ANALES DE LA
Academia de Ciencias de Cuba

CIENCIAS SOCIALES Y HUMANÍSTICAS
Artículo de revisión

Herramientas de código abierto para el análisis estadístico en investigaciones científicas

Rudibel Perdigón Llanes [1*] https://orcid.org/0000-0001-7288-6224
María Teresa Pérez Pino [2*] https://orcid.org/0000-0001-5923-204X

1 Empresa Frutas Selectas, Ministerio de la Agricultura. Pinar del Río, Cuba
2 Universidad de las Ciencias Informáticas. La Habana, Cuba

**Autor para la correspondencia:* rudibel@frutaspr.co.cu; rperdigon90@gmail.com

Revisores

Ramón Quiza Sardiñas
Universidad de Matanzas. Matanzas, Cuba

Enrique Juan Marañón Reyes
Universidad de Oriente. Santiago de Cuba, Cuba

Editor

Amanda Gómez Bahamonde
Academia de Ciencias de Cuba. La Habana, Cuba

Traductor

Yoan Karell Acosta González
Academia de Ciencias de Cuba. La Habana, Cuba

RESUMEN

Introducción. En esta investigación se realizó un análisis de las potencialidades y aplicabilidad de las principales herramientas de análisis estadístico de código abierto, existentes hasta el momento, en el ámbito de la investigación científica, en aras de contribuir a fomentar su utilización por investigadores cubanos. **Métodos**. Se desarrolló una revisión de la literatura donde se emplearon como métodos científicos el analítico sintético, el histórico lógico, la triangulación teórica y el análisis documental. **Resultados y Discusión.** Se evidenció la escasa utilización de soluciones libres para el análisis estadístico en investigaciones científicas publicadas por autores cubanos en revistas del catálogo Scielo Cuba durante el período 2015-2020. Se efectuó un análisis comparativo de las soluciones libres GNU/Octave, GNU-PSPP, el entorno informático R y la herramienta comercial IBM SPSS que evidenció la viabilidad y superioridad del entorno R para el tratamiento estadístico de los datos y su representación gráfica en investigaciones científicas.

Palabras clave: metodología de investigación; paquete estadístico; software libre

Open source tools for statistical analysis in scientific research

ABSTRACT

Introduction. In this research, an analysis of the potentialities and applicability of the main open source statistical analysis tools, existing so far in the field of scientific research, was carried out, in order to contribute to promoting their use by Cuban researchers. **Methods.** A literature review was carried out; the synthetic analytical method and the logical historical one, as well as theoretical triangulation and documentary analysis were used as scientific methods. **Results and Discussion.** A result was that there is scarce use of free solutions for statistical analysis in scientific research published by Cuban authors in journals of Scielo Cuba catalog during the period 2015-2020. A comparative analysis of the free solutions GNU/Octave, GNU-PSPP, the computer environment R, and the commercial tool IBM SPSS was carried out, which evidenced the viability and superiority of the R environment for the statistical treatment of data and its graphic representation in scientific research.

Keywords: research methodology; statistical package; free software

INTRODUCCIÓN

Una de las fases del proceso de investigación científico es el análisis de datos. [1-3] Este análisis se realiza mediante la aplicación de diferentes técnicas estadísticas que permiten a los investigadores medir, examinar y comprender la realidad que los rodea. [1-4]

La estadística ocupa un lugar importante dentro de la investigación científica porque brinda a los investigadores la posibilidad de evaluar cuantitativamente hipótesis de investigación, desarrollar modelos predictivos, estimar parámetros, confeccionar instrumentos de investigación y analizar experimentos. [3,5,6,7] Sin embargo, independientemente de las métricas empleadas para el análisis de datos, los académicos se enfrentan a diferentes dilemas para seleccionar las herramientas estadísticas apropiadas a utilizar en sus investigaciones. [1,8] Según, Azman *et. al.* [1] esto se ha debido a las afectaciones que puede ocasionar la falta de precisión durante el análisis de datos en los resultados científicos.

La introducción de las tecnologías digitales y las herramientas de análisis estadístico como parte de estas en los procesos de enseñanza, aprendizaje e investigación es una necesidad incuestionable para alcanzar un buen desempeño científico y académico. [9] Estas herramientas han constituido un componente esencial en el diseño de las investigaciones científicas y han contribuido a facilitar su reproducibilidad. [10] Varios autores han considerado pertinente emplear herramientas informáticas durante el análisis de datos cuantitativos en las investigaciones científicas para facilitar su registro, depuración, tratamiento, transformación de variables, su procesamiento numérico o estadístico y su representación gráfica. [3,10] No obstante, el uso impropio de estas herramientas por muy sofisticadas que parezcan, es uno de los factores que han conducido a la obtención de resultados incorrectos en investigaciones científicas. [3]

La selección del mejor *software* y el análisis estadístico apropiado puede representar una actividad compleja para los investigadores que depende en gran medida del tipo de investigación desarrollada y de las características de estas herramientas. [1,8,11] Los autores Abbasnasab *et. al.* y Sued [11,12] han considerado que su ámbito de desarrollo, su tipo de licencia de uso, la complejidad de su interfaz de usuario, sus funcionalidades para el análisis estadístico y el manejo de los datos han constituido criterios relevantes para seleccionar la herramienta de análisis adecuada.

Los autores Avello *et. al.* [5] identificaron que las herramientas de procesamiento estadístico que se han establecido en la investigación científica cubana, han sido opciones comerciales con precios bien altos en licencias. Esta situación ha dificultado la reproducibilidad de los resultados y el desarrollo soberano de la ciencia cubana debido a las limitaciones económicas y financieras impuestas al país que han impedido la adquisición de licencias de uso para estas aplicaciones. Además, las limitantes anteriores han obligado a no pocos académicos cubanos a violar principios éticos por el uso no autorizado de estas soluciones durante sus investigaciones. Por esta razón, autores como Avello *et. al.* y Torres [5,13] han considerado beneficioso para la sociedad científica cubana el empleo de herramientas libres y de código abierto para el análisis estadístico en su práctica científica, con el fin de contribuir a fortalecer el crecimiento sostenible y soberano de la ciencia en el país.

Los paquetes de análisis estadístico de código abierto son libres de costo, permiten a los investigadores comprobar la idoneidad y veracidad de sus algoritmos subyacentes y adecuarlos a las necesidades específicas de sus investigaciones. [14] En la literatura consultada se identificaron diversos estudios que evidenciaron la pertinencia de emplear herramientas de análisis estadístico de código abierto en campos como la psicología, la medicina, las ingenierías, las ciencias pedagógicas y la agrícola. [1,5,13,15-17] Asimismo, Campanioni *et. al.* [9] determinaron la viabilidad de sustituir asistentes matemáticos y estadísticos con licencias privativas, por soluciones libres en las universidades cubanas. Sin embargo, el uso de estas soluciones por los académicos cubanos aún es incipiente. [5,13] Según Abbasnasab *et. al.* [11] las publicaciones orientadas a evaluar la usabilidad, capacidades y funcionalidades técnicas de las herramientas de análisis estadístico basadas en software libre son escasas. El objetivo de esta investigación consistió en analizar las potencialidades y la aplicabilidad de las principales herramientas de análisis estadístico de código abierto, existentes hasta el momento en el ámbito de la investigación científica, en aras de contribuir a fomentar su utilización por investigadores cubanos.

DESARROLLO

Métodos

Se realizó una revisión de la literatura donde se emplearon métodos científicos del nivel teórico y empírico. Para la búsqueda de información se utilizaron las bases de datos *Google Scholar*, *Scielo* y *ScienceDirect* que son motores de búsqueda gratuitos y abarcan gran cantidad de artículos académicos. Como criterios de búsqueda se emplearon los términos: *software*, paquete, aplicación, herramienta, análisis estadístico y analítica de datos en idioma inglés y español. La combinación de los términos anteriores se realizó mediante los operadores lógicos AND y OR.

Se utilizaron como métodos teóricos el analítico-sintético, histórico-lógico y triangulación que permitieron el estudio de

la bibliografía existente relacionada con el objeto de estudio, analizar su evolución y disminuir el sesgo en la investigación. Del nivel empírico se empleó el análisis documental que posibilitó identificar mediante la revisión de diferentes fuentes, las herramientas de análisis estadístico más populares en la actualidad y las más empleadas por investigadores cubanos durante los últimos 5 años. Para lograr este propósito, se realizó una búsqueda en revistas científicas indexadas al catálogo *Scielo* Cuba hasta enero de 2021.

Este catálogo ha aglutinado revistas de calidad y de circulación internacional de países de América Latina, el Caribe, España y Portugal y posee una marcada importancia para la visibilidad de la ciencia en Cuba. [18] Su utilización en la presente investigación estuvo fundamentada por los resultados obtenidos por Galbán-Rodríguez *et al.* [19] que establecieron que alrededor de un 77 % de las investigaciones realizadas por científicos cubanos son publicadas en revistas nacionales. Se seleccionaron los artículos publicados durante el período comprendido entre enero de 2015 y diciembre de 2020.

Herramientas de análisis estadístico más utilizadas

Para identificar las herramientas de análisis estadístico más populares en la actualidad, se realizó un estudio de las principales tendencias del mercado. Estos resultados fueron descritos en la figura 1.

Aunque las soluciones comerciales han dominado ampliamente el mercado, se han identificado herramientas de código abierto como *RStudio* (entorno informático R) y GNU/*Octave* con un posicionamiento positivo dentro de este.

La herramienta *RStudio* es un entorno de desarrollo integrado (IDE, por sus siglas en inglés) para el lenguaje de programación orientado a objetos R, dedicado a la computación estadística y los gráficos que incluye una consola, un editor de sintaxis que apoya la ejecución de código y herramientas para el trazado, depuración y gestión del espacio de trabajo. [20] La capacidad de análisis estadístico de *RStudio* puede ampliarse mediante los paquetes desarrollados para el entorno R disponibles a través de la familia de sitios de Internet denominada Red Integral de Archivos R (CRAN, por sus siglas en inglés) que cubren una amplia gama de estadísticas modernas. [21,22]

Por su parte, GNU/*Octave* es un lenguaje de alto nivel diseñado principalmente para cálculos numéricos y algebraicos mediante una interfaz de línea de comandos. [23] Esta herramienta es fácilmente extensible y personalizable a través de funciones definidas por el usuario escritas en el propio lenguaje de *Octave* o usando módulos cargados dinámicamente. [23]

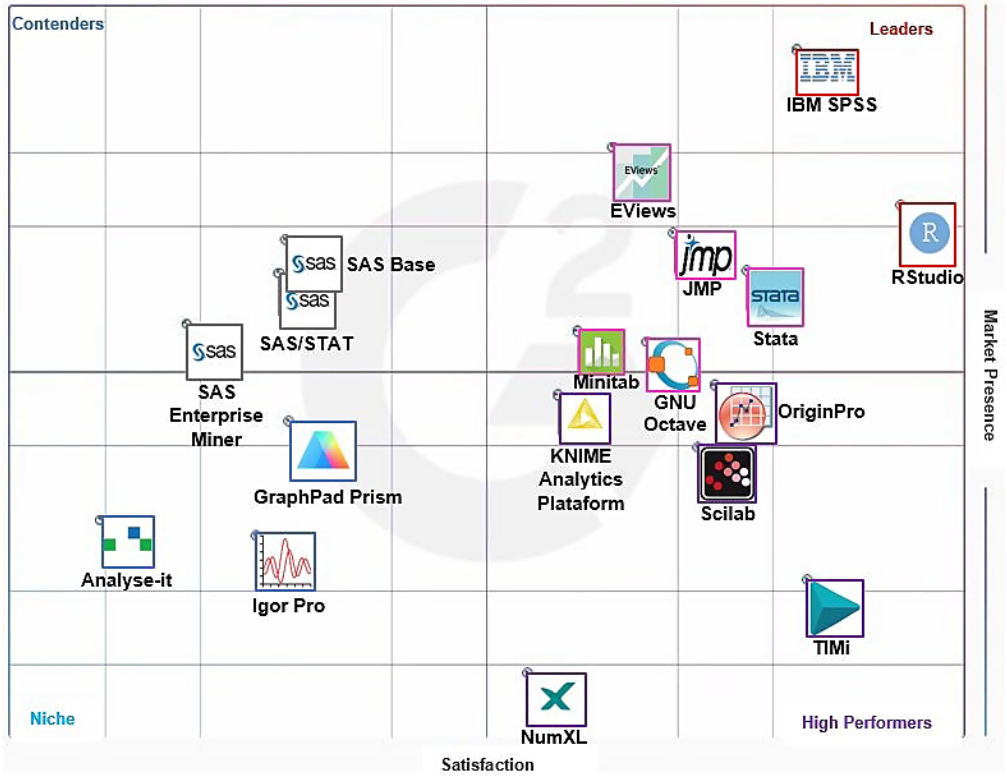

Fig. 1. Herramientas de análisis estadístico líderes en el mercado en 2021.

Fuente: G2 Crowd, Inc.

Herramientas de análisis estadístico empleadas por investigadores cubanos

Con el objetivo de identificar las soluciones de análisis estadístico más utilizadas por investigadores cubanos, se realizó una búsqueda en artículos publicados en revistas académicas. El cúmulo de revistas científicas existentes en la actualidad y de trabajos publicados por profesionales extranjeros en coautoría con investigadores cubanos, ha complejizado el desarrollo de la tarea mencionada. Por esta razón, solo fueron considerados trabajos indexados al catálogo *Scielo* Cuba, donde se hayan sido utilizadas estas herramientas y el autor principal fuera cubano.

Como resultado del proceso de búsqueda se obtuvieron un total de 1805 artículos, que correspondían a las siguientes áreas del conocimiento: Ciencias Médicas (71,1 %), Ciencias Agrícolas (25,48 %), Ciencias Sociales (2,22 %) e Ingenierías (1,2 %). Las revistas con mayor cantidad de publicaciones donde se identificó la aplicación de paquetes de análisis estadístico fueron: *Cultivos Tropicales* (165 publicaciones), *MEDISAN* (144 artículos), *Archivo Médico de Camagüey* (96 trabajos), *Revista de Ciencias Médicas de Pinar del Río* (90 publicaciones) y *Revista de Producción Animal* (71 investigaciones). A continuación, se muestran en la figura 2 los paquetes estadísticos más utilizados en las investigaciones identificadas.

Los resultados descritos en la figura 2 demuestran que las soluciones comerciales fueron ampliamente utilizadas por los científicos cubanos en sus investigaciones, fundamentalmente las herramientas IBM SPSS, STATGRAPHICS y STATISTICA. Se determinó que el uso de herramientas de código abierto para el análisis estadístico en la práctica científica cubana durante el período analizado fue prácticamente nulo, estas representaron solo el 0,33 % del total de las soluciones identificadas. Se evidenció que *RStudio* y GNU-PSPP fueron las únicas aplicaciones de código abierto utilizadas.

Análisis comparativo de las herramientas identificadas

La triangulación de los resultados obtenidos permitió identificar que *RStudio* es la herramienta de análisis estadístico de código abierto más popular en la actualidad. En relación a este hallazgo, se realizó un análisis comparativo de las potencialidades que ha brindado esta solución respecto a sus similares libres GNU/*Octave* y GNU-PSPP para el análisis estadístico en investigaciones científicas. En el análisis se incluyó la herramienta privativa IBM SPSS por ser el paquete de análisis estadístico más utilizado en Cuba y el mundo.

Según Niu *et. al.* [21] el entorno informático R ha permitido la realización de análisis estadísticos, manejo y visualización de datos, analítica de textos, aprendizaje de máquinas, entre otras funciones, aplicables en diferentes áreas del conocimiento como la medicina, [24] ciencias agrícolas, [25] ciencias empresariales, [21] ciencias sociales [26,27] e ingenierías. [28]

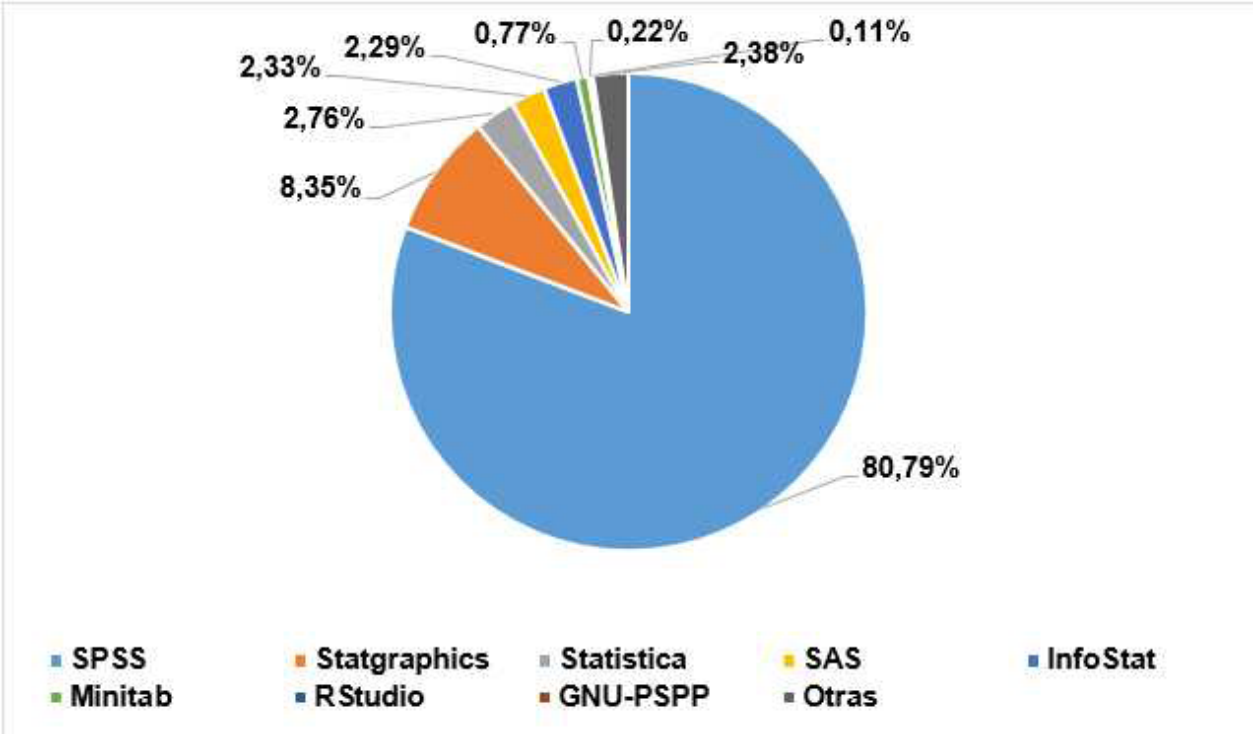

Fig. 2. Herramientas de análisis estadístico utilizadas por investigadores cubanos durante el período 2015-2020.

Fuente: elaboración propia.

A su vez, GNU/*Octave* ha sido utilizado por la comunidad científica para realizar analítica de datos, procesamiento de imágenes, investigaciones económicas, análisis estadístico, minería de datos, procesamiento de señales y aprendizaje de máquina. [29] Por su parte, GNU-PSPP ha sido ampliamente utilizado en investigaciones de las ciencias sociales debido a su similitud con el *software* comercial IBM SPSS [30].

Se analizaron elementos como: el tipo de interfaz de usuario (GUI, por sus siglas en inglés) de cada herramienta, su compatibilidad con diferentes formatos de datos, tipos de pruebas estadísticas soportados y sus capacidades para la representación gráfica de los datos, que han constituido criterios relevantes para evaluar este tipo de aplicaciones. [11,31] Se contemplaron los tipos de pruebas estadísticas más utilizadas en investigaciones científicas según los criterios de Oliveira Neto *et. al.* [4] Flores Ruiz *et. al.* [32] y Hernández *et al.* [33] Se emplearon las últimas versiones estables disponibles de las soluciones mencionadas hasta el momento de realizada la presente investigación, estas fueron: el entorno R (4.0.3), GNU/Octave (6.0.1), GNU-PSPP (1.4.1) e IBM SPSS (27.0.1). La tabla 1 muestra las características de estas herramientas, la información fue obtenida de los sitios oficiales de cada solución.

La tabla anterior demuestra las potencialidades del entorno informático R como herramienta de análisis estadístico en investigaciones científicas, respecto a GNU/*Octave*, GNU-PSPP y la popular herramienta comercial IBM SPSS. Se evidenció que el entorno informático R ha permitido importar y exportar datos en disímiles formatos compatibles inclusive con diferentes asistentes estadísticos comerciales como IBM SPSS, SAS y STATA, al contrario de GNU/*Octave* y GNU-PSPP que poseen limitaciones en este aspecto. La herramienta GNU-PSPP se ha destacado por el acabado de su GUI que posee un aspecto muy similar a IBM SPSS. No obstante, R ha destacado por la disponibilidad de diferentes paquetes y librerías que han facilitado el trabajo con este lenguaje más allá de la línea de comandos como es el caso de *R-Commander*, *RKWard*, *jamovi* y *Deducer*, este último también ha simulado la apariencia de IBM SPSS.

Respecto a la representación gráfica de los datos, el entorno informático R ha permitido elaborar una amplia variedad de gráficos como son los dendogramas, mapas de calor, gráficos de burbujas y nubes de palabras que han resultado de gran utilidad para el análisis de datos. Los dendogramas posibilitan realizar análisis de *cluster* jerárquicos y visualizar la cercanía entre diferentes unidades de análisis, los gráficos de burbujas facilitan el estudio de la producción bibliométrica de determinado autor, los mapas de calor ilustran los niveles de correlación bivariada entre variables y las nubes de palabras permiten realizar valoraciones preliminares sobre las principales categorías de un marco teórico-referencial mediante analítica de textos. [13,34] En este aspecto, R ha destacado respecto a IBM SPSS porque este último ha requerido del asistente IBM *Modeler* para elaborar gráficos de nubes de palabras durante la analítica de textos.

Se evidenció que al igual que IBM SPSS, el entorno informático R y GNU-PSPP ha permitido realizar la totalidad de las pruebas estadísticas utilizadas con frecuencia en las investigaciones científicas. La tabla 2 muestra algunos paquetes y funcionalidades de R que han posibilitado la aplicación de las diferentes pruebas estadísticas abordadas con anterioridad.

Además de los paquetes recogidos en la tabla anterior, R ha posibilitado la creación, modificación y reutilización de sus librerías para efectuar el análisis de datos según las necesidades específicas de los investigadores de diferentes áreas del conocimiento.

En el texto Abbasnasab *et al.* [11] se realizó un análisis comparativo de la usabilidad y las funcionalidades referentes a la interacción hombre-computador que brindan SPSS y el lenguaje informático R a través de los paquetes *RStudio*, *R-Commander* y *jamovi*. Los resultados obtenidos evidenciaron que R mediante el paquete *jamovi*, ofrecía el mayor potencial para su utilización por parte de investigadores noveles de las ciencias sociales.

A su vez, en Sanoja *et. al.* [31] se evaluaron las soluciones comerciales SPSS, SAS, BMDP, STATGRAPHICS y STATISTIX respecto a las soluciones libres IDAMS y el entorno informático R. Estos autores identificaron la superioridad de las soluciones comerciales para la realización de análisis estadísticos respecto a las soluciones libres, fundamentalmente para la aplicación de pruebas no paramétricas.

El autor Salas [35] realizó una comparación de las características generales de 2 programas estadísticos comerciales de amplia utilización en ciencias ecológicas (SPSS y SAS) con el entorno informático R. Se evaluaron cualitativamente aspectos como la amigabilidad de la interfaz de usuario, capacidad para la manipulación de los datos, calidad de los gráficos, control de procesos, costos de licencias de uso, variedad de análisis estadístico de cada herramienta, soporte técnico, documentación disponible y compatibilidad con diferentes sistemas operativos. Se catalogó a R como un excelente programa estadístico para ser empleado en docencia e investigación. [35]

En su investigación, Nayak *et. al.* [17] presentaron algunos paquetes de R de utilidad en la evaluación psicológica, relacionados con la tecnología psicométrica bajo la teoría de respuesta a los *ítems* (TRI), análisis de correspondencias, desarrollo de modelos de ecuaciones estructurales, escalamiento multidimensional y teoría clásica de *tests* (TCT).

Tabla 1. Potencialidades de R, GNU/*Octave*, GNU-PSPP e IBM SPSS para el análisis estadístico en investigaciones científicas

Elemento a evaluar	Entorno informático R	GNU/Octave	GNU-PSPP	IBM SPSS
Compatibilidad	*Windows, Linux, Mac*	*Windows, Linux, Mac*	*Windows, Linux, Mac*	*Windows, Linux, Mac*
Tipo de licencia	libre	libre	libre	comercial
Formato datos de entrada	.txt, .doc, .csv, .xls, .sav, .sas, .dat, .xpt, .dta, .rda, .tab, .rec, .mtp	.dat, .xls, .txt, .csv, .xpt	.txt, .xls, .csv, .odt, .sav	.sav, .xls, .csv, .txt, .dat, .xpt, .tab, .dta, .sas
Interfaz de usuario	GUI (*RStudio, Deducer, WinEdit, Tinn-R, Emacs*) y Consola	GUI (*Qt4*) y Consola	GUI y Consola	GUI y Consola
Formato datos exportados	.txt, .xls, .dta, .xml, .html, .pdf, .sav	.xpt, .xls, .pdf, .html	.txt, .pdf, .html, .odt, .csv, .ps	.txt, .xls, .html, .pdf
Pruebas de normalidad de los datos				
Shapiro-Wilk	Sí	Sí	Sí	Sí
Kolmogorov-Smirnov	Sí	Sí	Sí	Sí
T de Kendall	Sí	Sí	Sí	Sí
Pruebas estadísticas paramétricas				
Correlación de Pearson	Sí	Sí	Sí	Sí
Análisis de varianza (ANOVA) de uno o varios factores	Sí	Sí	Sí	Sí
T de Student	Sí	Sí	Sí	Sí
F de Fisher	Sí	Sí	Sí	Sí
Pruebas estadísticas no paramétricas				
Correlación de Spearman	Sí	Sí	Sí	Sí
U de Mann-Whitney	Sí	Sí	Sí	Sí
Kruskal-Wallis	Sí	Sí	Sí	Sí
Prueba de rango de signos de Wilcoxon	Sí	Sí	Sí	Sí
Prueba de Friedman	Sí	Sí	Sí	Sí
Distribución X^2	Sí	Sí	Sí	Sí
Rachas de Wald-Wolfowitz	Sí	Sí	Sí	Sí
Jonckeere-Terpstra	Sí	No	Sí	Sí
Reacciones extremas de Moses	Sí	No	Sí	Sí
Prueba de los signos	Sí	Sí	Sí	Sí
Prueba de la mediana	Sí	Sí	Sí	Sí
Q de Cochran	Sí	Sí	Sí	Sí
Prueba de McNemar	Sí	Sí	Sí	Sí
Análisis de fiabilidad y consistencia interna				
Alfa de Cronbach		Sí	Sí	Sí
Elaboración de gráficos				
Barras	Sí	Sí	Sí	Sí

Líneas	Sí	Sí	No	Sí
Áreas	Sí	Sí	No	Sí
Circular	Sí	Sí	Sí	Sí
Dispersión	Sí	Sí	Sí	Sí
Histogramas	Sí	Sí	Sí	Sí
Dendogramas	Sí	Sí	No	Sí
Mapas de calor	Sí	Sí	No	Sí
Gráfico de burbujas	Sí	No	No	Sí
Nubes de palabras	Sí	No	No	No

Tabla 2. Algunos paquetes y funciones de R orientados al análisis estadístico

Prueba estadística	Paquete / comando
Shapiro-Wilk	rio / *shapiro.test*
Kolmogorov-Smirnov	rio / *ks.test*
Correlación de Pearson	stats / *cor.test* con la opción method= "pearson"
Correlación de Spearman	pspearman / *spearman.test*
T de Kendall	Kendall / *Kendall*
Análisis de varianza (ANOVA) de un factor	StatCharrms / *basicAnova*
Análisis de varianza (ANOVA) de varios factores	car / *Anova* con la opción Type= "II" ó "III"
T de Student	stats / *t.test*
F de Fisher	stats / *var.test*
U de Mann-Whitney	asht / *wmwTest*
Kruskal-Wallis	stats / *kruskall.test*
Prueba de rango de signos de Wilcoxon	MASS / *wilcox.test*
Prueba de Friedman	stats / *friedman.test*
Distribución X^2	rio / *chi.test*
Rachas de Wald-Wolfowitz	randtests / *runs.test*
Jonckeere-Terpstra	DescTools / *JonckheereTerpstraTest*
Reacciones extremas de Moses	DescTools / *MosesTest*
Prueba de los signos	BSDA / *SIGN.test*
Prueba de la mediana	RVAideMemoire / *mood.medtest*
Q de Cochran	DescTools / *CochranQTest*
Prueba de McNemar	exact2x2 / *mcnemar.exact*
Alfa de Cronbach	ltm / *cronbach.alpha*

Los autores Ozgur *et. al.* [36] realizaron un análisis de las potencialidades y aplicabilidad de las soluciones Excel, SPSS, SAS y R para el trabajo con el *big data*. Estos autores destacaron el funcionamiento adecuado de SPSS, SAS y R en proyectos donde se ha requerido del manejo de grandes cantidades de datos y denotaron las facilidades que ha brindado R por ser una herramienta libre.

En su estudio Welbers *et. al.* [37] identificaron las potencialidades de R como herramienta para el análisis computacional de textos y su valor como paquete estadístico aplicable en investigaciones relacionadas con las ciencias de la comunicación.

Se ilustró la transformación de la escala asociada a la TCT de un instrumento de investigación empleado para medir el escalamiento de los puntajes de una variable-producto, por otra representativa de TRI y se ejemplificó la aplicabilidad del entorno R para la realización de análisis estadísticos mediante ANOVA, la prueba de la mediana y para los análisis multivariados mediante la utilización de modelos jerárquicos lineales. Se demostró la aplicabilidad del entorno R para la realización

de gráficos de tipo nubes de palabras y mapas de calor con el objetivo de facilitar el análisis semántico de palabras clave en la conformación de un marco teórico-referencial tras la elaboración de un diseño teórico metodológico y para analizar la correlación bivariada existente entre determinadas variables de un estudio pedagógico. [13]

Se realizó un análisis comparativo de las herramientas SPSS, SAS, STATA y R en relación a sus capacidades para la manipulación de los datos, la realización de análisis estadísticos, la calidad de los gráficos que generan, la compatibilidad con diferentes sistemas operativos y su curva de aprendizaje. Se identificaron como principales fortalezas de SPSS y STATA sus capacidades para la realización de análisis estadísticos de manera rápida y sencilla y se determinó que las potencialidades del entorno informático R han radicado en su versatilidad para confeccionar los gráficos, licencias gratuitas y en la enorme comunidad de desarrolladores que han contribuido a su evolución. [38]

La herramienta R se empleó para calcular las puntuaciones de los factores que han integrado un cuestionario para medir la inteligencia emocional de las personas y analizar su correlación. Aunque el trabajo con R es un poco más complejo que con otras herramientas comerciales, esta solución ha facilitado la reproducibilidad de las investigaciones y ha permitido exportar los resultados estadísticos obtenidos a disímiles formatos. [22]

En contraposición a lo planteado por los autores Sanoja *et. al.* [31] y López [38], los resultados obtenidos en la presente investigación demostraron las capacidades que brinda el entorno informático R para el análisis estadístico, el tratamiento y la representación gráfica de los datos, incluso frente a soluciones privativas bien establecidas como SPSS. Esto se debe al desarrollo y nivel de madurez alcanzado por esta solución.

Conclusiones

En la investigación se identificaron las principales herramientas de análisis estadístico empleadas por académicos cubanos durante el período 2015-2020 en investigaciones publicadas en revistas del catálogo *Scielo* Cuba. Se evidenció que la solución comercial IBM SPSS fue la más utilizada y que el uso de aplicaciones de código abierto fue prácticamente nulo durante el quinquenio analizado.

Se realizó un análisis comparativo entre diferentes paquetes estadísticos de código abierto y la herramienta IBM SPSS. Se identificó que GNU-*Octave*, GNU-PSPP y el entorno R poseían diversas funcionalidades que posibilitan el procesamiento y la analítica de datos. Sin embargo, se evidenció que el entorno informático R ha brindado marcadas capacidades para el procesamiento, análisis estadístico y representación gráfica de los datos en investigaciones científicas, superior a GNU-*Octave*, GNU-PSPP e incluso al líder privativo IBM SPSS. Además, se comprobó en las fuentes bibliográficas consultadas la aplicabilidad del entorno informático R para el análisis estadístico en investigaciones de diferentes áreas del conocimiento.

El presente estudio ha contribuido a solventar las limitantes legales y económicas que han atentado contra el desarrollo sostenible y soberano de la ciencia en países del tercer mundo. El empleo de soluciones libres y de código abierto para el análisis estadístico en investigaciones científicas en Cuba deberá favorecer el desarrollo de estos estudios en el sentido de facilitar su reproducibilidad por autores y evaluadores externos.

En aras de potenciar la utilización de paquetes estadísticos libres en la práctica científica cubana se propuso incluir el trabajo con estas herramientas en los planes de estudio de la asignatura Estadística en la enseñanza universitaria y en los cursos de Metodología de la investigación científica impartidos en el posgrado.

REFERENCIAS BIBLIOGRÁFICAS

1. Azman MH, Puteh F. Quantitative Data Analysis: Choosing Between SPSS, PLS and AMOS in Social Science Research. Int Interdiscip J Sci Res [Internet]. 2017;3(1):14-25. Disponible en: https://iijsr.org/data/frontImages/gallery/Vol._3_No._1/3._14-25.pdf
2. Pérez Grenier O, Collazo Acosta E. Estadística inferencial en la actividad científica de la residencia de Medicina General Integral en Artemisa. Revista Cubana de Medicina General Integral [Internet]. 2017 [citado 11/11/2021]; 33(3):331-341. Disponible en: http://www.revmgi.sld.cu/index.php/mgi/article/view/347
3. Enrique FM, Peña M. Improcedencias al usar la estadística en las investigaciones sociales. Varona Rev Científico Método [Internet]. 2020 [citado 25/06/2021]; 70:13-8. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1992-82382020000100013&lng=es&nrm=iso&tlng=en
4. De Oliveira Neto FG, Torkar R, Feldt R, Gren L, Furia CA, Huang Z. Evolution of statistical analysis in empirical software engineering research: Current state and steps forward. Journal of Systems and Software [Internet]. 2019 [citado 14/11/2021]; 156:246-67. DOI: https://doi.org/10.1016/j.jss.2019.07.002
5. Avello R, Seisdedo A. El procesamiento estadístico con R en la investigación científica. MediSur [Internet]. 2017 [citado 25/06/2021];15(5):583-6. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1727-897X2017000500001&lng=es&nrm=iso&tlng=es
6. Rodrigues C F, Camello F J, Timbó F. Importance of using basic statistics adequately in clinical research. Brazilian Journal of Anesthesiology [Internet]. 2017 [citado 11/11/2021]; 67(6):619-25. DOI: https://doi.org/10.1016/j.bjane.2017.01.011
7. Weihs C, Ickstadt K. Data Science: the impact of statistics. International Journal of Data Science and Analytics [Internet]. 2018 [citado 11/11/2021]; 6,189-94. DOI: https://doi.org/10.1007/s41060-018-0102-5
8. Masuadi E, Mohamud M, Almutairi M, Alsunaidi A, Alswayed AK, Aldhafeeri OF. Trends in the Usage of Statistical Software

and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis. Cureus [Internet]. 2021 [citado 25/06/2021];13(1). DOI: https://doi.org/10.7759/cureus.12639
9. Companioni A, Cuesta E, Hernández Y, Orovio VO, Días S. Entorno integrado para el trabajo con GNU/Octave. Revista Cubana de Ciencias Informáticas [Internet]. 2012;6(4):9. Disponible en: http://www.redalyc.org/articulo.oa?id=378343677001
10. Kramer MH, Paparozzi ET, Stroup WW. Best Practices for Presenting Statistical Information in a Research Article. HortScience [Internet]. 2019 [citado 25/06/2021];54(9):1605-9. DOI: https://doi.org/10.21273/HORTSCI13952-19
11. Abbasnasab S, Brown GTL, Denny P. Comparing four contemporary statistical software tools for introductory data science and statistics in the social sciences. Teaching Statistics [Internet]. 2021 [citado 12/11/2021]; 43(S1): 157-72. DOI: https://doi.org/10.1111/test.12274
12. Sued GE. Repertorio de técnicas digitales para la investigación con contenidos generados en redes sociodigitales. PAAKAT Revista de Tecnología y Sociedad [Internet]. 2020 [citado 25/06/2021];10(19). DOI: https://doi.org/10.32870/Pk.a10n19.498
13. Torres PA. Lo que todo investigador educativo cubano debiera conocer: el entorno informático R. Atenas [Internet]. 2018 [citado 25/06/2021];4(44): 1-27. Disponible en: https://www.redalyc.org/jatsRepo/4780/478055154001/index.html
14. Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen J, Ly A, Gronau Q F, Šmíra M, Epskamp S, Matzke D, Wild D, Knight P, Rouder JN, Morey RD, and Wagenmakers EJ. JASP: Graphical statistical software for common statistical designs. Journal of Statistical Software [Internet]. 2019 [citado 12/11/2021]; 88(2):1-17. DOI: https://doi.org/10.18637/jss.v088.i02
15. Alvarenga H, Sampaio A. Teaching Introductory Statistical Classes in Medical Schools Using RStudio and R Statistical Language: Evaluating Technology Acceptance and Change in Attitude Toward Statistics. J Stat Educ [Internet]. 2020 [citado 25/06/2021];28(2):212-9. DOI: https://doi.org/10.1080/10691898.2020.1773354
16. Ruiz Ruano AM, Puga JL. R como entorno para el análisis estadístico en evaluación psicológica. Papeles del Psicólogo [Internet]. 2016; 37(1):74-9. Disponible en: http://www.redalyc.org/articulo.oa?id=77844204010
17. Nayak P, Mukherjee AK, Pandit E, Pradhan SK. Application of Statistical Tools for Data Analysis and Interpretation in Rice Plant Pathology. Rice Science [Internet]. 2018 [citado 13/11/2021]; 25(1), 1-18. DOI: https://doi.org/10.1016/j.rsci.2017.07.001
18. Santin D, Caregnato S. Participación del Caribe en la ciencia regional: una mirada general y un análisis comparado de Cuba, Jamaica y Trinidad y Tobago. Revista Cubana de Información en Ciencias de la Salud [Internet]. 2020 [citado 10/11/2021]; 31(4) Disponible en: http://www.rcics.sld.cu/index.php/acimed/article/view/1605
19. Galbán-Rodríguez E, Torres-Ponjuán D, Martí-Lahera Y, Arencibia-Jorge R. Measuring the Cuban scientific output in scholarly journals through a comprehensive coverage approach. Scientometrics [Internet]. 2019 [citado 10/11/2021]; 121(2): 1019–1043. DOI: https://doi.org/10.1007/s11192-019-03233-6
20. Çetinkaya-Rundel M, Rundel C. Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum. The American Statistician [Internet]. 2017 [citado 10/11/2021]; 72(1): 58–65. DOI: https://doi.org/10.1080/00031305.2017.1397549
21. Niu G, Segall RS, Zhao Z, Wu Z. A Survey of Open Source Statistical Software (OSSS) and Their Data Processing Functionalities. International Journal of Open Source Software and Processes [Internet]. 2021 [citado 16/11/2021]; 12(1): 1-20. DOI: https://doi.org/10.4018/IJOSSP.2021010101
22. Stiglic G, Watson R, Cilar L. R you ready? Using the R programme for statistical analysis and graphics. Res Nurs Health [Internet]. 2019 [citado 25/06/2021];42(6):494-9. DOI: https://doi.org/10.1002/nur.21990
23. About [Internet]. GNU Octave. 2021 [citado 25/06/2021]. Disponible en: https://www.gnu.org/software/octave/about
24. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes & Diseases [Internet]. 2017 [citado 14/11/2021];4(3): 138-48. DOI: https://doi.org/10.1016/j.gendis.2017.06.001
25. Piepho HP, Edmondson RN. A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. Journal of Agronomy and Crop Science [Internet]. 2018 [citado 14/11/2021]; 204: 429-55. DOI: https://doi.org/10.1111/jac.12267
26. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics [Internet]. 2017 [citado 14/11/2021];11(4): 959-75. DOI: https://doi.org/10.1016/j.joi.2017.08.007
27. Welbers K, Van Atteveldt W, Benoit K. Text Analysis in R. Communication Methods and Measures [Internet]. 2017 [citado 14/11/2021];11(4): 245-65. DOI: https://doi.org/10.1080/19312458.2017.1387238
28. Madeyski L, Kitchenham B. Would Wider Adoption of Reproducible Research Be Beneficial for Empirical Software Engineering Research? Journal of Intelligent & Fuzzy Systems [Internet]. 2017 [citado 16/11/2021]; 32(2): 1509-21. DOI: https://doi.org/10.3233/JIFS-169146
29. Pajankar A, Chandu S. GNU Octave by Example. A Fast and Practical Approach to Learning GNU Octave. APRESS; 2020. 179p. DOI: https://doi.org/10.1007/978-1-4842-6086-9
30. De Oliveira SB, da Motta RA, da Costa SR, Calvosa MV, de Oliveira AS, Romanini D. Em busca de um Software de Apoio a Pesquisas Qualitativas: o caso de uma unidade de ensino e pesquisa de uma universidade pública brasileira. Revista Ibérica de Sistemas e Tecnologias de Informação [Internet]. 2021 [citado 16/11/2021]; E41: 164-77. Disponible en: https://www.proquest.com/openview/4fc3316c510c3b226fcfbb9b6264ce46/1?pq-origsite=gscholar&cbl=1006393
31. Sanoja J, Ortiz J. Paquetes tecnológicos para el tratamiento de datos en investigación en educación matemática. Paradígma [Internet]. 2007 [citado 25/06/2021];28(1):215-34. Disponible en: http://ve.scielo.org/scielo.php?script=sci_abstract&pid=S1011-22512007000100011&lng=es&nrm=iso&tlng=es
32. Flores Ruiz E, Miranda-Novales MG, Villasís-Keever MÁ, Flores-Ruiz E, Miranda Novales MG, Villasís Keever MÁ. El protocolo de investigación VI: cómo elegir la prueba estadística adecuada. Estadística inferencial. Rev Alerg México [Internet]. 2017 [citado 25/06/2021];64(3):364-70. DOI: https://doi.org/10.29262/ram.v64i3.304
33. Hernández R, Hernández C, Baptista P. Metodología de la investigación. 6.a ed. México D. F.: Mc Graw Hill Education; 2014. 634 p.

34. Sieger T, Hurley CB, Fišer K, Beleites C. Interactive Dendrograms: The R Packages idendro and idendr0. Journal of Statistical Software [Internet]. 2017 [citado 14/11/2021];76(10): 1-22. DOI: https://doi.org/10.18637/jss.v076.i10
35. Salas C. ¿Por qué comprar un programa estadístico si existe R? Ecol Austral [Internet]. 2008 [citado 25/06/2021];18(2):223-31. Disponible en: http://ojs.ecologiaaustral.com.ar/index.php/Ecologia_Austral/article/view/1389
36. Ozgur C, Dou M, Li Y, Rogers G. Selection of Statistical Software for Data Scientists and Teachers. Journal of Modern Applied Statistical Methods [Internet]. 2017;16(1). DOI: https://doi.org/10.22237/jmasm/1493599200
37. Welbers K, Atteveldt WV, Benoit K. Text Analysis in R. Communication Methods and Measures [Internet]. 2017 [citado 25/06/2021];11(4):245-65. DOI: https://doi.org/10.1080/19312458.2017.1387238
38. López MA. Uso de TIC en la investigación: Herramientas informáticas para la recolección y análisis de la información. En: Situaciones y Retos de la investigación en Latinoamérica [Internet]. Fondo Editorial Universidad Católica Luis Amigó; 2018. p. 124-40. DOI: https://doi.org/10.21501/9789588943381

Recibido: 02/07/2021
Aprobado: 24/04/2022

Conflicto de intereses
Los autores declaran que no existen conflictos de interés

Financiación
Los autores declaran que no recibieron financiación para desarrollar la investigación.

Cómo citar este artículo
Perdigón-Llanes R, Pérez-Pino MT. Herramientas de código abierto para el análisis estadístico en investigaciones científicas. AnAcadCienc Cuba [internet] 2022 [citado en día, mes y año]; 12(3): e1120. Disponible en: http://www.revistaccuba.cu/index.php/revacc/article/view/1120.