# Si una imagen vale más que mil palabras: ¿cuánto puede decir un gráfico de cajas?

## If an image is worth than thousand words: how much a box plot can say?

Dennis Denis Ávila[1,*] y Víctor Manuel Ramírez-Arrieta[2]

[1]Facultad de Biología, Universidad de La Habana, Calle 25, Nº 455, e/ J e I, Vedado, Plaza de la Revolución, La Habana, Cuba. CP. 10400. [2]Instituto de Ciencias del Mar (ICIMAR), Calle Loma, Nº 14, e/ 35 y 37, Alturas del Vedado, Plaza de la Revolución, La Habana, Cuba. CP. 11600.
*Autor para correspondencia (e-mail: dda@fbio.uh.cu).

**RESUMEN**
Las visualizaciones gráficas de datos son una parte fundamental del Análisis Exploratorio de Datos y preceden al análisis estadístico. Los gráficos de cajas son de los métodos gráficos más ampliamente utilizados para representar los estadísticos descriptivos de una muestra y visualizar comparaciones. Sin embargo, si estos no son aplicados correctamente, pueden distorsionar marcadamente las interpretaciones de los datos. Por esta razón, en el presente trabajo se describen los problemas asociados a estos gráficos y se mencionan varias maneras en que pueden ser mejorados, con el uso de la flexibilidad que ofrece el entorno *R* de programación. A través de la adición de dispersiones de puntos, histogramas y curvas de densidad de distribuciones se obtienen variantes gráficas que han sido descritas en la literatura reciente con nombres como gráficos de violín, de pirata, de nubes y lluvia, de vainas, *sinaplots*, entre otros. Se presenta, además, la aplicación *Extended Boxplot Graphics* que permite utilizar las potencialidades gráficas del entorno *R*, sin necesidad de dominar su programación, para producir de forma sencilla variantes mejoradas de los gráficos de cajas para la representación de datos científicos.

**Palabras clave**: exploración de datos, figuras científicas, gráficos de cajas y bigotes, lenguaje *R*

**ABSTRACT**
Graphic data visualization is essential to Exploratory Data Analysis and precedes any statistical analysis. Box plots are one of the most widely used graphical representation to show descriptive statistics of data in a sample and to represent multiple comparisons. This graphic, however, if careless used can markedly distort data interpretations, and for this reason in this paper we describe box plot pitfalls and show several ways to improve them using *R* programming flexibility. By adding jittered raw data, histograms and density curves graphic variants can be produced, which had been described in recent literature with other names such as violin plots, pirate plots, raincloud plots, beanplots, sinaplots, among others. We also present the app *Extended Boxplot Graphics* that facilitates the use of potent graphic flexibilities of *R* without mastering the programming language. With this app in a simple way several enhanced variations of box plots can be produced to represent scientific results.

**Keywords**: exploratory data analysis, scientific figures, box and whisper plot, *R* language

**Citación**: Denis, D. & Ramírez-Arrieta, V.M. 2020. Si una imagen vale más que mil palabras: ¿cuánto puede decir un gráfico de cajas? *Revista Jard. Bot. Nac. Univ. Habana* 41: 57-69.

**Recibido**: 3 de junio de 2020. **Aceptado**: 22 de junio de 2020. **Publicado en línea**: 26 de septiembre de 2020. **Editor encargado:** José Angel García-Beltrán.

### INTRODUCCIÓN

La revista americana *LIFE* tenía una consigna que se hizo universal: "Una imagen vale más que mil palabras". La idea subyacente es que las imágenes contienen una alta densidad de información, que es procesada visualmente con mayor rapidez que la leída o escuchada. Los pasos fundamentales de la investigación científica diaria incluyen: procesar datos, detectar en ellos cosas inusuales y buscar vínculos o relaciones aparentes. Inicialmente, todo ello se hace por medio de visualizaciones que dan poco peso a la aleatoriedad, a ecuaciones de modelos estocásticos o a parámetros poblacionales, en lo que se llama el Análisis Exploratorio de los Datos (AED) (Hoaglin & *al.* 1983, 1985). Esta es una rama del análisis de datos muy importante que, por ser relativamente joven, ha sido tradicionalmente omitida en los libros y cursos de Estadística, los que tienden a presentar una división dicotómica de dicha disciplina: descriptiva e inferencial.

El AED fue desarrollado formalmente por Tukey (1977) y describe el acto de observar los datos y buscar que aparentan decir, actividad donde las visualizaciones gráficas ocupan un rol de primer orden. Las figuras permiten usar las ventajas de la interpretación visual para comprender el comportamiento de los datos, incluso los más complejos, y ahorrar tiempo al analista, a la vez que permiten crear hipótesis sobre la variabilidad, escala, patrones y tendencias en estos (DuToit & *al.* 1986). En el campo científico, el análisis gráfico de datos constituye una de las ramas de mayor importancia al preceder a la aplicación de los demás procesamientos estadísticos. Además, en las publicaciones se reconocen a las tablas y figuras como uno de sus componentes más importantes (Hubbard & Dunbar 2017), ya que acaparan la mayoría de las veces el foco de atención.

En la actualidad, en la era del *big-data*, de las múltiples "-ómicas" en la Biología y de las ciencias computacionales integrativas, la necesidad de métodos sencillos y claros para visualizar datos se hace cada vez más importante, para poder comprender los patrones subyacentes en ellos (Sidiropoulos & *al.* 2018). Si las figuras son creadas de forma efectiva y soportadas con un

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

pie de figura bien articulado, portan información compleja y suficiente que puede permitir llegar rápidamente a conclusiones, incluso sin leer los detalles de la narrativa (Hubbard & Dunbar 2017). Por ello, el campo del AED tiene gran influencia e incluye un fuerte desarrollo de los métodos gráficos, la minería de datos, el aprendizaje de máquina y otros métodos modernos (Morgenthaler 2009) que se han integrado en la llamada Ciencia de Datos (*Data Science*). Los modelos matemáticos y estadísticos más formales quedan para una fase más adelantada del análisis, que sería la estadística tradicionalmente empleada.

Durante el AED la visualización de los datos primarios es una herramienta esencial, y se refiere a cualquier forma de presentarlos: tablas, figuras, mapas o esquemas. Por la importancia de este paso, se han realizado múltiples esfuerzos en la búsqueda de nuevas formas de representar los datos (Buja & *al.* 1996). Estas representaciones se han llamado tradicionalmente "figuras estadísticas", aunque muchas veces presentan solamente los datos primarios y no contienen aspectos del campo de la estadística formal. Los métodos de visualización complementan los resúmenes numéricos ya que permiten detectar características que se pueden perder durante la descripción estadística.

Las tendencias actuales muestran cambios fundamentales en la forma en que se presentan las visualizaciones de datos. En la época de las publicaciones impresas se asumieron reglas estilísticas asociadas a las limitaciones logísticas, como el empleo de figuras en blanco y negro, y más utilización de líneas y puntos que de áreas. Tufte (1983) presentó un índice para medir la cantidad de información irrelevante en un gráfico, conocido como Razón Datos - Tinta = Tinta de los datos / Total de tinta del gráfico. Este se puede interpretar como el porcentaje de tinta del gráfico que no puede ser eliminado sin afectar la comunicación. Ello fue criticado por Wainer (1990), quién defendió no solo la eficiencia en la transmisión de los datos, sino también la elegancia y estética. Actualmente, al desaparecer las limitaciones de impresión y con las ventajas del entorno digital, se retoman los colores, íconos y la combinación de figuras de datos con esquemas, fotografías y dibujos en una nueva reestructuración de criterios estéticos que afecta a todos los tipos de representaciones, incluidas las del AED.

Uno de los métodos gráficos más ampliamente utilizados para el AED son los gráficos de cajas o *box plot* (también llamados de cajas y bigotes, *box and whisker plots*), introducidos por el propio Tukey (1977), aunque la descripción de un gráfico similar ya había sido dada por Spear (1952). En su forma básica, son una representación del llamado "resumen de cinco números", que contiene las propiedades principales que describen a un conjunto de datos: tendencia central, dispersión y valores extremos. Desde su aparición, muchas modificaciones fueron sugeridas (*e.g.*, McGill & *al.* 1978, Velleman & Hoaglin 1981, Chambers & *al.* 1983, Frigge & *al.* 1989) y se ha mantenido como un buen ejemplo de procedimiento exploratorio: versátil, comunicativo y fácil de hacer. Sin embargo, es un tipo de figura que clasificaría como "débilmente buena" según el criterio de Wainer (1990), ya que no es totalmente comprensible de forma instintiva y sin conocimientos estadísticos previos, es decir, requiere de cierto nivel de entrenamiento, explicación o ayuda adicional para interpretarlos. En este sentido, se evidencia lo que decía John W. Tukey: *"A picture may be worth a thousand words but it may take a hundred words to do it"* [un gráfico puede valer mil palabras, pero puede tomar cientos de palabras para lograrse] (Tukey 1986).

A pesar de la popularidad de los diagramas de cajas, no todos los investigadores que los emplean están conscientes de los fallos que pueden tener, al ser formas de representación que, en ocasiones, distorsionan la información de maneras muy marcadas. Por esta razón, en la actualidad se han comenzado a hacer populares diversas formas de mejorarlos a través de la inclusión o modificación de elementos complementarios como la dispersión de los datos primarios o combinaciones con histogramas o estimados de distribución de densidades. Estas formas derivadas del gráfico de cajas primario han recibido otros nombres (en el idioma inglés en el cual fueron descritos) como *violin plots* (Hintze & Nelson 1998), *beanplots* (Kampstra 2008), *raincloud plots* (Allen & *al*. 2018), *pirate plots* (Phillips 2016), *sinaplots* (Sidiropoulos & *al.* 2018), entre otros.

En la presente comunicación se describen los problemas asociados a los gráficos estadísticos de cajas y varias de las maneras que pueden ser mejorados con el uso de la flexibilidad que ofrece el entorno *R* de programación, con lo cual su interpretación puede hacerse más informativa y exacta. Como producto adicional, se emplea este análisis para presentar un código de *R* llamado *Extended Boxplot Graphics,* que permite a los investigadores con poco conocimiento de este entorno de programación, utilizar de forma sencilla todas las potencialidades de tal lenguaje en la producción de variantes mejoradas de este tipo de representación de datos científicos.

## ANATOMÍA DE LOS GRÁFICOS DE CAJAS

Al representar este tipo de gráfico, en su versión básica, se proyectan en un sistema de coordenadas los estadísticos de resumen de un conjunto de datos (Benjamini 1988). Inicialmente se desarrolló con el llamado resumen de cinco números: una secuencia que describe completamente un conjunto de datos (Figura 1): los dos valores extremos, los cuartiles y la mediana. Estrictamente, en sus primeras versiones en lugar de los cuartiles se utilizaron los puntos de inflexión o *hinges*, que no son más que las medianas de cada mitad de los datos, con la inclusión de la mediana global. Algunas aplicaciones actuales, como *R*, todavía usan los *hinges* en lugar de los cuartiles (para los cuales aún no existe un método universal de obtención o cálculo), lo cual hace que en algunas ocasiones puede variar la posición de la caja, aunque la mayoría de las veces coincide entre métodos (Krzywinski & Altman 2014). La distancia entre cuartiles (o la separación entre *hinges* conocida como *H-spread*) es indicadora de la variabilidad de la muestra. Por su definición, estos números dividen la muestra en cuatro intervalos, entre los cuales los datos se distribuyen de forma relativamente equitativa.

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

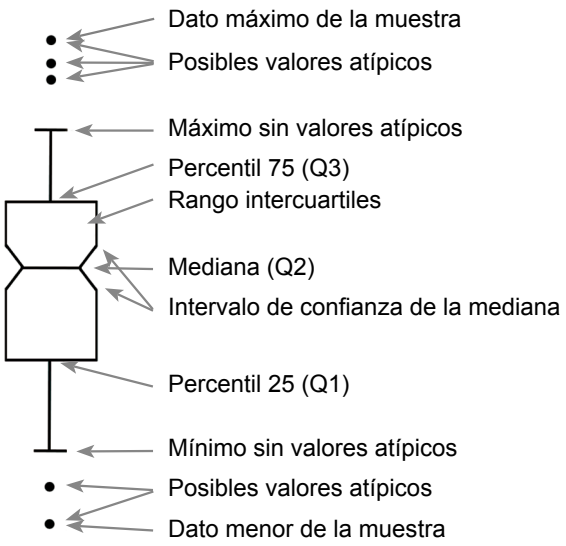


**Fig. 1.** Composición básica de un diagrama de cajas en su concepción original (estilo de Tukey). Q: cuartiles.

***Fig. 1.*** *Basic composition of a box plot in its original conception (Tukey´s style). Q: quantile.*

Entre las variaciones básicas está el uso de los datos extremos para la barra central (estilo de Spear) o la sustitución de estos extremos por el rango sin valores atípicos (estilo de Tukey). Estos son identificados por estar fuera del intervalo cuartil ± ½ (rango intercuartil) y son marcados como valores extremos. Por su ubicación relativa pueden diferenciarse en valores adyacentes o valores atípicos (Cleveland 1993). La caja puede incluir una muesca (*notch*) que indica las zonas de máxima probabilidad inferencial. Los estadísticos representados en cada elemento del gráfico de cajas pueden ser también paramétricos: media, desviaciones estándares, errores estándares e intervalos de confianza que, al asumir implícitamente una distribución normal, generan un gráfico simétrico.

Los gráficos de cajas, aunque pueden emplearse en muestras muy pequeñas, se sugiere que nunca sean menores de cinco muestras e indicar siempre el tamaño de estas. También se deben evitar las muescas de inferencia en la caja (intervalos de confianza de la mediana), a menos que la muestra sea suficiente y estas estén completamente contenidas en el rango intercuartiles (Krzywinski & Altman 2014).

## DESVENTAJAS DE LOS GRÁFICOS DE CAJAS

Un principio básico para la presentación de los datos, dado por Tufte (1983), es que no se distorsione la información, o sea, que la interpretación de la representación visual sea consistente con la interpretación de la información numérica. Varias de las figuras más conocidas han sido criticadas por este efecto. Por ejemplo, en la Figura 2A se muestra la comparación de valores entre dos localidades en una variable X, para demostrar la distorsión que puede aparecer través de un gráfico de barras. Las grandes diferencias que se observan entre las barras son un reflejo distorsionado de las diferencias en las distribuciones de datos que, a través de los estadísticos descriptivos básicos, se pueden notar.

A pesar de ser muy utilizados, se ha recomendado fuertemente evitar el uso de gráficos de barras, incluso con líneas de error (Weissgerber & *al.* 2015). Estas figuras codifican la magnitud de los valores por la longitud de las barras y puede ser una representación muy precisa para datos de conteos, proporciones o porcentajes, pero no tanto así para promedios. En este último caso, las barras también producen la percepción de que las medias se relacionan a la altura más que a la posición del extremo superior. Al originarse siempre en cero, el rango de recorrido de las barras cubre zonas de valores del eje Y que nunca aparecieron en los datos, lo cual es contradictorio (Streit & Gehlenborg 2014). Además, el empleo de distintas líneas base (es decir, cuando los ejes no parten del origen de coordenadas) puede interferir en la evaluación visual de las magnitudes de error. También en sus formas típicas estas barras muestran solo una rama del error (la externa) y la evaluación de las superposiciones para comparar muestras se hace difícil. El empleo de cortes de ejes o escalas no lineales dificulta aún más su comprensión (Krzywinski & Altman 2014).

En el caso de la Figura 2B, al compararse el sitio A con el B, puede verse como las “grandes diferencias” que las barras indicaban son dadas tan solo por unos pocos puntos que distorsionan el cálculo de la media como tendencia central paramétrica. Sin tener en cuenta estos valores, las distribuciones en ambas localidades muestran cierto grado de superposición. Pero nuevamente, esta mayor similitud puede ser modificada, si en lugar de estadísticos descriptivos no paramétricos como la mediana y el rango intercuartiles se empleasen la media, el error estándar y el intervalo de confianza al 95 %, como en la Figura 2C. Véase que en este caso las cajas y bigotes se separan completamente y en la primera localidad son tan estrechos que muchos de los puntos de los datos quedan por fuera de ellos. Esto refleja el efecto del tamaño de muestra en la precisión aparente de esta representación, una de las razones por las cuales, indicar el tamaño de muestra es fundamental para una interpretación adecuada de un gráfico de cajas (al igual que en otras figuras y tablas).

Cuando en las muestras aparecen datos con amplios rangos de dispersión, se suelen utilizar transformaciones del eje, como la escala logarítmica (Figura 2D). Véase que en esta figura las cajas se distienden y son más fácilmente visibles, de modo que las diferencias se hacen aparentemente menores. Sin embargo, solo es un efecto visual, dado que los datos son los mismos. Esto ejemplifica cómo un cambio de la escala del eje de estos gráficos puede manipular las percepciones. Un cambio mucho más radical se hace cuando la transformación no se hace a la escala del eje sino a los propios datos (Figura 2E). La transformación logarítmica de los datos previa a su representación en un diagrama de cajas resulta en una figura mucho más simétrica y compacta, pero se dificulta fuertemente su interpretación, ya que los estadísticos descriptivos sobre los datos transformados ya no tienen las mismas propiedades que sobre los datos originales. Así por ejemplo, un intervalo de la media más menos la desviación

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

estándar, aunque es simétrica ya no contiene la misma cantidad de valores originales dentro, debido a que las distancias de cada dato a la media se modifican de manera diferente en la mitad superior que en la inferior. Es decir, el intervalo aparentemente simétrico con los datos transformados no se interpreta igual sobre los datos originales. Por ello, cuando se realizan transformaciones de escala, se recomienda la representación de los datos sin transformar para una mejor interpretación de los resultados (Pérez 2018).

Todos los estadísticos descriptivos conllevan resumir la información y, por consiguiente, perder detalles que pudieran ser importantes. En este sentido, la pérdida más fuerte asociada a los gráficos de cajas está en que no brindan información sobre la distribución de los datos que subyacen a la inferencia de los estadísticos que representa. En la Figura 3 se demuestra este problema a través de otro ejemplo: la comparación de una variable entre tres localidades. Una inspección superficial de la Figura 3A sugiere diferencias "evidentes" que, en realidad

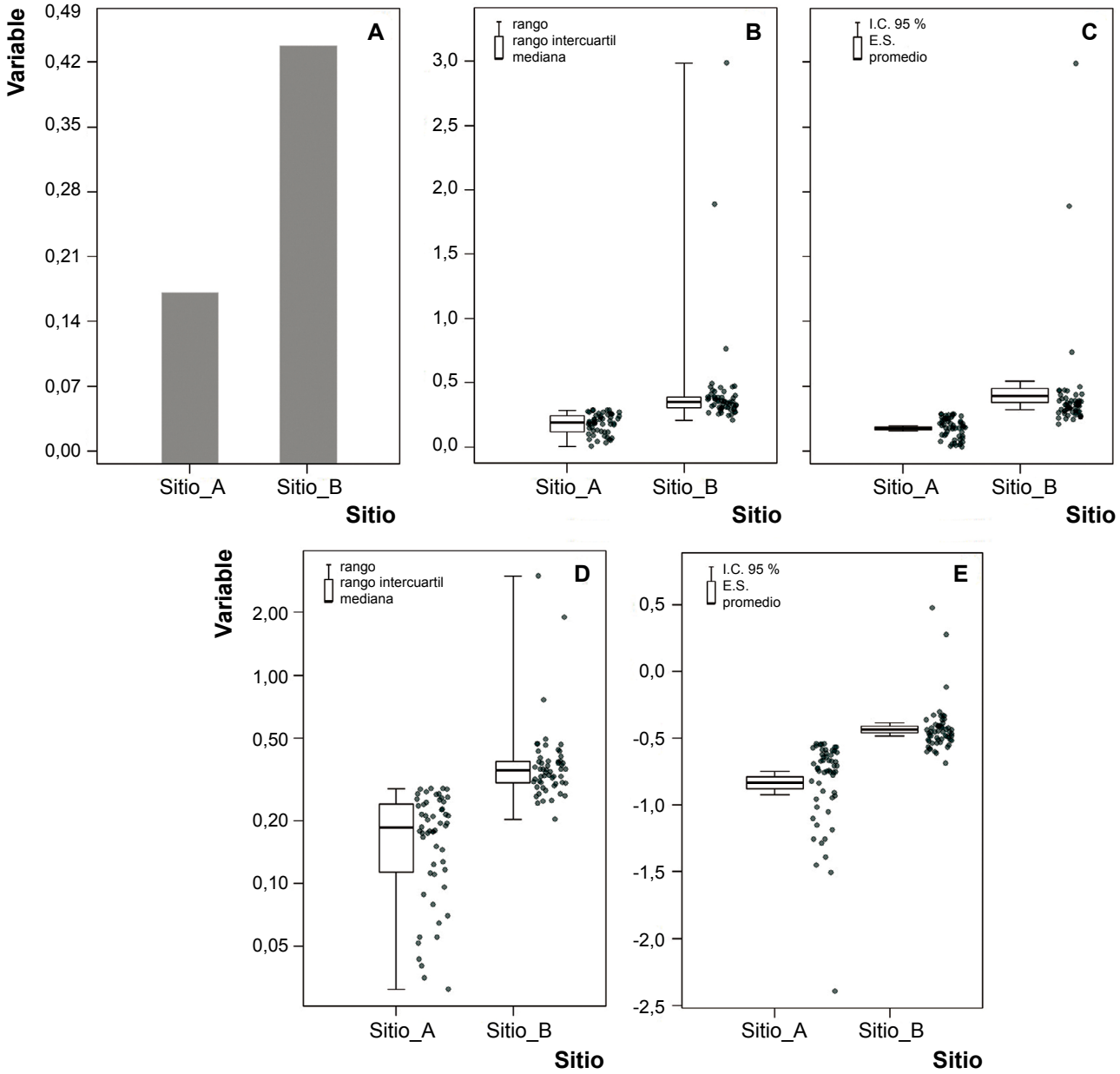


**Fig. 2.** Representaciones gráficas de la comparación entre dos sitios, descritos con un mismo conjunto simulado de datos, para evidenciar las potenciales distorsiones en las interpretaciones que son sugeridas por las distintas formas de representación. **A**. Gráfico de barras de las medias. **B**. Gráfico de cajas con descriptivos no paramétricos. **C**. Gráfico de cajas con descriptivos paramétricos. **D**. Gráfico de cajas con eje en escala logarítmica. **E**. Gráfico de cajas con datos transformados a logaritmos.

***Fig. 2.*** *Graphical representation of a comparison among two sites using the same simulated datasets to demonstrate potential for distorted interpretations suggested by data graphic visualization.* ***A.*** *Bar plot.* ***B.*** *Box plot with non-parametric descriptors.* ***C.*** *Box plot with parametric descriptor.* ***D.*** *Box plot with logarithmic scale axis.* ***E.*** *Box plot with logarithm transformed data.*

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

son falsas, lo cual se demuestra al comparar los histogramas de frecuencia de los datos (Figura 3B). Los sitios A y B tienen una fuerte superposición entre los datos que ponen en cuestionamiento las diferencias. Nuevamente, el principal efecto está dado por las distorsiones que las diferencias en tamaño de muestra pueden hacer sobre los estadísticos paramétricos.

La situación contraria también es posible (Figura 3C). Un diagrama de cajas que sugiere una apariencia de igualdad muy fuerte entre muestras, enmascara profundas diferencias en las distribuciones de los datos (Figura 3C). Nótese que, en este ejemplo, la distribución de datos del sitio A es asimétrica la izquierda y en el sitio C es bimodal (Figura 3D).

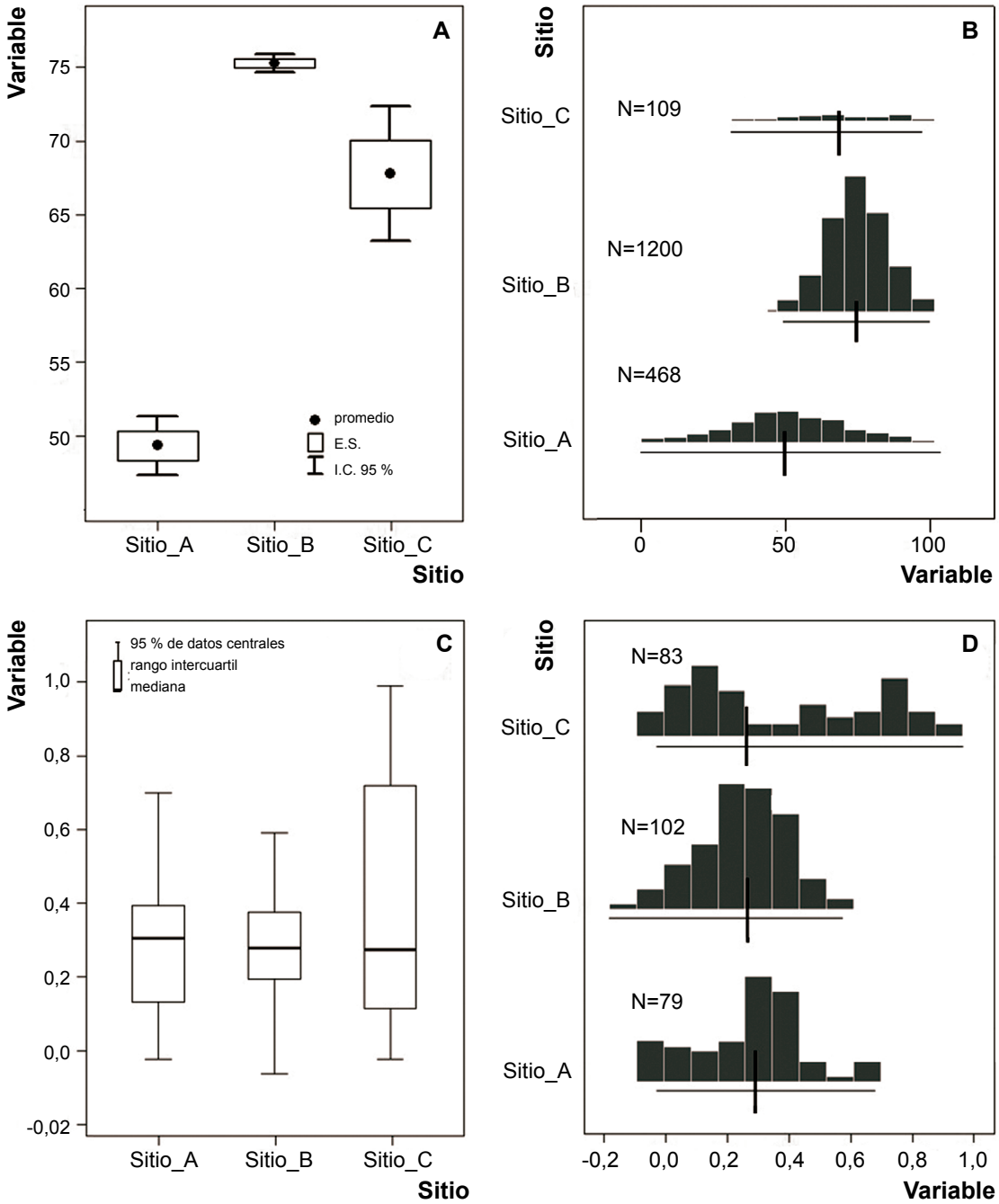


**Fig. 3.** Representaciones que demuestran las desventajas de los gráficos de cajas al ocultar la forma de la distribución de los datos y al emplear estadísticos paramétricos que asumen implícitamente una distribución probabilística normal. A y B son representaciones de los mismos conjuntos de datos, al igual que C y D. El gráfico A sugiere diferencias no reales y C sugiere similitudes tampoco reales, como demuestran los histogramas B y D, respectivamente. La línea horizontal bajo los histogramas representa el rango y la vertical, la media de los datos.

***Fig. 3.*** *Graphics demonstrating pitfalls of box plots because of hiding data distributions and due to the inappropriate use of parametric descriptive statistics that implicitly assume a normal distribution. A and B are representations of the same dataset, as well as C and D. Plot A suggest non real differences and plot C suggest unreal similarities, as shown by B and D histograms. Lines under histograms represent mean and data range.*

## VARIANTES O MEJORAS ACTUALES A LOS GRÁFICOS DE CAJAS

Para vencer las deficiencias de los gráficos de cajas tradicionales se han creado numerosas variantes, sobre todo, facilitadas por las flexibilidades gráficas que ofrece el entorno de programación *R*. Las variantes de mejoras incluyen la modificación o adición de elementos que pueden ser clasificados en tres grupos: estadísticos, datos crudos, o representaciones de distribuciones de frecuencia o probabilidad.

Primeramente, la selección de los estadísticos representados en los componentes del gráfico debe hacerse cuidadosamente, por las distorsiones mencionadas previamente. Con propósitos exploratorios la mediana es más recomendable que la media como tendencia central, ya que no está asociada a la distribución normal y es muy resistente a los valores atípicos. Los intervalos de confianza paramétricos pueden ser sustituidos por intervalos no paramétricos, obtenidos por métodos de Montecarlo, con lo cual se evita el problema de la falsa simetría y la inclusión de valores irreales, como sucede cuando se incluyen en el intervalo valores negativos en variables que no los poseen. El rango puede ser sustituido por el intervalo sin valores atípicos o simplemente por el intervalo que contiene el 90 % de los datos centrales o cualquier otro porcentaje determinado por el investigador.

Cuando el tamaño de muestra no es demasiado grande, adicionar los puntos de los datos originales (*raw data*) es una muy buena opción para visualizar la distribución subyacente a los descriptivos del gráfico de cajas (Figura 4A). Para ello se adiciona un factor de dispersión que evita la superposición de valores coincidentes (en inglés *jitter*), y de esta forma se hace mucho más informativo y preciso el gráfico. Solo con esta adición se hacen evidentes los patrones en los datos y la interpretación de las figuras puede cambiar radicalmente. En la Figura 4A se observa con claridad que el tamaño de muestra y la amplia dispersión de los datos de la localidad C no permiten sustentar ninguna diferencia, y que tienen una distribución marcadamente bimodal. Por esta razón, el valor medio no describe realmente la tendencia central ya que cae en una zona de muy baja probabilidad de observar ese dato.

En relación a esta estrategia de mejora se ha planteado que la visión humana no juzga de forma muy precisa las diferencias en densidades, los momentos estadísticos y las distribuciones de nubes de puntos dispersos (Bobko & Karren 1979, Zylberberg & *al.* 2014, Spence & *al.* 2016). Además, la interpretación también depende fuertemente del tipo y tamaño de punto seleccionado, y su utilidad se limita severamente cuando el tamaño de muestra es grande. En este último caso, la visualización de todos los datos no es recomendable ya que tiende a producir un efecto de amontonamiento que dificulta la interpretación gráfica. Por ello, en estos casos se pueden sustituir los datos crudos por representaciones de la distribución de frecuencias o probabilidad de los valores.

La forma más tradicional de describir la distribución de un conjunto de datos es a través de un histograma, y estos se pueden acoplar al gráfico de cajas (Figura 4B). En este caso, es importante recordar que la forma de un histograma es fuertemente dependiente del ancho del intervalo de clase que se utilice, por lo que para hacer comparaciones de sus formas entre figuras se tiene que utilizar el mismo intervalo o se distorsiona la información. Los histogramas, además, desde el punto de vista visual pueden ser cargantes y se pierde la información de la distribución de valores dentro de cada intervalo.

Las desventajas de los histogramas causaron que Tapia & Thompson (1978), Parzen (1979), Silverman (1986) y Scott (1992) propusieran y definieran otras alternativas de estimadores de densidad, en forma de curvas continuas. Existen diferentes métodos para obtener estas curvas. Una de ellas es la propuesta por Chambers & *al.* (1983) como la fracción de valores de los datos por unidad de medida, que aparecen en un intervalo centrado en cada valor. También se emplean las distribuciones de Pareto (Ultsch 2005) o las funciones *kernel* (Silverman 1986) (Figura 4C), estas últimas de las más empleadas. Algunos autores que no temen a estilos más recargados pueden emplear ambas formas de distribución (histogramas y curvas) simultáneamente (Figura 4D).

Aunque estas curvas son bastante intuitivas para poder interpretar totalmente, se debe conocer cómo se construyen. La idea básica de una curva *kernel* es asumir a los valores observados como representantes de otros valores no observados en su vecindad. Por tanto, se calcula asociado a cada uno de ellos una verosimilitud relativa, normalmente distribuida, en determinado rango a su alrededor (Figura 5). Luego, para cada punto del eje se suman los valores de probabilidades asociados a cada valor y la curva resultante se re-escala a un valor de área igual a 1 (Scott 1992).

Las curvas de densidad, tienen como opción el ancho de los intervalos, que determinará cuan suavizada o no será la figura resultante, lo que tiene un efecto semejante al de la selección del tamaño de clase de un histograma. La selección de este valor depende del conocimiento de los datos y del mensaje que se quiera transmitir con la figura. No se debe seleccionar a la ligera porque un valor demasiado alto producirá resultados sobre-suavizados que ocultarán patrones importantes dentro de la muestra y transmitirá la ilusión de un conocimiento perfecto de la distribución de probabilidades. Pero un valor demasiado pequeño mostrará un sobreajuste a los datos, con muchas irregularidades en respuesta a ruidos o variaciones muy locales. En general, porcentajes entre 10 y 40 % del rango son recomendables. Estas curvas, tienden a sobrepasar el rango de los propios datos, y es una opción del investigador permitir esta inferencia extendida o cortarlas a la altura de los datos extremos.

Al combinar estos elementos, puntos, histogramas y curvas, se pueden producir figuras de alto grado de profesionalismo, con un gran contenido de información y que permiten visualizar todas las propiedades de la muestra de datos.

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

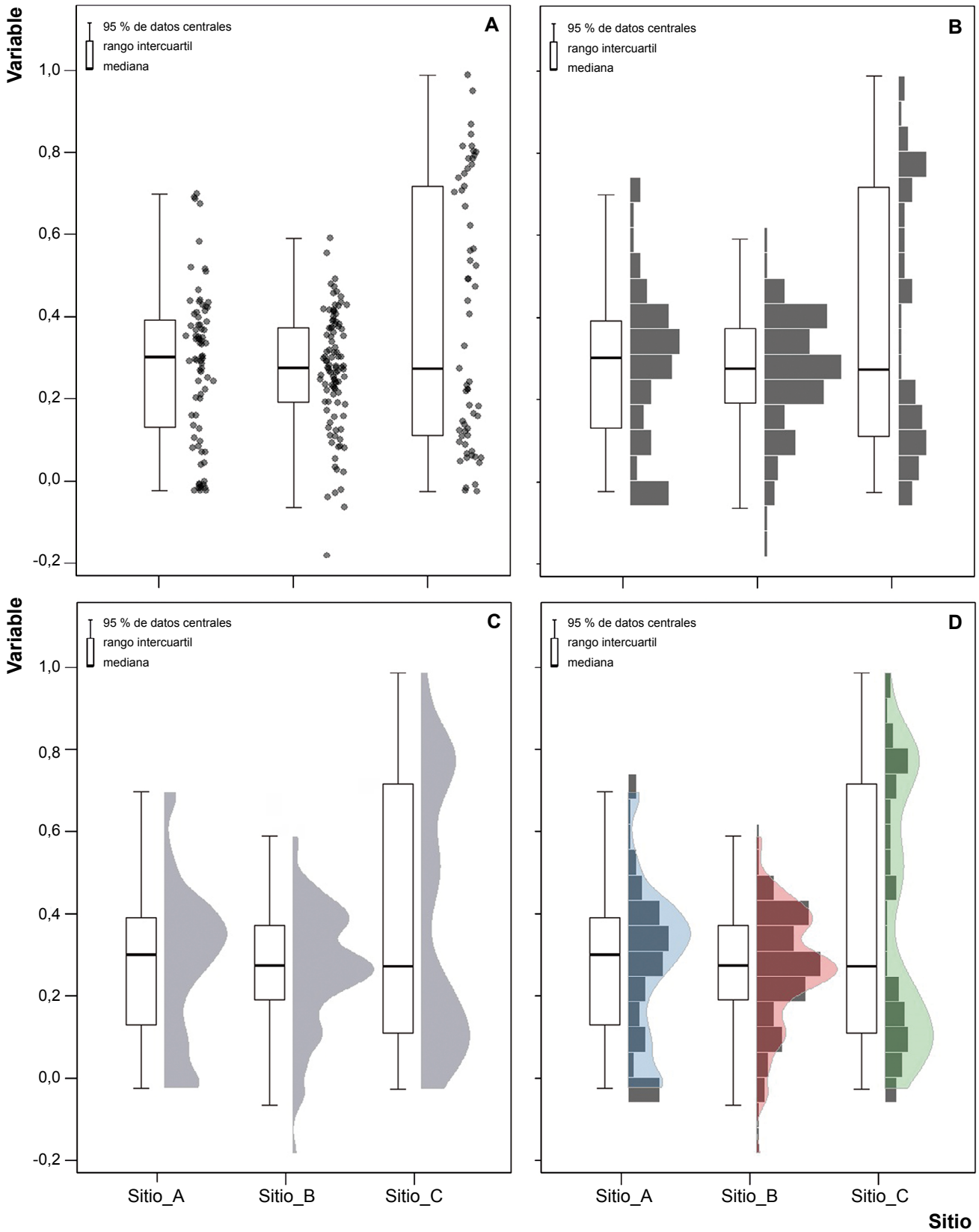


**Fig. 4.** Variantes básicas de gráficos de cajas con representación adicional de la distribución de los datos por diferentes vías. **A**. Con datos originales dispersos. **B**. Con el histograma de los datos. **C**. Con una curva *kernel* de distribución de densidad. **D**. Mezcla de las representaciones B y C.

***Fig. 4.*** *Basic enhancements of box plots by incorporating a representation of data distribution by different ways.* ***A****. Adding jittered row data.* ***B****. Adding the histogram.* ***C****. Adding a kernel density distribution curve.* ***D****. Mixture of previous B and C graphics.*

Algunas combinaciones estándares han sido publicadas bajo distintos nombres: gráficos de violín, gráficos de vainas, *sinaplot*, gráficos de pirata, etc. (Figura 6).

Cuando se emplean las curvas de densidad o histogramas suavizados duplicados a ambos lados del eje central, aparece el gráfico de violín (Hintze & Nelson 1998, Adler 2005) (Figura 6A), una de las primeras variantes de gráfico de cajas que incluyó estos elementos para caracterizar la distribución subyacente de los valores. Con esta representación, las diferencias en la distribución de valores se hacen muy evidentes pero los datos individuales no son visibles y por tanto, no hay indicador del número de observaciones. El nombre proviene de su apariencia y han sido criticados por no cumplir exactamente la regla de la tinta mínima de Tufte (1983), por aquellos autores que se cuestionan la necesidad de duplicar la información de la densidad de la distribución a ambos lados de la figura.

Kampstra (2008) propuso una combinación entre un gráfico de banda (*stripchart*) y una curva de densidad, figura a la cual llamó gráfico de vaina (*beanplot*) (Figura 6B). El gráfico de banda es un diagrama de dispersión unidimensional,

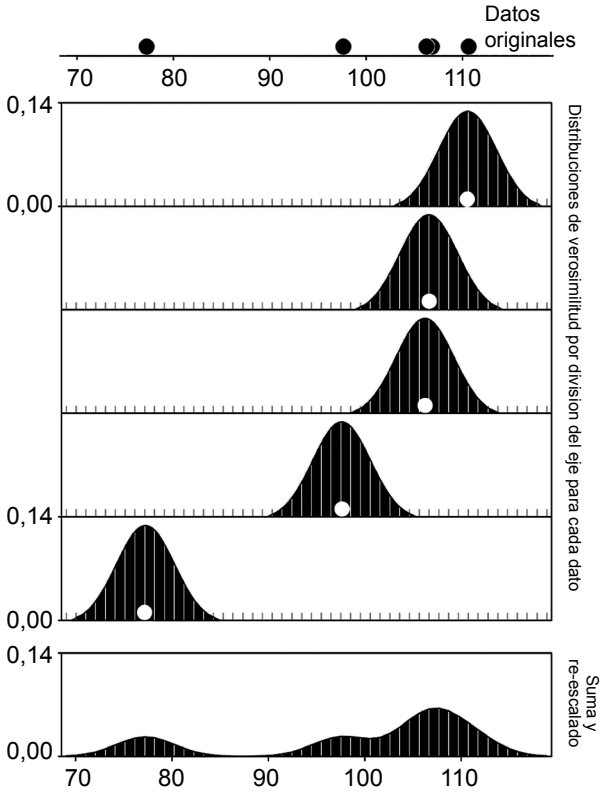


**Fig. 5.** Procedimiento general para obtener una curva *kernel* de distribución de la densidad de valores de una muestra. Las verosimilitudes asociadas a cada división del eje, calculadas para cada dato de la muestra (representados por los círculos), se suman y se re-escala la curva a área igual a 1.

***Fig. 5.*** *General procedure to obtain a kernel distribution curve for a data sample. Associated likelihood in each axe division, estimated for each data of the sample (represented by the dots), are summed and the curve area is re-escalated to 1.*

en el que solo se presentan los puntos de los datos individuales (*e.g.,* Box & *al.* 1978), y no se provee ningún estadístico de resumen. El gráfico de vaina incluye un único estadístico de resumen, el promedio. La curva de densidad es reflejada a ambos lados, por lo cual externamente es similar al gráfico de violín. Como el uso del mismo ancho de banda para todas las muestras pudiera tener un efecto de distorsión en aquellas con muy pocos datos, en las cuales el ancho de la curva estaría exagerado, en las muestras con menos de 10 datos los anchos son re-escalados linealmente (es decir, una muestra de 3 datos tendría solo 3/10 de su ancho normal). Los gráficos de vaina son una buena alternativa para comparaciones pareadas, ya que pueden hacerse asimétricos, es decir de forma que cada mitad pertenezca a una muestra de cada par comparado.

Una variante reciente son los llamados gráficos de pirata (*pirate plot*) (Phillips 2016) (Figura 6C), que incluye componentes de las tres categorías: datos originales, estadísticos de resumen y estadísticos inferenciales. Este gráfico agrupa cuatro elementos principales: puntos de datos, tendencia central, curva de densidad y un intervalo de inferencia que pueden ser intervalos de confianza frecuentistas o intervalos bayesianos de máxima densidad. Esta visualización es realmente una extensión del *beanplot* (Kampstra 2008), mezclado con otras opciones.

El gráfico de nubes y lluvia (*raincloud plot*) emplea ambos recursos: curvas de densidad y valores de los puntos (Allen & *al.* 2018). Recibe su nombre de la apariencia que tienen cuando se grafican de manera horizontal (Figura 6D) y combina un "medio violín" por encima del gráfico de cajas y los datos dispersos por debajo. Como sus autores reconocen, no es una propuesta totalmente nueva, sino una combinación potente de los mismos recursos de las anteriores.

El *sinaplot* (Sidiropoulos & *al.* 2018) también representa una simplificación máxima derivada inicialmente de los gráficos de cajas pero que prácticamente pierde todos sus componentes iniciales (Figura 6E). Este se basa simplemente en representar la dispersión de los datos, restringida dentro de la curva de densidad de probabilidades (un gráfico de banda con dispersión en sus puntos [*jitter*]). De una manera simple, mantiene la información de la distribución de los datos, tamaño de muestra, modalidad y valores atípicos, pero se desliga totalmente de los estadísticos formales. En cierta forma es una sinergia entre el gráfico de violín y el gráfico de banda. Formas más complejas de dispersión gráfica de los puntos han dado origen al gráfico de enjambre (*beeswarm plot*) (Eklund 2016) (Figura 6F).

Para comparaciones de múltiples muestras, una variante atractiva es la llamada figura de cadenas montañosas o graficos de cordilleras (*mountain range plot*) (Figura 6G), que puede describir muy bien variaciones temporales y que a diferencia de las anteriores se orienta en sentido horizontal. Esta también es válida para mostrar resultados de diseños complejos, como los de medidas repetidas (Figura 6H).

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

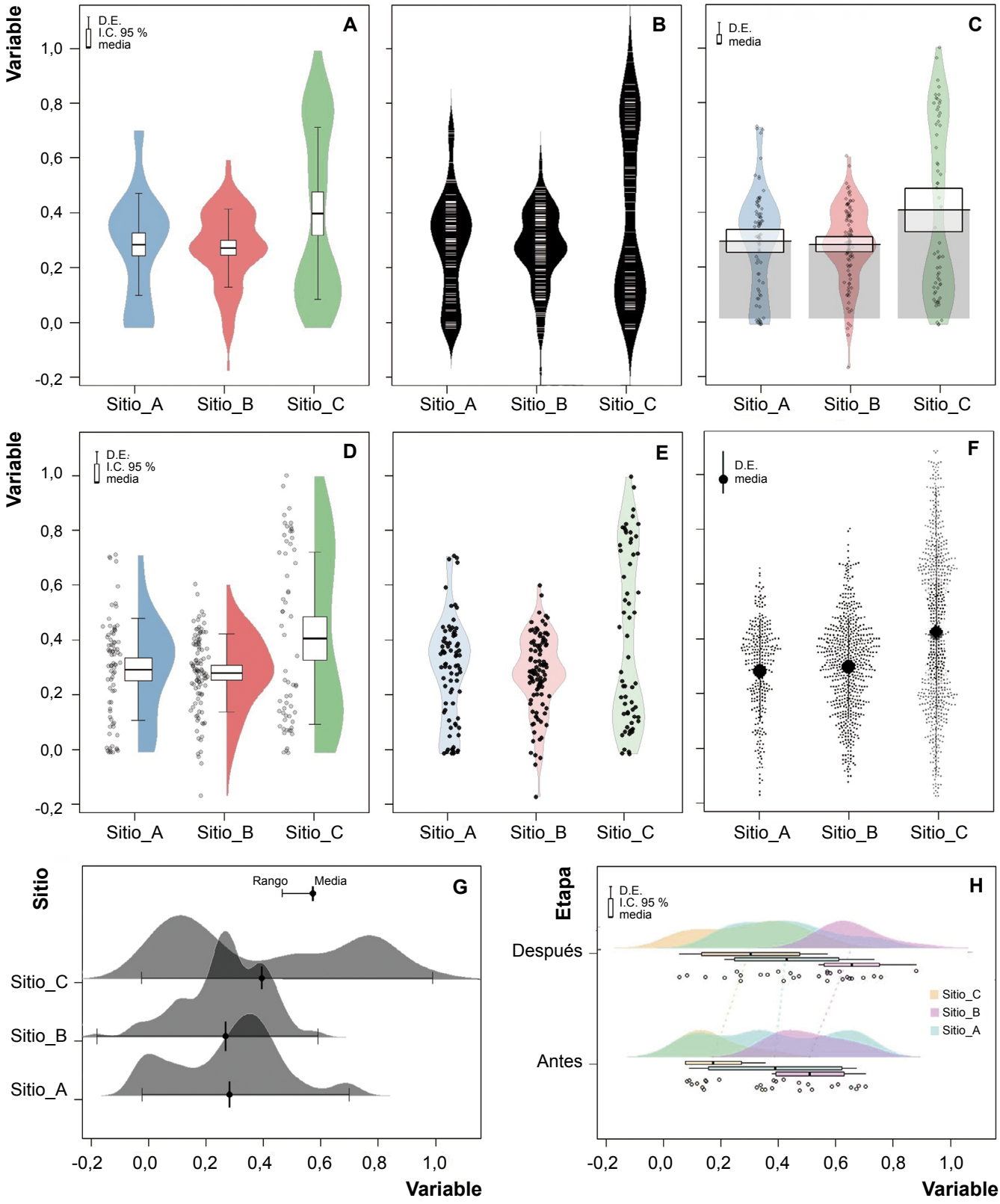


**Fig. 6.** Variantes actuales y derivadas de gráficos de cajas que incorporan elementos de mejora. **A**. Gráfico de violín. **B**. Gráfico de vaina. **C**. Gráfico de pirata. **D**. Gráfico de nubes y lluvia. **E**. *Sinaplot*. **F**. Gráfico de enjambre **G**. Gráfico de cordillera. **H**. Gráfico para un diseño de medidas repetidas

***Fig. 6.*** *Modern variations and derivatives of box plots that incorporate improvements elements.* ***A****. Violin plot.* ***B****. Beanplot.* ***C****. Beeswarm plot.* ***D****. Pirate plot.* ***E****. Raincloud plot.* ***F****. Sinaplot.* ***G****. Mountain range plot.* ***H****. Graphic for a repeated measurement design.*

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

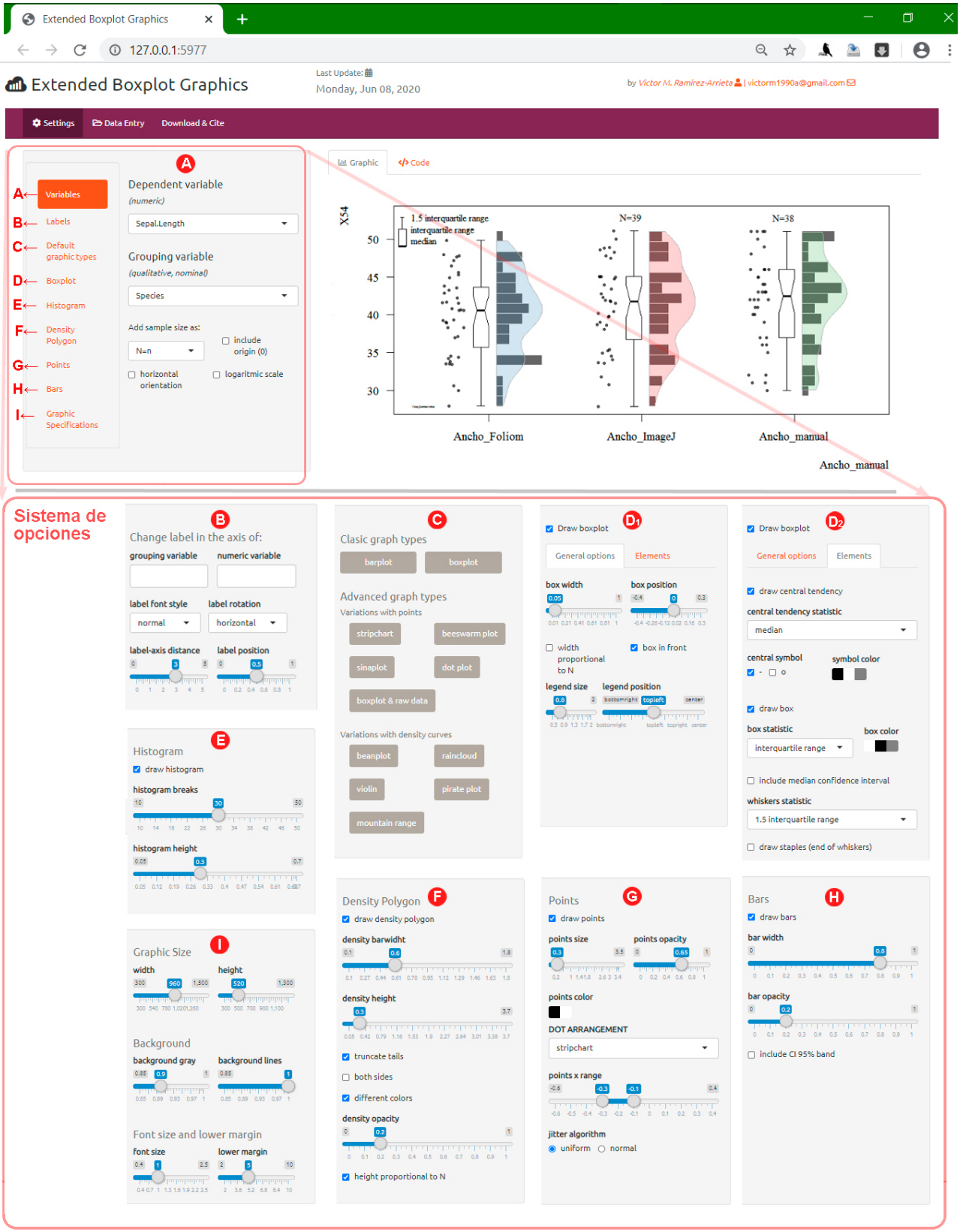


**Fig. 7.** Vista inicial y sistema de opciones de la aplicación *Extended Boxplot Graphics* programada en el entorno *R* y con interfaz *html* diseñada con el paquete *shiny* para facilitar la creación de gráficos de cajas mejorados (disponible en: https://vmra.shinyapps.io/univariados/).

***Fig. 7.*** *View and option system in the Extended Boxplot Graphics app programmed in* R *language with an* html *interface in* shiny, *to facilitate creation of enhanced boxplots, box plots (available in: https://vmra.shinyapps.io/univariados/).*

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

Estos tipos de figuras no aparecen en los programas estadísticos tradicionales y se han hecho posible gracias a las ventajas del lenguaje modular y la gramática gráfica del lenguaje *R* (Wickham & Chang 2008). Como una herramienta de ayuda a aquellos que no dominan la creación o manejo de códigos de *R*, aquí se presenta la herramienta *Extended Boxplot Graphics* que puede ser utilizada *online* en el sitio https://vmra.shinyapps.io/univariados/ (DOI: 10.13140/RG.2.2.23249.76643) sin necesidad de instalar el programa *R*, o puede ser descargada para su empleo local en máquinas con el programa.

Esta aplicación lee los ficheros de datos en formatos *txt* o *csv*, con cualquier tipo de separador y decimales marcados con puntos o comas. Luego de seleccionar las variables a graficar, a través de un sistema de opciones (Figura 7), se pueden seleccionar las formas gráficas típicas por defecto (gráfico de barras con líneas de error, gráfico de cajas, *stripchart*, *sinaplot*, *dotplot*, *beanplot*, *raincloud plot*, *violin plot*, *pirate plot*) o se pueden combinar y personalizar los elementos de forma independiente: histogramas, polígonos de densidad, puntos y barras (también se incluyen otras dos figuras menos relacionadas a los diagramas de cajas que son el *beeswarm plot* y el *mountain range plot*). Las figuras pueden generarse de tamaños regulables hasta de 1500 × 1300 píxeles y se extraen simplemente mediante la copia y el pegado en los documentos de trabajo.

Una ventaja adicional, incorporada para que la aplicación sea útil para conocedores del lenguaje *R*, es la capacidad que posee para que, una vez diseñada la figura con las opciones deseadas, se genere de forma independiente el código de *R* que la crea. Esto puede ayudar a hacer otras modificaciones no implementadas en la aplicación o a la publicación de los códigos como información suplementaria de los artículos, lo que aumenta su replicabilidad (Denis 2020). Una aplicación desarrollada con objetivo similar, fue presentada por Spitzer & *al.* (2014) (http://boxplot.tyerslab.com/), pero la desarrollada para el presente trabajo muestra ventajas en términos de actualización, flexibilidad y posibilidades de personalización.

## CONSIDERACIONES FINALES

Pocos trabajos han hecho énfasis en los errores de interpretación que se producen por la forma en que se presentan los gráficos en los medios de comunicación (*e.g.*, Moore & *al.* 1979, Wainer 1984, Weissgerber & *al.* 2015), pues la gran mayoría no se preocupan por ello y los usan indiscriminadamente. Muchos elementos de forma pueden ser personalizados en estos gráficos, por lo que es muy importante que se describa al detalle la manera de representarlos, a menos que los datos y códigos estén libremente disponibles.

A pesar de las deficiencias de los gráficos de cajas básicos, con las mejoras actualmente empleadas vuelven a posicionarse como una de las visualizaciones de datos más completas e informativas para los análisis exploratorios y para complementar análisis inferenciales. Estos aspectos ratifican la metáfora inicial de que, en sus versiones mejoradas, un gráfico de cajas puede ser equivalente a muchas palabras. Ello se debe a que contienen un volumen condensado de información estadística muy superior a otros tipos de figuras (como los gráficos de dispersión o los diagramas de barras). También permiten combinar un examen gráfico exploratorio de los datos individuales, con estadísticos descriptivos numéricos que resumen las características de la muestra en su conjunto e incluso con estadísticos inferenciales. Si son empleados sobre valores de tamaños de efecto, los gráficos de cajas pueden llegar a sustituir pruebas estadísticas de comparación de hipótesis nulas (Cumming 2007). Su inspección permite evaluar de forma inicial el cumplimiento de asunciones subyacentes (por ejemplo, simetría, homocedasticidad, normalidad). Finalmente, son muy buenos para identificar posibles valores atípicos (Dai & Genton 2018, Hussain 2019). Para ayudar a las inferencias se pueden utilizar estadísticos relacionados, como intervalos de confianza, o densidades de probabilidades bayesianas *a posteriori* u otros estimadores de parámetros (Ho & *al.* 2018).

Con el advenimiento de herramientas de modulares flexibles para hacer gráficos como el *ggplot2* de *R* (Wickham & Chang 2008, Wickham 2010, Patil 2018), todos los componentes de estas variantes de gráfico de cajas pueden usarse de manera complementaria y de seguro aparecerán nuevas variantes. Con las facilidades de *R* los investigadores continúan en los intentos de producir visualizaciones de datos más robustas, transparentes e intuitivas, por ejemplo: *estimation plots* (Ho & *al.* 2018).

El incremento actual en el nivel de detalles, cantidad de información y complejidad de los datos de las investigaciones puede sobresaturar la capacidad de visualización y comprensión humana (Carr 2002). La visión del ser humano no está adaptada naturalmente a la comprensión óptima de figuras muy complejas y tiende a pasar por alto muchos detalles. Puede pensarse que la búsqueda de formas de visualizar los datos es algo sencillo, pero no lo es. Encontrar una representación ideal para un conjunto específico de datos que balancee funcionalidad, interpretabilidad y complejidad, sin sacrificar la estética, puede ser un reto intelectual profundo que lleva a tener en cuenta muchos aspectos teóricos, metodológicos, de diseño gráfico y hasta psicológicos para poder obtener visualizaciones capaces de cumplir eficientemente su objetivo. Los gráficos de la familia derivada de los diagramas de cajas, si son usados apropiadamente, representan un arsenal invaluable para la comunicación de la ciencia por los investigadores.

## CONTRIBUCIÓN DE LOS AUTORES

D. Denis concibió la idea original, diseñó la aplicación y escribió la primera versión del manuscrito. V.M. Ramírez-Arrieta programó en *R* la aplicación y depuró el código. Ambos autores compilaron la información presentada, y revisaron la aplicación y el manuscrito.

## CUMPLIMIENTO DE NORMAS ÉTICAS

*Conflicto de intereses*: Los autores declaran que no existen conflictos de intereses.

*Consentimiento para la publicación*: Todos los autores han dado su consentimiento para publicar este trabajo.

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

## REFERENCIAS BIBLIOGRÁFICAS


Adler, D. 2005. vioplot: Violin Plot. R package version 0.2. http://CRAN.R-project.org/package=vioplot.

Allen, M., Poggiali, D., Whitaker, K., Rhys, T. & Kievit, R. 2018. Raincloud lots: a multi-platform tool for robust data visualization. *PeerJ Preprints*, doi.org/10.7287/peerj.preprints.27137v1.

Benjamini, Y. 1988. Opening the Box of the Box Plot. *Am. Stat.* 42: 257-262.

Bobko, P. & Karren, R. 1979. The Perception of Pearson Product Moment Correlations from Bivariate Scatterplots. *Pers. Psych*. 32(2): 313-325.

Box, G.E.P., Hunter, W.G. & Hunter, J.S. 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons. Hoboken, USA.

Buja, A., Cook, D. & Swayne, D.F. 1996. Interactive highdimensional data visualization. *J. Comp. Graph. Stat.* 5: 78-99.

Carr, D. 2002. Graphical displays. Pp. 933-960. En: El-Shaarawi, A.H. & Piegorsch, W.W. (eds.). Encyclopedia of Environmetrics, Volumen 2. John Wiley & Sons, Ltd. Chichester, UK.

Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. 1983. Graphical Methods for Data Analysis. Wadsworth. Belmont, CA, USA.

Cleveland, W.S. 1993. Visualizing Data. Hobart Press. Hobart, Australia.

Cumming, G. 2007. Inference by Eye: Pictures of Confidence Intervals and Thinking About Levels of Confidence. *Teach. Stat.* 29(3): 89-93.

Dai, W. & Genton, M. 2018. Multivariate functional data visualization and outlier detection. *J. Comp. Graph. Stat*. 27(4): 923-934.

Denis, D. 2020. Las crisis actuales de la ciencia. *Revista Cub. Cienc. Biol.* 8(1): 1-16.

DuToit, S.H.C., Steyn A.G.W. & Stumpf R.H. 1986. Graphical Exploratory Data Analysis. Springer-Verlag Inc. New York, USA.

Eklund, A. 2016. beeswarm: the bee swarm plot, an alternative to stripchart. R package version 0.2.3.

Frigge, M., Hoaglin, D.C. & Iglewicz, B. 1989. Some Implementations of the Box Plot. *Am. Stat*. 43: 50-54.

Hintze, J.L. & Nelson, R.D. 1998. ViolinPlots: A Box Plot - Density Trace Synergism. *Am. Stat*. 52(2): 181-184.

Ho, J., Tumkaya, T., Aryal, S., Choi, H. & Claridge-Chang, A. 2018. Moving beyond P values: Everyday data analysis with estimation plots. *BioRxiv*. https://doi.org/10.1101/377978.

Hoaglin, D.C., Mosteller, F. & Tukey J.W. (eds). 1983. Understanding Robust and Exploratory Data Analysis. Wiley. New York, USA.

Hoaglin, D.C., Mosteller F. & Tukey J.W. (eds). 1985. Exploring Data Tables, Trends, and Shapes. Wiley. New York, USA.

Hubbard, K.E. & Dunbar, S.D. 2017. Perceptions of scientific research literature and strategies for reading papers depend on academic career stage. *PLOS ONE* 12:e0189753.

Hussain, I. 2019. Outlier Detection using Graphical and Nongraphical Functional Methods in Hydrology. *Int. J. Adv. Comp. Sci. App.* 10(12): 438.

Kampstra, P. 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *J. Stat. Soft.* 28(1): 1-9.

Krzywinski, M. & Altman, N. 2014. Points of significance: Visualizing samples with box plots. *Nat. Methods* 11(2): 119-120.

McGill, R., Tukey, J.W. & Larsen, W.A. 1978. Variation of boxplots. *Am. Stat.* 32: 12-16.

Moore, M.V., Nawrocki, L.H. & Simutis, Z.M. 1979. The instructional effectiveness of three levels of graphics displays for computer-assisted instruction. Report No. ARI-TP-359. U.S. Army Research Institute for the Behavioral and Social Sciences. Alexandria, Virginia, USA.

Morgenthaler, S. 2009. Exploratory data analysis. *WIREs Comp. Stat.* 1: 33-44.

Parzen, E. 1979. Nonparametric Statistical Data Modeling. *J. Am. Stat. Assoc.* 7(4): 105-131.

Patil, I. 2018. ggstatsplot: "ggplot2" Based Plots with Statistical Details. CRAN package. http://CRAN.R-Project.Org/Package=Ggplot2. junio de 2020.

Pérez, L. 2018. ¿Cómo proceder ante el incumplimiento de las premisas de los métodos paramétricos? o ¿cómo trabajar con variables biológicas no normales? *Revista Jard. Bot. Nac. Univ. Habana* 39: 1-12.

Phillips, N.D. 2016. The pirate plot (2.0)—the RDI plotting choice of R pirates. http://nathanieldphillips.com/ 2016/04/pirateplot-2-0-the-rdi-plotting-choice-ofr-pirates/. junio de 2020.

Scott, D.W. 1992. Multivariate Density Estimation; Theory, Practice and Visualization. Wiley. New York, USA.

Sidiropoulos, N., Sohi, S.H., Pedersen, T.L., Porse, B.T., Winther, O., Rapin, N. & Bagger, F.O. 2018. SinaPlot: an enhanced chart for simple and truthful representation of single observations over multiple classes. *J. Comp. Graph. Stat*. 1-12.

Silverman, B.W. 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall. New York, USA.

Spear, M.E. 1952. Charting Statistics. Editorial McGraw-Hill. New York, USA.

Spence, M.L., Dux, P.E. & Arnold, D.H. 2016. Computations underlying confidence in visual perception. *J. Exp. Psych.: Human Percep. Perf.* 42(5): 671-682.

Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. 2014. BoxPlotR: a web tool for generation of box plots. Correspondence. *Nat. Methods* 11(2): 121.

Streit, M. & Gehlenborg, N. 2014. Points of View: Bar charts and box plots. *Nat. methods.* 11(2): 117.

Tapia, R.A. & Thompson, J.R. 1978. Nonparametric probability density estimation. Johns Hopkins University Press. Baltimore, MD.

Tufte, E.R. 1983. The Visual Display of Quantitative Information. Graphics Press. Cheshire, UK.

Tukey, J.W. 1977. Exploratory Data Analysis. Readings, M.A. Addison-Wesley.

Tukey, J.W. 1986. Sunset salvo. *Am. Stat.* 40: 72-76.

Ultsch, A. 2005. Pareto density estimation: A density estimation for knowledge discovery. Pp. 91-100. En: Baier, D. & Werrnecke, K.D.

(eds.). Innovations in classification, data science, and information systems. Vol. 27. Springer. Berlin, Germany.

Velleman, P.F. & Hoaglin, D.C. 1981. Applications, basis and computing of Exploratory Data Analysis. Duxbury Press, Boston, USA.

Wainer, H. 1984. How to display data badly. *Am. Stat.* 38(2): 137-147.

Wainer, H. 1990. Graphical visions from William Playfair to John Tukey. *Stat. Sci.* 1: 340-346.

Weissgerber, T.L., Milic, N.M., Winham, S.J. & Garovic, V.D. 2015. Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology* 13(4): e1002128.

Wickham, H. & Chang, W. 2008. ggplot2: An implementation of the Grammar of Graphics. R Package Version 0.7. http://CRAN. R-Project. Org/Package=Ggplot2

Wickham, H. 2010. A layered grammar of graphics. *J. Comp. Graph. Stat*. 19(1): 3-28.

Zylberberg, A., Roelfsema, P. R. & Sigman, M. 2014. Variance misperception explains illusions of confidence in simple perceptual decisions. *Consc. Cognition* 27: 246-253.

 URL: www.rjbn.uh.cu ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)