desde la Ciencia de Datos puede decidir, el patrón "ideal o correcto" dado que existen múltiples patrones a comprenderse. Por su parte, si los programas estadísticos son incapaces de analizar cientos de miles de datos (carece de sentido al reconocerse las decisiones desde una muestra probabilística aleatoria y, por el contrario, no considerarse imposibilita hacer inferencias), entonces la posibilidad de representar la diversidad desde la Estadística es limitada, ya que existe una centralización en minimizar las sumas de las desviaciones al cuadrado medio y no comprender la diversidad que realiza la Ciencia de Datos. Se concluye, que la Estadística suma a la confiabilidad y validez, mientras que la Ciencia de Datos permite el desarrollo de metodologías que condicionan a la incorporación de tecnologías donde resulta difícil desmarcar la barrera entre la Estadística y la Ciencia de Datos, pues en algunas ocasiones son indistintas entre sí y en otros casos se comparte una asociación. Por tanto, el dominio del tratamiento de los datos desde la Estadística y el aprendizaje automático que facilita la Ciencia de Datos se requiere para la decisión, pero debe existir la formación en ambos campos de estudio.

Palabras clave: ciencia de datos – competencia profesional – decisiones – estadística

INTRODUCCIÓN

La Estadística se ocupa del tratamiento de datos y la inferencia poblacional mediante la observación de hechos. Es decir, define los resultados esperados, la población, unidades de observación, variables, métodos, plan de muestreo, tamaño de la muestra y factores, métodos estadísticos y el diseño de experimentos entre otros aspectos (Villarroel, 2002; Pérez, 2015; Villegas, 2019). Actualmente, la Estadística se aplica en todas las áreas del saber y entre ellas destaca la Ciencia de Datos, la cual influye en el desarrollo industrial y tecnológico (McNutt, 2014; Nachtsheim & Stufken, 2019; MacGillivray, 2021; Sardareh *et al.*, 2021). Sin embargo, en los últimos tiempos existe la discusión que se puede hacer ciencia y tomar decisiones hábiles desde la Ciencia de Datos (Diggle, 2015; Galeano & Pena, 2019), sin considerar una teoría estadística (Granville, 2014; Davison, 2018).

Según, Learner & Phillips (1993), existen cuatro componentes que destacan en la ciencia: 1ro) la teoría, 2do) los datos, 3ro) la metodología, y 4to) el problema. Pareciera erróneo, que exista una Ciencia de Datos, pues los datos forman parte de la ciencia y los mismos se procesan, a partir de sus registros (Phillips, 2017). La relevancia de la ciencia se comprende desde la significación de una zona de aceptación o rechazo (Figura 1) con margen de probabilidad, según una prueba de hipótesis: Ho (Fisher, 1992).

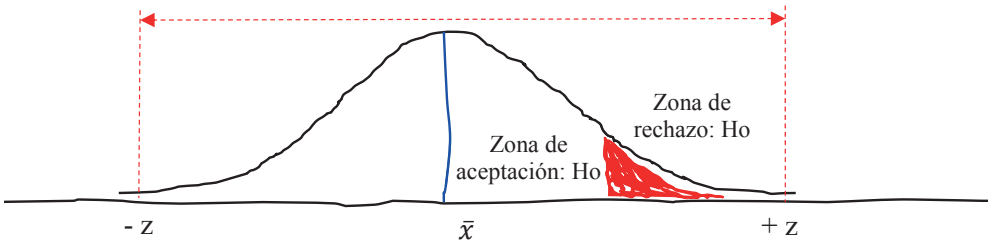

Figura 1. Significación estadística en la ciencia / zona de aceptación o rechazo.

Al mismo tiempo, se desea suponer que una institución universitaria integró, dos profesionales calificados a sus planes de investigación científica y además adquirió, una tecnología de última generación para generar y analizar múltiples datos lo cual se esperaría, un crecimiento por tres años de sus productos y/o servicios, al menos para dos clientes externos: A y B (Figura 2), así como la continuación del compromiso práctico que se tiene en la propia formación y competencia de sus estudiantes de pregrado.

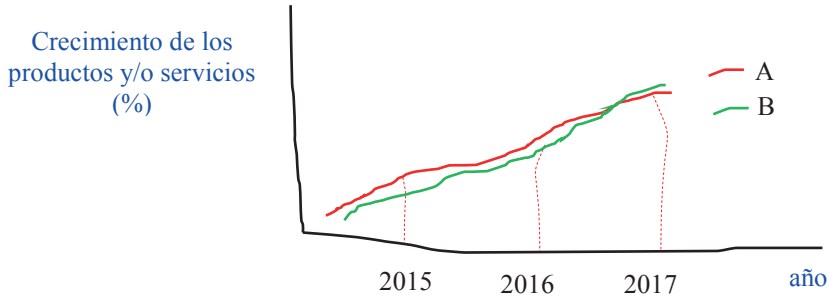

Figura 2. Aumento de productos y/o servicios. A = cliente externo / B = cliente externo.

Durante este periodo se realizó, un análisis de los resultados referidos a los productos y/o servicios y no hubo diferencias estadísticamente significativas (ej.: p≤0,05) entre A y B, aunque la tasa de rentabilidad fue mayor en A. De igual manera, la tecnología se usó por los estudiantes de pregrado, pues fue una condición de garantía, no solo para su formación, sino en la generación de datos, por parte de A y B. Sin embargo, al finalizar el año 2017, la curva empezó a descender (Figura 3), y existió preocupación, pues los datos disponibles, "no mostraron la predicción de tal efecto" y esta respuesta se evidenció con el crecimiento de productos y/o servicios durante los tres años previos.

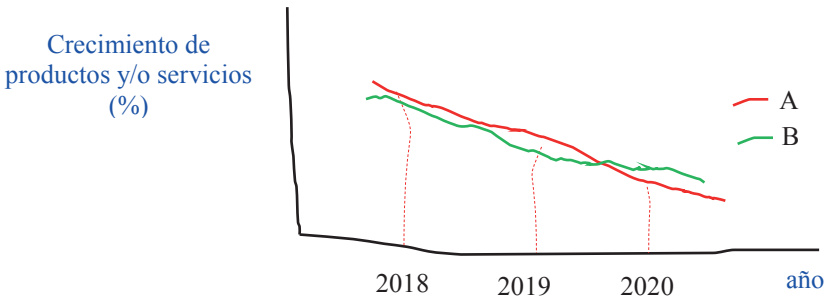

Figura 3. Disminución de productos y/o servicios A = cliente externo / B = cliente externo.

Según, la observación entre las Figuras 1 y 2, se derivan múltiples interrogantes, a partir de no visualizarse el cambio de oportunidad para mantener o mejorar los productos y/o servicios desde la argumentación estadística que se realizó. Una cuestión resulta la diferencia entre distinguir los grandes datos orientados a la investigación científica y otra es, que se orienten a la metodología de la Ciencia de Datos. Por consiguiente, la valoración contable sobre cualquier escenario futuro es circunstancial y pudiera

ser, una ficción la suposición ante cualquier valor total a esperarse. En este sentido, destaca para el análisis las preguntas siguientes:

1. ¿Si, la hipótesis indica una significación estadística y requiere de datos, qué representó la intensión humana desde la integración de los dos profesionales y la adquisión de una nueva tecnología?
2. ¿Solo la selección primaria de una agrupación de datos que se supuso para el crecimiento de productos y/o servicios permite el análisis del vacío del conocimiento desde la observación empírica a contrastar?
3. ¿El conjunto de datos que se tenía y que relacionaba el crecimiento de productos y/o servicios se probaron de forma independiente y dependiente?
4. ¿Se realizó desde la base de datos la visualización del análisis y la interacción de nuevos datos como lo propuso Tukey (1962)?
5. ¿Qué variable fue determinante a considerar y resultó ajena a la observación empírica, si los programas estadísticos son incapaces de analizar cientos de miles o millones de datos y esto carece de sentido al reconocer las decisiones desde una muestra probabilística aleatoria y, por el contrario, si no se considera tampoco pudiera hacerse inferencias a la población (Tsao *et al.*, 2016)?
6. ¿Cuál sería el significado desde la representación de la diversidad, pues la Estadística no reconoce el significado de las desviaciones debido a su centralización en minimizar las sumas de las desviaciones al cuadrado medio y no, a comprender la diversidad que realiza la Ciencia de Datos?
7. ¿Por qué no se comprobó una significación del potencial de valores atípicos y las probabilidades desde un análisis de la Ciencia de Datos (Phillip, 2017)?

Con razón al resultado de las figuras y posibles respuestas a las interrogantes, puede entenderse que el análisis de los datos hacia su inferencia o no, requiere la combinación de tareas sin segmentarse que la interpretación obedezca a la Estadística o Ciencia de Datos. Es decir, se necesita la formación adecuada para que los resultados sean satisfactorios (Sardareh *et al.*, 2021). Dado que, en diversas ocasiones los procesos metodológicos fallan (en la Estadística y la Ciencia de Datos), entonces resulta una fuga de datos y, en consecuencia, los resultados son erróneos. Se considera que la Ciencia de Datos, permite en algunos casos el desarrollo de metodologías para el análisis del cúmulo de datos (ej.: Lenguajes Python y R, técnicas de visualización y la inteligencia artificial), pero ninguna de las metodologías puede decidir, el patrón "ideal o correcto", dado que existen múltiples patrones a comprenderse. Ante, los múltiples errores que se presentan en

los entrenamientos o las pruebas de modelos, bien sea en la Estadística o Ciencia de Datos, es imprescindible que la correcta toma de decisiones sea mediante la participación formativa (Vandeput, 2020).

Es por ello, que existe la necesidad para la Estadística, incorporar la Ciencia de Datos lo que constituye ser más aplicable y accesible a una realidad global, y al mismo tiempo, que la Ciencia de Datos considere la significación desde los genuinos de la Estadística porque la precisión e interpretación de datos siempre será crucial en el aprendizaje humano para otorgar significado al mundo exterior, aunque muchos de los datos con aplicaciones a la ciencia son complejos y carecen de interpretaciones, a pesar que sean precisos (Makridakis *et al.*, 2020).

Se concluye, que la Estadística suma a la confiabilidad y validez (Carmichael & Marron, 2018), mientras que la Ciencia de Datos permite el desarrollo de metodologías que condicionan a la incorporación de tecnologías (Diggle, 2015), y resulta difícil desmarcar, la barrera entre la Estadística y la Ciencia de Datos, pues en algunas ocasiones son indistintas entre sí, pero en otros casos se comparte una asociación. Por tanto, se requiere el dominio del tratamiento de los datos desde la Estadística con el aprendizaje automático que facilita la Ciencia de Datos como un elemento fundamental de decisión formativa para la investigación científica.

REFERENCIAS BIBLOGRÁFICAS

Carmichael, I. & Marron, J.S. 2018. Data science vs. statistics: two cultures?. Japanese Journal of Statistics and Data Science, 1: 117–138.

Davison, J. 2018. *No, Machine learning is not just glorified statistics*. https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistic

Diggle, P.J. 2015. Statistics: a data science for the 21st century. Journal of the Royal Statistical Society, 178: 793-813.

Fisher, R.A. 1992. *Statistical methods for research workers*. In: Kotz, S. & Johnson, N.L. (eds) *Breakthroughs in statistics. Springer series in statistics.* (Perspectives in Statistics). Springer, pp. 66-70.

Galeano, P. & Pena, D. 2019. Data science, big data and statistics. Test, 28: 289-329.

Granville, V. 2014. *Data science without statistics is possible, even desirable*. Data Science Central. https://www.datasciencecentral.com/profiles/blogs/data-science-without-statistics-is-possible-even-desirable

Learner, D.B. & Phillips, F.Y. 1993. Method and progress in management science. Socio-Economic Planning Sciences, 27: 9-24.

MacGillivray, H. 2021. Statistics and data science must speak together. Teaching Statistics, 43: 5-10.

Makridakis, S.; Spiliotis, E. & Assimakopoulos, V. 2020. The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36: 54-74.

McNutt, M. 2014. Raising the Bar. Science, 345: 1- 9.

Nachtsheim, A.C. & Stufken, J. 2019. Comments on: Data science, big data and statistics. Test, 28: 345-348.

Pérez, D.R. 2015. Importancia de la estadística en los trabajos enviados para publicación. Revista Obstetricia Ginecología Venezolana, 75: 145-146.

Phillips, F. 2017. A perspective on 'Big Data'. Science and Public Policy, 44: 730-737.

Sardareh, S.A.; Brown, G.T.L. & Denny, P. 2021. Comparing four contemporary statistical software tools for introductory data science and statistics in the social sciences. Teaching Statistics, 43: 157-172.

Tsao, C.C.; Chang, P.C.; Fan, C.Y.; Chang, S.H. & Phillips, F. 2016. A patent quality classification model based on an artificial immune system. Soft Computing, 45: 1-10.

Tukey, J. 1962. The future of data analysis. Annals of Mathematical Statistics, 33: 1-67.

Vandeput, N. 2020. *Data science for supply chain forecasting*. De Gruyter, 2nd edition, pp. 4.

Villarroel, P.L.A. 2002. Rol de la estadística aplicada en investigación científica. Acta Nova, 2: 110-115.

Villegas, Z.D.A. 2019. La importancia de la estadística aplicada para la toma de decisiones en Marketing. Revista de Investigación & Negocios, 12: 29-42.

Received February 18, 2022.

Accepted April 5, 2022.