PROGRAMACIÓN COMPUTACIONAL Y ANÁLISIS DE DATOS EN EDUCACIÓN ESTADÍSTICA

PROGRAMMING COMPUTING, DATA ANALYSIS IN STATISTICAL EDUCATION

HUMBERTO CUEVAS
TECNOLÓGICO NACIONAL DE MÉXICO, MÉXICO
jesus.ca@chihuahua2.tecnm.mx
https://orcid.org/0000-0003-4142-8161

CRISTINA SOLÍS
UNIVERSIDAD TÉCNICA NACIONAL, COSTA RICA
csolis@utn.ac.cr
https://orcid.org/0000-0002-2859-6008

IXRAEL SILVA
ULSA CHIHUAHUA / TECNOLÓGICO NACIONAL DE MÉXICO, MÉXICO
ixrael.silva@itchihuahuaii.edu.mx
http://orcid.org/0000-0002-6873-3898

Fecha de recepción: 16 noviembre 2018
Fecha de aceptación: 15 enero 2019

RESUMEN

Se presentan los resultados de un estudio cuyo propósito fue conocer y comparar la opinión de un grupo de estudiantes de Ingeniería respecto de las características más representativas de los lenguajes de programación *R* y *Python* a la luz de su uso en el análisis estadístico de un caso de estudio real. A los 61 estudiantes participantes, quienes cursaban asignaturas de probabilidad y estadística; se les presentó un caso de estudio real y contextual para su examen. El 98,3% realizó el análisis y documentó sus resultados de acuerdo con los lineamientos establecidos; en el caso de *R*, el 81,97% optó por usar paquetes externos al núcleo básico para elaborar el informe reproducible, el 52,45% indicó usar *R Markdown* en detrimento de otra tecnología. En el caso de *Python*, el 88,52% usó las librerías *Scipy*, *Matplotlib*, *Numpy* y *Pandas* para el análisis; el 67,21% utilizó *Markdown-Python* para la redacción del informe. De *Python* destacaron la facilidad para escribir código; de *R* distinguieron su potencia para organizar, visualizar y efectuar cálculo estadístico. Se recomienda efectuar un estudio experimental que permita probar métodos pedagógicos que integren prácticas distintas a las predominantes durante las últimas tres décadas en la educación estadística.

PALABRAS CLAVE: educación estadística; programación computacional; análisis exploratorio de datos

ABSTRACT

We present the results of a study whose purpose was to know and compare the opinion of a group of engineering students in relation to the most representative characteristics of the R and Python

languages, especially in their use in the statistical analysis of a real and contextual study case. Sixty-one Engineering students took part in the topics of probability and statistics at the same time of the inquiry and were presented with a real study case for their exam. The 98.3% performed the analysis and documented the results according to the established guidelines; in the case of R, the 81.97% chose to use packages external to the basic core; to elaborate the reproducible report, the 52.45% indicated to use R Markdown and not another technology. In the case of Python, the 88.52% used the libraries Scipy, Matplotib, Numpy and Pandas for the analysis; the 67.21% used Markdown-Python to write the report. Python is excellent to write code; R was selected for the power to organize, visualize and perform statistical calculations. We recommended an experimental study that test a pedagogical method that integrates practices different from those that have dominated the spectrum of statistical education during the last three decades.

KEYWORDS: statistical education; programming computing; exploratory data analysis.

1. INTRODUCCIÓN

¿En qué medida es necesario utilizar lenguajes de programación computacional en la enseñanza y aprendizaje de la estadística?, ¿cuáles son las ventajas y desventajas en el uso de los lenguajes de programación R y Python en el tratamiento estadístico de datos y su representación gráfica en el campo de la ingeniería? La incorporación masiva de tecnologías de información y comunicación en el ámbito educativo hacen necesario plantearse las interrogantes anteriores, especialmente, por los posibles beneficios añadidos en la formación académica y el desarrollo de habilidades intelectuales que pueden generarse.

Actualmente, la comunidad científica internacional cuenta con una variedad extensa de software especializado para el tratamiento y análisis estadístico de datos; entre los más representativos se encuentran *SAS, SPSS, STATA, Statistica; Minitab, EViews, Statgraphics*, entre otros. Todos incluyen herramientas sofisticadas de análisis de Interfaces Gráficas de Usuario (GUI, por sus siglas en inglés) que permite su uso en entornos con necesidades diversas, como ocurre en los ámbitos científico, industrial y académico.

Este tipo de software especializado es licenciado por las empresas que los desarrollan y para su uso legal es necesario pagar derechos que ascienden desde cientos hasta decenas de miles de dólares (USD). Por consiguiente, con regularidad quedan fuera del alcance del usuario medio y, únicamente, algunas organizaciones, universidades y centros de investigación pueden sufragar los costos muy entusiastas que, regularmente, no obtienen retribución económica alguna.

La formación de profesionales en el campo de la ingeniería exige un dominio exhaustivo en disciplinas formales y fácticas, adiestramiento en el uso de métodos cuantitativos y tecnología computacional para efectuar cálculo numérico-simbólico, modelado y visualización de datos. Estas exigencias se reflejan en los planes y programas de estudio, destacando las que incorporan tópicos relacionados con la probabilidad, la estadística y el análisis de datos. Una muestra de lo anterior es que sus mallas curriculares incorporan una o más asignaturas especializadas de manera directa y, al menos, otras dos integran tópicos estocásticos de forma indirecta. Un fenómeno similar se presenta en el área de la informática, en el cual, Rüde, Willcox, McInnes & Sterck (2018) destacan que su

utilización se ha masificado en la investigación básica y aplicada que se realiza en el ámbito académico, industrial y en centros avanzados de investigación.

Respecto de la tecnología computacional sugerida en los programas de estudio, destacan los programas de cómputo diseñados exprofeso y hojas electrónicas de cálculo. Los primeros ofrecen herramientas sofisticadas de análisis, visualización, tabulación y modelado de datos; las segundas proveen opciones y rutinas de propósito general que permiten explorar información y efectuar cálculos de forma simple. Sin embargo, no es frecuente encontrar lenguajes de programación computacional como sugerencias principales; pareciera que su recomendación está condicionada al estudio de carreras profesionales específicas (v. gr. Sistemas Computacionales, Robótica, Sistemas inteligentes, entre otras) y no como herramienta para solucionar problemas o medio para desarrollar habilidades intelectuales propias del pensamiento lógico, algorítmico y científico.

Es probable que la importancia de la estadística y la computación en la formación profesional se siga incrementando, porque su combinación permite dar respuesta a interrogantes de investigación, que ni la realización de experimentos rigurosos o teorías reconocidas por su consistencia interna pueden responder. La generación masiva de datos en organizaciones pequeñas, medianas y grandes, de carácter público o privado, exigen el diseño, ejecución y monitoreo de procesos para sistematizarlos, analizarlos y crear modelos matemático-estadísticos con propósitos explicativos y predictivos para fenómenos de índole diversa.

No obstante, en el ámbito educativo es conveniente prevenir el uso abusivo de *prácticas motivadoras y software matemático-estadístico*, especialmente en el ejercicio de la docencia. En un estudio reciente, Zetterqvist (2017) cuestiona el considerar la mera existencia de datos y tecnología computacional como eje central en el desarrollo de aprendizajes, es decir, el uso de paquetes especializados o lenguajes de programación *per se* no garantiza su logro. Sin embargo, no utilizarlos puede impedir el desarrollo de habilidades fundamentales en el ejercicio futuro de una profesión.

De igual manera, el profesorado debe conocer los mecanismos que utilizan los estudiantes para aprender, así como la percepción y opinión de ellos respecto de los instrumentos para aproximarse y clasificar la información. Además, es indispensable que los docentes se percaten del movimiento del consumo de datos hacia su producción y uso en situaciones distintas a las establecidas en los programas de estudio.

La literatura relacionada con el razonamiento estadístico se ha incrementado de manera notable las últimas dos décadas; sin embargo, pareciera que los avances son magros respecto de la cantidad de estudios realizados. Aún cuando se han logrado avances significativos, las metas y retos planteados por Ben-Svi y Garfiel (2004), Gal (2004) y una década después por Watson (2013), continúan teniendo vigencia casi dos décadas después.

En un estudio que sometió a prueba un método de trabajo basado en representaciones visuales y cómputo estadístico realizado con estudiantes preuniversitarios, Van Dijke-Droogers, Drijvers & Tolboom (2017) señalan que los currículos escolares continúan imposibilitados en la aplicación de sus saberes en entornos fuera del aula. Es

importante destacar que dichos currículos y didácticas a menudo se enfocan en estudiantes con desempeño promedio, lo que puede ocasionar que los más dotados intelectualmente pierdan la oportunidad de desarrollar habilidades superiores. En relación con el análisis de datos, la detección de valores atípicos representa una de las habilidades más necesarias en la formación estadística universitaria, especialmente las últimas dos décadas. Según Zimek & Filzmoser (2018) el interés por la necesidad de perfeccionar -y modificar- los procedimientos existentes para detectar este tipo de valores entre comunidades epistémicas que cultivan la minería de datos; enfatizan la importancia del razonamiento estadístico como punto de partida. Los autores reconocen la importancia de los métodos de trabajo para promover dicho razonamiento, pero indican que ya no responden a las necesidades y retos actuales. Así, describen una propuesta de trabajo alterna -y holista- sustentada en la unión de la disciplina con la computación para coadyuvar en la interpretación de valores atípicos y promover el desarrollo de habilidades de razonamiento desde otra perspectiva. En este mismo sentido, López y Ramírez (2018) recomiendan promover el pensamiento estadístico desde otra perspectiva en el ámbito universitario y tomar en cuenta la opinión del alumnado; en un estudio exploratorio consultaron la opinión de estudiantes universitarios respecto de la identificación y clasificación de actividades orientadas al "saber hacer" y el "para qué hacer" en su proceso de formación profesional. Los autores encontraron diferencias significativas entre las opiniones clasificadas por las regiones donde cursaban sus estudios, aun cuando el modelo educativo y sus directrices correspondientes eran las mismas.

En el caso específico del Análisis Exploratorio de Datos (AED) desarrollado por Tukey (1977), se ha incrementado el interés por su utilización en diferentes escenarios. Actualmente, se reconoce la importancia del AED en la formación académica de diversas profesiones, especialmente las relacionadas con la Ingeniería. Sin embargo, al no existir directrices metodológicas para su incorporación en otras disciplinas, comúnmente se equipará con la descripción estadística de valores de corte confirmatorio.

En relación con los objetivos del AED, entre los más importantes se pueden mencionar los siguientes: [1] Identificar patrones de comportamiento en los datos; [2] Detectar datos atípicos; [3] Descubrir factores que puedan tener efecto en el caso de bajo estudio; [4] Localizar relaciones entre variables; [5] Permitir la generación de interrogantes e hipótesis susceptibles de ser probadas empíricamente; [6] Coadyuvar en la elección apropiada de métodos estadísticos y herramientas computacionales; [7] Elaborar representaciones gráficas y tabulares para comunicar los resultados producto de la exploración; [8] Plantear y evaluar supuestos en que se basarán las posibles diferencias estadísticas y [9] Ofrecer un punto de inicio para estudios de mayor envergadura de tipo observacional o experimental.

Respecto del protocolo de trabajo en un AED, también existen diversos enfoques. A continuación, se muestran una propuesta de tipo holista, a saber, la primera etapa consiste en obtener los datos, clasificarlos y almacenarlos en un medio digital para respetar la naturaleza de dichos datos. En una segunda etapa, es necesario procesaros a través del filtrado, separación o agrupamiento en función de uno o más atributos. Enseguida se recomienda examinar la existencia de valores duplicados, faltantes, registrados de manera

errónea y corregirlos para tener una base de datos *limpia.* En esta parte del proceso, es posible realizar una exploración inicial, a través del uso de representaciones gráficas y tabulares con el objetivo de detectar valores atípicos, inferir su variación, descubrir sesgos, tendencias, entre otras características y medidas. De forma paralela, el modelado estadístico es a la vez complemento y secuencia de la actividad anterior, porque implica efectuar cálculos y pruebas como medidas de tendencia central, de dispersión, regresión y correlación. Posteriormente, se sugiere elaborar un informe digital y reproducible con los resultados obtenidos en la exploración y comunicarlos.

Como se infirió líneas atrás, la estadística y la tecnología computacional tienen un vínculo bidireccional en términos de su uso. Regularmente, la segunda se utiliza como apoyo a la primera. La programación computacional está retomando el lugar en la formación académica que casi perdió en las tres últimas décadas. No obstante, los lenguajes para programar han evolucionado de manera significativa, ofreciendo variadas alternativas de elección.

Los lenguajes de programación *R* y *Python* son utilizados de manera amplia en el ámbito académico, centros de investigación y organizaciones privadas y públicas. Su desarrollo es auspiciado por instituciones universitarias, organismos colegiados, empresas privadas multinacionales y, especialmente, por usuarios entusiastas.

R es un lenguaje de programación creado en la última década del siglo XX por dos científicos neozelandeses: Ross Ihaka y Robert Gentleman (Chambers, 2008; Ihaka, 1998; Ihaka & Gentleman, 1996). El interés principal de sus creadores fue desarrollar una herramienta computacional para efectuar análisis de datos y generación de gráficas en sus cursos de estadística universitaria. Para Kabacoff (2018) y Kassamabara (2017), *R* es una plataforma idónea para el análisis de datos por su capacidad para crear cualquier tipo de representación gráfica con un esfuerzo mínimo.

Su desarrollo fue lento al inicio, sin embargo, su potencia de cálculo y capacidad gráfica atrajeron a usuarios de comunidades de aprendizaje en los campos de estocástica, computación y matemáticas, quienes comenzaron a migrar desde otras plataformas tecnológicas, accedieron al código fuente y comenzaron a crear paquetes especializados. La cantidad de paquetes creció de manera exponencial, extendiendo su capacidad para usarse en otras disciplinas científicas –y sus subdisciplinas correspondientes– como modelado matemático, economía, finanzas, inteligencia de negocios, aprendizaje máquina, entre otras.

Entre las características más importantes del lenguaje *R*, destaca que no es necesario pagar una licencia por su uso; es posible crear librerías externas al núcleo porque es de código abierto; permite la reproducibilidad de la investigación y sus beneficios añadidos, como la transparencia intelectual; tiene una colección de librerías altamente sofisticadas -aprobadas por pares expertos- y organizadas para el análisis estadístico y la visualización de datos.

Actualmente, *R* cuenta con más de 13100 paquetes organizados por una Fundación para el Cómputo Estadístico creada en 1997, cuya sede se encuentra en Viena, Austria (R Core Team, 2018). Esta fundación promueve su desarrollo y aplicación en instituciones públicas y privadas a través del diseño de metodologías de enseñanza que integren

disciplinas formales -Matemática, Lógica, Estadística y Computación Teórica -con disciplinas fácticas- Física, Química, Biología, Economía, Educación, Lingüística, Psicología, entre otras-.

Python, por su parte, es un lenguaje de alto nivel e interpretado, cuyo origen se remonta a inicios de la década de 1990 para realizar actividades de propósito general. Según Johansson (2015), este lenguaje permite escribir código en varios estilos de programación -imperativa y orientación a objetos- con una sintaxis fácil de leer y actualizar. De acuerdo con Solano (2011), entre las ventajas más importantes del uso del lenguaje se encuentran su sintaxis elegante y legible; el entorno de ejecución que coadyuva en la detección de errores de codificación; la posibilidad de escribir código orientado a objetos; su vasta estructura de datos y la expresividad que permite crear programas cortos y funcionales. Por otra parte, Toomey (2018) destaca la importancia de las plataformas de trabajo diseñadas para el análisis de datos en tiempo real con *Python* y el beneficio añadido de mezclar código de otros lenguajes de programación para potenciar sus capacidades analíticas.

El objetivo de esta indagación fue conocer y comparar la opinión de un grupo de estudiantes de Ingeniería respecto de las características más representativas de los lenguajes de programación *R* y *Python* a la luz de su uso en el análisis estadístico de un caso de estudio real y contextual. Fue de interés especial conocer su opinión respecto de paquetes especializados para este tipo de análisis y sus correspondientes entornos de desarrollo integrado.

Esta indagación forma parte de un estudio longitudinal de cuyo propósito es someter a prueba un método de trabajo para formar profesionales de la Ingeniería en Gestión Empresarial en el campo de la Estadística, la Probabilidad y el Análisis de Datos desde una perspectiva holística y cuyas premisas son las siguientes, a saber, [1] la programación computacional es un coadyuvante del desarrollo del pensamiento algorítmico; [2] los alumnos aprenden mejor cuando interactúan de forma activa en el análisis de casos reales que les exija examinar problemas e interactuar con sus pares y profesionales externos en la generación y prueba de opciones de solución; [3] el estudiante universitario debe participar de manera activa en el planteamiento de casos de estudio y emitir opiniones respecto de las alternativas de análisis, métodos de trabajo y herramientas computacionales por utilizar; [4] es factible y viable articular didácticas particulares que integren análisis estadístico, programación computacional en la solución de problemas sociales y del sector productivo en el ámbito de la profesión en que se forman.

2. MÉTODOS Y MATERIALES

Para la realización del estudio se contó con la participación de una muestra de 61 estudiantes universitarios de ambos sexos (31 mujeres y 30 hombres) que cursaban asignaturas de probabilidad y estadística. Se utilizó un muestreo no aleatorio y por conveniencia debido a que no fue posible seleccionar al azar a los participantes. Los restantes 11 estudiantes que integraban la población de quienes cursaban estas asignaturas

no pudieron participar debido a que la actividad se efectuó en una fecha y horario distinto a la programación oficial.

Todos recibieron capacitación equivalente en el uso de ambos lenguajes de programación y su aplicación en el análisis cuantitativo de datos.

A los participantes se les presentó un caso de estudio real y vinculado con el peso en kilogramos de un producto elaborado por una empresa alimenticia, debido a que se les solicitó un reporte escrito, digital y reproducible, fue necesario el uso de métodos estadísticos y desarrollo computacional en los lenguajes de programación *R* y *Python*.

La interacción con *R* se realizó a través del Entorno de Desarrollo Integrado (IDE) *RStudio* versión 1.456. Su elección se realizó debido a los motivos siguientes: [1] fue construido especialmente para *R*; [2] la compañía propietaria ofrece una versión de uso libre con documentación masiva de acceso abierto; [3] permite la depuración eficiente de código; [4] es posible activar el completado automático de instrucciones; [5] tiene actualizaciones continuas. En el caso de *Python*, se usó la Distribución Anaconda porque es de uso gratuito, permite redactar código en varios lenguajes de programación en los entornos de desarrollo *Spyder* y *JupyterLab*, además, de contar con el soporte de una comunidad científica que impulsa su desarrollo constantemente.

Para garantizar los requisitos mínimos de consistencia y validez del caso presentado, se recopilaron los datos *in situ* a través de un muestreo aleatorio estratificado. Un representante del Departamento de Control de Calidad en la empresa verificó el cumplimiento estricto del proceso de selección al azar de los paquetes, su pesaje y registro digital de datos. También, se elaboraron y revisaron las interrogantes más significativas que se requirió plantear y la estructura global del texto. A continuación, se presenta el caso de estudio suministrado:

Uno de los factores más importantes que revisa una empresa que elabora, empaca y vende cereal es la variación de los pesos en cada una de sus presentaciones. De acuerdo con la normatividad interna, el peso neto de los paquetes debe ubicarse entre 500 ± 12 gr. Durante la última semana, se han presentado discrepancias en este factor, provocando retrasos y retrabajos adicionales que constituyen, a la postre, costos y pérdida de competitividad.

En este contexto, se decide investigar la cantidad de producto que contienen los paquetes. Para ello, de la producción total en un turno de trabajo, se seleccionó una muestra aleatoria de 300 paquetes, provenientes de tres lotes (A1L1, A1L2 y A1L3). Posteriormente, se confeccionó una base de datos con el propósito de dar respuesta a las interrogantes siguientes:

[a] De acuerdo con los pesos de los 300 paquetes, ¿en qué medida es representativo el peso promedio?

[b] ¿Qué tan dispersos están los pesos?

[c] ¿En qué grado se está cumpliendo con la normatividad interna?

[d] ¿Cuáles son las ventajas y desventajas de efectuar un análisis de datos separado para cada lote?

[e] Si se realiza un análisis para cada lote, ¿existen diferencias significativas entre ellos?

[f] ¿Cuál es el diagnóstico general?

[g] ¿Cuáles son sus recomendaciones?

[h] Respecto del método de trabajo seguido en el análisis de caso, ¿en qué medida consideran ustedes pertinente utilizar métodos bayesianos?

También, se les suministraron dos libros digitales y especializados en Análisis de Datos. El primero titulado *R for Data Sciencie: import, tidy, transform, visualize and model data* escrito por *Hadley Wickham* y *Garret Grolemoud* (Wickham & Grolemund, 2016). El título del segundo fue *Python for Data Analysis: Data wrangling with Pandas, Numpy, and Ipython*, elaborado por *Wes McKinney* de manera individual (McKinney, 2012). Ambos textos tienen un amplio reconocimiento internacional y se promueven como lectura obligada en entornos educativos, de investigación y desarrollo profesional independiente.

El estudio se efectuó en las instalaciones del Centro de Información de la Universidad, porque se asignó un área de trabajo, lo suficientemente amplia, para que los participantes colocaran computadoras portátiles, calculadora científica y material de trabajo que consideraran pertinente. Así, el protocolo seguido en el estudio se integró con las siguientes actividades:

Un día antes de iniciar el estudio se les comunicó de manera verbal las actividades por efectuar. Posteriormente, en una reunión plenaria se realizó una presentación del caso a examinar; se hizo énfasis en describir los estándares mínimos de cumplimiento que debía reunir el reporte final en términos de formato y estructura. También, se indicó que contaban con cinco horas como duración máxima para la entrega del informe y un cuestionario cuyo propósito fue copiar información que permitiera comparar las características de ambos lenguajes (Anexo 1). Para evitar interpretaciones erróneas en el protocolo por seguir, se realizó una sesión de preguntas y respuestas.

La reunión plenaria tuvo una duración aproximada a los 35 minutos y acto seguido, se dio inicio al análisis del caso.

3. Resultados

La actividad se realizó de acuerdo con el plan de trabajo. No hubo ausencia de participantes ni valores faltantes en el cuestionario. Entre los resultados más representativos se pueden enunciar los siguientes:

Con excepción de un participante, el resto terminó y entregó el reporte del análisis del caso en un período menor a las cinco horas. El participante que no terminó en el plazo previsto señaló que su demora se debió a problemas de compatibilidad de algunos paquetes de *R* y *LaTeX*; también, subrayó que debió descargarlos de la red e instalarlos de nuevo.

El ítem 1 solicitó información relacionada con los paquetes usados para el cálculo estadístico, la representación gráfica y tabular con el lenguaje R. únicamente 11 participantes (18,03%) utilizaron los paquetes incluidos en el núcleo de R; los restantes 50 (81,97%) prefirieron usar librerías externas, destacando *DataExplorer* (Boxuan, 2018) que fue empleado por 16 (26,22%). En la figura 1 se presenta un resumen global del ítem y su comparativo por sexo.

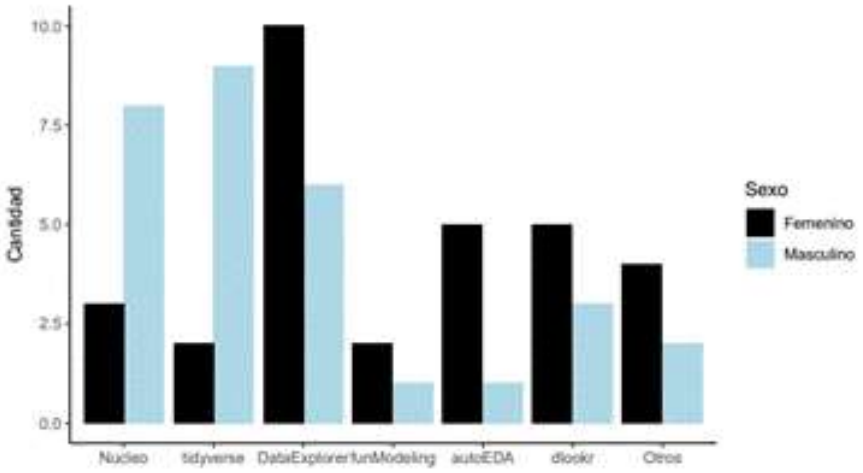

Figura 1: Paquetes de R utilizados para cálculo estadístico, representación gráfica y tabular

En relación con el uso de *R* para elaborar el informe reproducible, treinta y dos (52,45%) participantes indicaron usar *R Markdown*, 23 (37,70%) *LaTeX – R*, cinco (8,91%) otra tecnología y un participante del sexo masculino (1,62%) respondió *Bookdown*. La distribución de las respuestas puede observarse en la Figura 2.

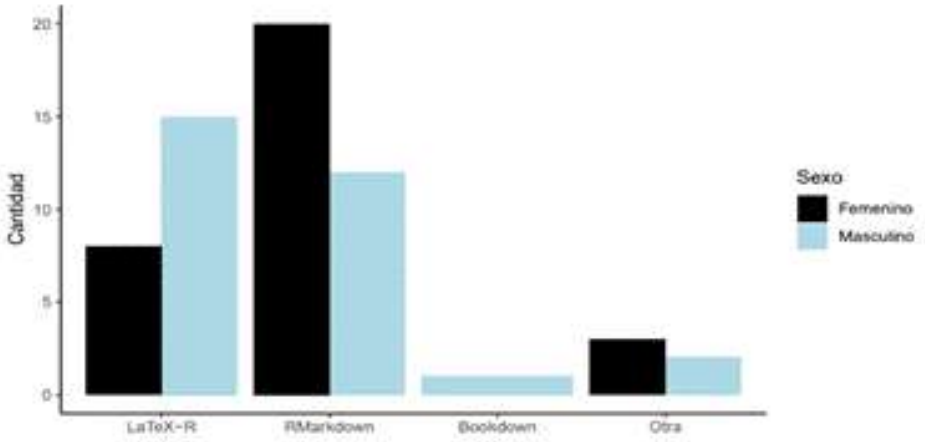

Figura 2: Paquetes de R utilizados para elaborar el informe reproducible

Es necesario señalar que los cinco participantes que respondieron haber utilizado otra tecnología para redactar el informe, no mencionaron cuál fue.

Por otra parte, en el caso del lenguaje de programación Python, se encontró que 54 (88,52%) utilizaron las librerías *Scipy*, *Matplotlib*, *Numpy* y *Pandas*. La proporción por sexo fue equivalente como se puede observar en la Figura 3. Debe decirse que el texto *Python for Data Analysis: Data Wranling with Pandas, Numpy, and Ipython* que fue escrito por *Wes McKinney* (McKinney, 2012), promociona el uso de esas librerías para efectuar análisis estadísticos formales.

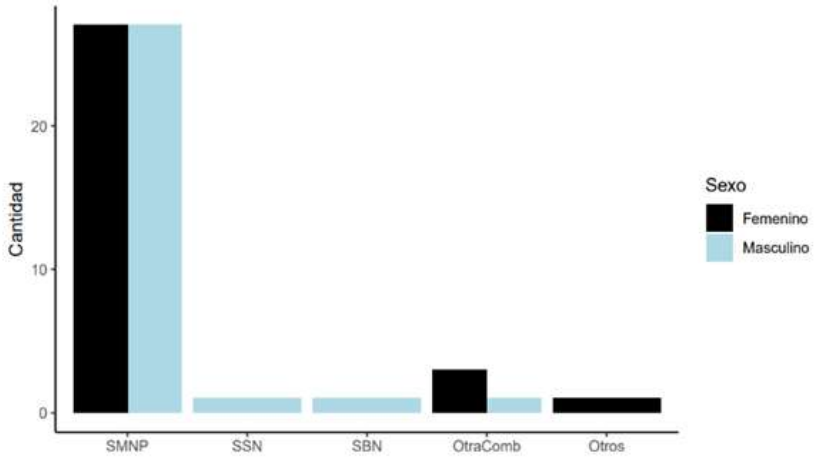

Figura 3: Librerías de Python utilizadas para cálculo estadístico, representación gráfica y tabular

Para redactar el informe con atributos de reproducibilidad, 41 (67,21%) mencionó *Markdown Python*, 19 (13,14%) la triada *Markdown-Python-LaTeX* y un participante (1,63%) una combinación diferente sin especificar cuál.

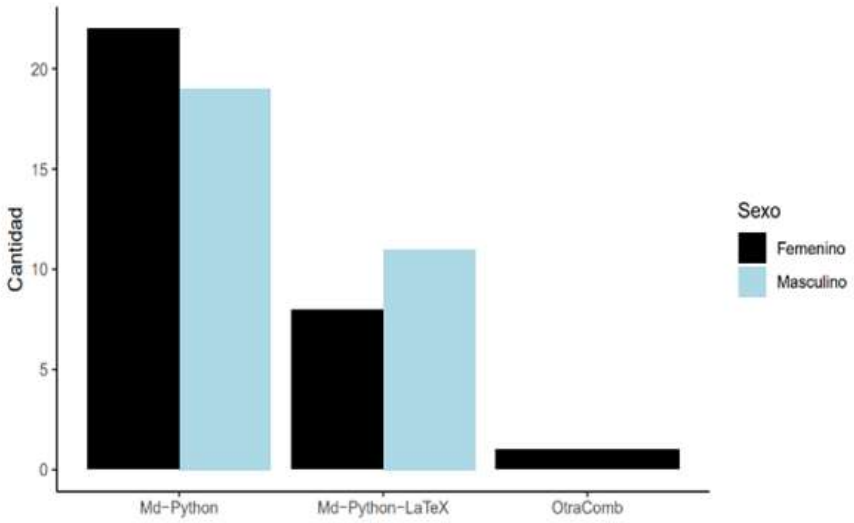

Figura 4: Librerías de Python utilizadas para elaborar informe reproducible

También, se solicitó una valoración global en el análisis del caso de estudio ambos lenguajes de programación y sus paquetes asociados. Dicha valoración se integró con las categorías siguientes: [1] Facilidad de uso; [2] Estructura de datos; [3] Potencia gráfica; [4] Organizar tabular; [5] Paquetes para Análisis Exploratorio de Datos (AED); [6] Integrar un informe reproducible. De manera particular se les pidió que emitieran una calificación en la escala del 1 al 10.

Los resultados indican que los participantes se decantaron por *Python* en las categorías relacionadas con la facilidad de uso, la importación y organización de datos. La redacción de código tiene similitud con el lenguaje natural, trayendo consigo alta legibilidad y mantenimiento en etapas subsecuentes. De igual manera, sus librerías para organizar información proveniente de base de datos masivas constituyen uno de los resultados más importantes del lenguaje.

Respecto de la potencia gráfica, ambos lenguajes fueron evaluados con puntajes altos. Sin embargo, los paquetes gráficos de *R* y la variedad tan extensa de librerías como *tidyverse, DataExplorer, funModeling, dlookr, autoEDA*, …, inclinaron las preferencias a favor de este lenguaje.

La capacidad para elaborar informes reproducibles fue puntuada más alto en *R* que en *Python*. *JupyterLab* es un entorno potente, versátil y en desarrollo continuo que permite redactar código en varios lenguajes, entre ellos *R* y *Python*. El soporte de otros lenguajes es menor en *R Markdown*, no obstante, la lógica subyacente para elaborar documentos dinámicos y reproducibles es superior.

En la Figura 5, se representa una matriz de correlaciones entre las categorías de análisis señaladas antes.

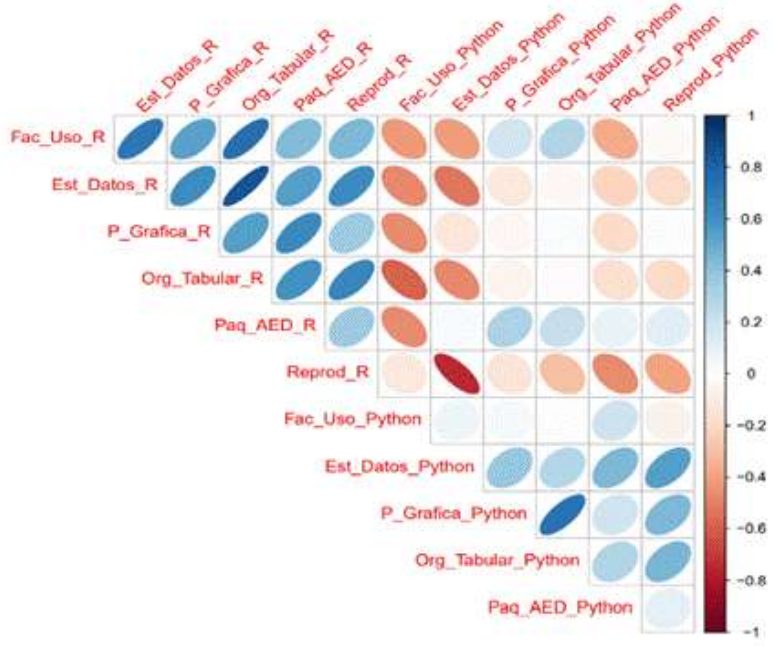

Figura 5: Matriz de correlaciones

Cuevas, H., Solís, C. y Silva, I. (2019). Programación computacional y análisis de datos en educación estadística. *Areté. Revista Digital del Doctorado en Educación de la Universidad Central de Venezuela*, 5 (9), 11 – 27.

Las correlaciones más fuertes (0,859 y 0,758) se presentaron entre la organización de *R*, la facilidad de uso y la capacidad para estructurar datos. También, puede observarse la ponderación alta emitida a los paquetes para AED en *R*. En el caso de *Python*, la correlación positiva más alta (0,730) se presentó entre la organización tabular y su potencia gráfica.

En la parte final del cuestionario, se les planteó una situación hipotética a los participantes en que debían realizar un análisis exploratorio de datos. Se presentaron cuatro opciones únicas de respuesta: [1] R; [2] Python; [3] Uno u otro; [4] ambos. *R* fue elegido por 27 participantes (44,3%), especialmente, por los participantes del sexo masculino. En el caso de *Python*, fue seleccionado por 19 (31,1%) especialmente por las mujeres. Una cantidad reducida, 9 (14,8%) señaló que usaría uno de los dos de manera indistinta; una cantidad menor respondió que usaría ambos; esta respuesta es interesante porque cuatro de los informes entregados se crearon mezclando código de *R* y *Python* de manera paralela. En la Figura 6 puede observarse un panorama global de las respuestas.

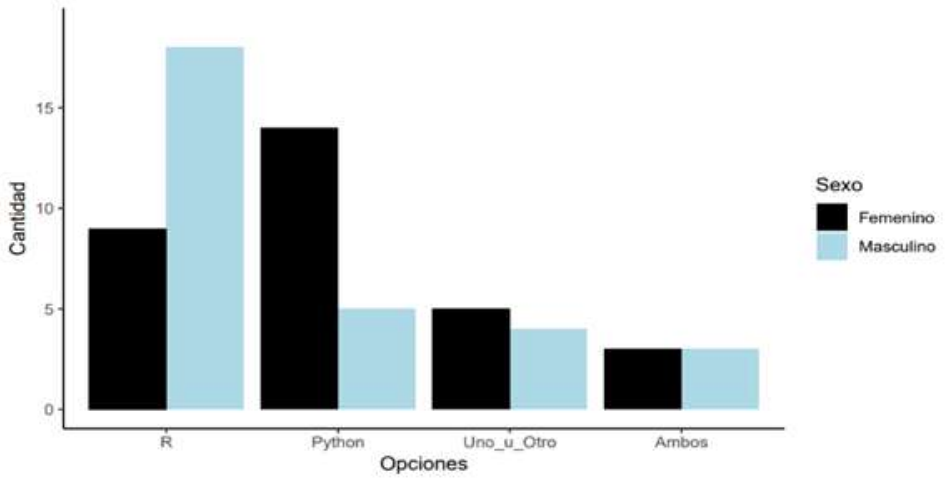

Figura 6: Lenguaje de programación preferente para análisis datos

4. CONCLUSIONES Y RECOMENDACIONES

El objetivo de esta indagación fue conocer y comparar la opinión de un grupo de estudiantes de Ingeniería respecto de las características más representativas de los lenguajes de programación *R* y *Python* a la luz de su uso en el análisis estadístico de un caso de estudio real y contextual.

El estudio fue de tipo exploratorio, en consecuencia, los resultados obtenidos son válidos única y exclusivamente en la comunidad donde se realizó la indagación. No se probaron conjeturas propiamente dichas; en todo caso permitió plantear y reconfigurar interrogantes de investigación entre las que sobresalen las siguientes: ¿en qué medida es necesario desarrollar didácticas particulares en la enseñanza y aprendizaje de la estadística que se alejen de las corrientes teóricas y metodológicas tradicionales dominantes durante

las últimas tres décadas?, ¿en qué grado es factible articular un método pedagógico basado en el aprendizaje basado en retos, la programación computacional y el análisis de datos?

En relación con la actividad desarrollada por los participantes, 60 de ellos la concluyeron. Los reportes digitales cumplieron los criterios normativos de formato establecidos. Como se indicó anteriormente, más que evaluar la habilidad de los participantes en explorar los datos desde la perspectiva cuantitativa-analítica, interesó conocer y comparar las opiniones vertidas en relación con paquetes especializados y entornos de desarrollo integrado creados *exprofeso*.

De acuerdo con los resultados obtenidos, ambos lenguajes fueron valorados con puntajes altos. La curva de aprendizaje para *Python* es menos pronunciada que la de *R*, sin embargo, el 44,3% elegiría el segundo para efectuar análisis exploratorios. Los participantes del sexo masculino se decantaron en su mayoría por *R*; las del sexo femenino por *Python*.

Los resultados, también, indican que los participantes del sexo masculino prefieren usar los paquetes que vienen integrados en el núcleo de *R*. Las integrantes del sexo femenino prefirieron usar paquetes externos como *DataExplorer, funModeling, autoEDA y dlookr* para sus análisis. Dichos paquetes integran herramientas para visualizar y tabular información de generar resúmenes estadísticos.

La reproducibilidad computacional es una característica irreductible de cualquier informe de investigación que promueva la transparencia y certidumbre intelectual. El lenguaje de marcas *Markdown* se ha convertido en un estándar de facto para redactar textos por su análisis; *R Markdown* es un paquete que articula la versatilidad de este lenguaje con la potencia de cálculo y visualización de *R*. Los participantes optaron por esa combinación en detrimento del uso de *LaTeX*. De igual manera, la vinculación *Markdown-Python* constituyó en su momento la contraparte de este campo.

Aún cuando el propósito central del estudio no fue examinar la pertinencia de utilizar lenguajes de programación computacional, en procesos de mediación enseñanza-aprendizaje-estadística, los resultados obtenidos, la observación de la interacción de los participantes y comentarios vertidos por escrito, son solicitud expresa; dan indicios positivos de su incorporación. La implementación de lenguajes como *R* y *Python* ofrece más ventajas que desventaja para el aprendizaje de la estadística, la probabilidad y el análisis de datos. El desarrollo de habilidades de pensamiento lógico y algorítmico constituye una hipótesis repetible de ser probada empíricamente.

Se recomienda efectuar una indagación que vaya más allá de la exploración y descripción. Un estudio de corte experimental en que se someta a prueba un método pedagógico que integre prácticas diferentes a las que han dominado el espectro de la educación estadística durante las últimas tres décadas. El aprendizaje basado en retos es una alternativa real que ofrece varias ventajas en su incorporación en la educación en todos sus niveles, especialmente en el universo.

REFERENCIAS

Ben-Zvi, D., y Garfield, J. (2004) Statistical Literacy, Reasoning, and Thinking: Goals, Definitions, and Challenges. En: Ben-Zvi D., Garfield J. (eds.) The Challenge of Developing Statistical Literacy, Reasoning and Thinking. Springer, Dordrecht.

Boxuan, C. (2018). *Package DataExplorer (software).* Tomado de https://CRAN.R-project.org/package=DataExplorer.

Chambers, J. (2008). *Software for data analysis: programming with R. Springer Science & Business Media.*

Gal, I. (2004). Statistical literacy. In The challenge of developing statistical literacy, reasoning and thinking (pp. 47-78). Springer, Dordrecht.

Ihaka, R. (1998). R: Past and future history. *Computing Science and Statistics, 392396.*

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*. 5(3), 299-314.

Johansson, R. (2015). *Numerical Python. A Practical Techiques Approach for Industrial*. Uruyasu, Japan: Apress.

Kabacoff, R. (2018). Data Visualization with R. EEUU: Wesleyan University.

Kassambara, A. (2017). *R Graphics Essentials for Great data Visualization*. México: IPP.

López, C. C., & Ramírez, M. M. O. (2018). La opinión de los estudiantes sobre el uso de las metas de aprendizaje de la estadística en cursos introductorios en la universidad veracruzana. Investigación Operacional, 39(2), 181-191.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.

R Core Team, F. (2018). *Writing R Extensions. Viena, Austria: R Foundation Project.* Tomado de https://cran.r-project.org/doc/manuals/r-release/R-exts.html

Rüde, U., Willcox, K., McInnes, L., & Sterck, H. (2018). Research and Education in Computational Science and Engineering. SIAM Review, 60 (3), 707-754. Tomado de https://doi.org/10.1137/16M1096840

Solano, J. (2011). *Introducción a la programación en Python*. Costa Rica: Editorial Tecnológica del ITCR.

Toomey, D. (2018). *Jupyter Cookbook*. Birmingham, UK: Packt Publishing.

Tukey J. (1977). Exploratory Data Analysis. UK: Pearson

Van Dijke-Droogers, M., Drijvers, P., & Tolboom, J. (2017). Enhancing statistical literacy. In CERME 10

Watson, J. (2013). Statistical literacy at school: Growth and goals. Routledge.

Wickham, H. & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. O' Reilly Media, Inc.

Zetterqvist, L. (2017). Applied problems and use of technology in an aligned way in basic courses in probability and statistics for engineering students-a way to enhance understanding and increase motivation. *Teaching Mathematics and its Applications:*

An International Journal of the IMA. 36 (2), 108-122. Tomado de https://doi.org/10.1093/teamat/hrx004

Zimek, A., & Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(6), 1280.

Humberto Cuevas. Ingeniero Industrial, Especialista en Docencia, Maestro en Ciencias y Doctor en Educación. Sus líneas de trabajo son Estadística Computacional, Educación Estadística e Innovación y Desarrollo Tecnológico en Educación. Tiene más de 50 participaciones como conferencista, ponente y tallerista en foros académicos realizados en Centroamérica, Sudamérica y Europa, así como publicaciones en revistas arbitradas e indexadas. Actualmente se desempeña como profesor-investigador en el Tecnológico Nacional de México.

Cristina Solís Moreira. Docente y Filóloga por la Universidad de Costa Rica, Licenciada en Ciencias de la Educación con énfasis en Enseñanza del Español por la Universidad de las Ciencias y el Arte de Costa Rica y candidata al grado de Maestría en Tecnología Educativa por la Universidad Estatal a Distancia. Las líneas de investigación que cultiva son Lingüística Computacional, Diseño de Recursos Educativos Virtuales, Análisis en Modismos del Lenguaje.

Ixrael Silva Contreras. Ingeniero Industrial por el Instituto Tecnológico de Chihuahua y Maestro en Ciencias en Matemática Educativa por el Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México. Sus líneas de investigación son Tecnología Computacional Aplicada a la Educación y Educación Matemática Universitaria. Actualmente trabaja como profesor-investigador en el Instituto Tecnológico de Chihuahua II del Tecnológico Nacional de México y la Universidad La Salle Campus Chihuahua.

ANEXO

Cuestionario

El propósito de este cuestionario es obtener información respecto del procedimiento estadístico–computacional que siguió para dar respuesta a las interrogantes planteadas en el caso de estudio.

Instrucciones: Dé respuesta a cada uno de los ítems que se presentan a continuación.

Sexo: _____
Carrera: _____

Lenguaje de programación R
R1. Para efectuar cálculo estadístico, representación gráfica y tabular utilicé:

[ ] Módulos del núcleo de R.
[ ] tidyverse (ggplot2, tibble, tidyr, readr, purrr, dplyr,...)
[ ] DataExplorer
[ ] funModeling
[ ] autoEDA
[ ] dlookr
[ ] Otros: _____

R2. En la redacción del informe reproducible utilicé:

[ ] LaTeX--R
[ ] R Markdown
[ ] Bookdown

[ ] Otra combinación: _____

Lenguaje de programación Python
Py1. Para efectuar cálculo estadístico, representación gráfica y tabular utilicé:

[ ] Scipy, Matplotlib, Numpy, Pandas
[ ] Scipy, Seaborn, Numpy
[ ] Scipy, Bokeh, Numpy
[ ] Otra combinación: _____
[ ] Otros: _____

Py2. En la redacción del informe reproducible utilicé:

[ ] Markdown--Python
[ ] Markdown--Python--LaTeX
[ ] Otra combinación: _____

Valoración global

Lea con detenimiento cada categoría presentada en la tabla siguiente y califique en la escala del 1 al 10:

Tabla 1. Valoración global de ambos lenguajes.

Categoría	R	Python
Facilidad de uso		
Estructura de datos		
Potencia gráfica		
Organización tabular		
Paquetes para AED		
Integrar un informe reproducible		

Para realizar un Análisis Exploratorio de Datos, elijo como primera opción:
[ ] R
[ ] Python
[ ] Uno u otro
[ ] Ambos

Cuevas, H., Solís, C. y Silva, I. (2019). Programación computacional y análisis de datos en educación estadística. *Areté. Revista Digital del Doctorado en Educación de la Universidad Central de Venezuela*, 5 (9), 11 – 27.