

3^a
Emisión

DATA SCIENCE

Módulo 03

Importancia de la estadística en ciencia de datos

Dr. Roberto Bárcenas Curtis



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Dirección General de Cómputo y de Tecnologías de información y Comunicación
Dirección de Docencia en TIC



Educación
Continua
1971 - 2021

Presentación

En algunos contextos, la "ciencia de los datos" se piensa como un término útil para designar métodos y conceptos originales del campo de la estadística, pero que suelen aplicarse en un espectro más amplio.

Aunque esto se ha ido transformando, también existe una oportunidad mucho mayor para volver a enfatizar la relevancia universal de la probabilidad y la estadística como métodos para la cuantificación de la incertidumbre, y explotar su potencial para el análisis de datos.

Objetivo

El participante identificará la importancia de la probabilidad y la estadística en el campo de la Ciencia de Datos.

Además, reconocerá sus aplicaciones en distintas áreas del conocimiento, conectando las ideas adquiridas con otras disciplinas.

Contenido

- Relevancia de la probabilidad y estadística en la Ciencia de Datos.
- Antecedentes
- Proceso y componentes

Ciencia de Datos

Es el estudio de los procesos involucrados en el análisis, procesamiento y modelación de información con el fin de construir vía evidencia, productos basados en datos para la toma de decisiones.

“Data Science is the discipline of making data useful”

Cassie Kozyrkov

Chief Decision Scientist - Google

Ciencia de datos

La ciencia de datos es una ***disciplina*** científica ***influenciada*** por la informática, la ciencia de la computación, las matemáticas, la investigación de operaciones y la estadística, así como las ciencias aplicadas.

Weihs, C., & Ickstadt, K. (2018).

BIG DATA (Hay muchos datos)

Los mecanismos para la recolección de datos (por ejemplo, sensores) y el desarrollo de tecnología para bases de datos, ha hecho posible que una gran cantidad de información pueda estar disponible en bases de datos, almacenes de datos y otros repositorios de información.

Por ello, existe la necesidad de convertir estos datos en conocimiento e información.

Antecedentes

Parte de los inicios de la Ciencia de Datos se encuentran en los trabajos e ideas de John Tukey.

En su trabajo *The future of data analysis* dice: “El análisis de datos, y las partes de las estadísticas que se adhieren a él, deben asumir las características de la ciencia... ya que es intrínsecamente una disciplina empírica... .

Antecedentes

John Chambers (1993) creía necesario poner más énfasis en la preparación y la presentación de datos, que en el modelado estadístico;

Leo Breiman (2001) distingue entre quines se enfocan en la predicción, antes que en la inferencia;

Cleveland (2001), incluso, sugirió el nombre “*Ciencia de datos*” para esta disciplina emergente.

Recorrido histórico

Año	Autor	Hecho	Publicación
1974	Peter Naur	En el prefacio, la ciencia de datos se definió como "la ciencia de tratar con datos, una vez que se han establecido, mientras que la relación de los datos con lo que representan se delega a otros campos y ciencias".	Concise Survey of Computer Methods. Studentlitteratur, Lund, Sweden.
1996		Se utilizó por primera vez el término Data Science en el título de una conferencia estadística .	"International Federation of Classification Societies (IFCS) "Data Science, classification, and related methods"
1997	Jeff Wu	Propone cambiar el nombre de Estadística a Ciencias de Datos y de estadístico a científico de datos.	Inaugural lecture for the H. C. Carver Chair in Statistics at the University of Michigan, Professor C. F. Jeff Wu.
2001	Cleveland	Describe un plan para ampliar las principales áreas de trabajo técnico del campo de la estadística. Debido a que el plan es ambicioso e implica un cambio sustancial, sugiere que el campo ampliado se llame "Ciencia de Datos".	Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistics Review, 69, 21-26.
2002		Inicia la publicación de "Data Science Journal" y de "Journal of Data Science"	"Data Science Journal" y de "Journal of Data Science"

Componentes ciencia de datos

$$\text{Ciencia de datos} = \frac{\text{estadística} + \text{informática} + \text{cómputo} + \text{comunicación} + \text{sociología} + \text{administración}}{\text{datos} + \text{entorno} + \text{pensamiento}}$$

En esta especificación, la sociología representa los aspectos sociales (datos+entorno+pensamiento), lo cual significa que todas esas disciplinas actúan sobre la base de los datos, el entorno y el pensamiento de datos para la creación conocimiento.

Referencias

1. Weihs, C., & Ickstadt, K. (2018). Data science: the impact of statistics. *International Journal of Data Science and Analytics*, 6(3), 189-194.
2. Diggle, P. J. (2015). Statistics: a data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(4), 793-813.
3. Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1), 1-67.
4. Wu, J. (1997) Inaugural lecture for the H. C. Carver Chair in Statistics at the University of Michigan. <http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf>
5. Chambers, J. M. (1993). Greater or lesser statistics: a choice for future research. *Statistics and Computing*, 3(4), 182-184.
6. Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199-231.
7. Cleveland, W. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistics Review*, 69, 21-26.

Contacto

Dr. Roberto Bárcenas Curtis

rbarcenas@ciencias.unam.mx