Uso de R para la enseñanza de buenas prácticas de visualización de datos en investigación biológica

Anónimo

Palabras clave: visualización de datos - investigación biológica - gráficos

Introducción

La visualización de datos comprende al conjunto de herramientas necesarias para conseguir una representación gráfica de cualquier tipo de información de una manera clara y efectiva. Es una práctica que desempeña un papel fundamental tanto en el análisis exploratorio de datos como en la comunicación de los resultados de una investigación científica. Sin embargo, no resulta trivial implementarla adecuadamente y muchas veces es un tema difícil de enseñar.

El presente trabajo nace a partir de una presentación realizada en el marco del *VI Encuentro de Jóvenes Biofísicxs* en el mes de noviembre de 2022. A lo largo de nuestra experiencia formativa y de la actividad docente que desempeñamos dentro del área estadística de la institución de la que formamos parte, advertimos que, si bien la importancia de la visualización de datos para la investigación es incuestionable, muchos investigadores carecen de una formación adecuada en el tema y la literatura científica desborda de representaciones gráficas deficientes. Esto nos motivó a proponer, para nuestra intervención en las jornadas mencionadas, una revisión de distintas herramientas gráficas utilizadas en el análisis exploratorio de datos vinculados a la biofísica y la biología molecular, exponiendo las ventajas y desventajas de cada una de ellas, así como errores de construcción comunes. Dicha revisión fue elaborada usando diversas herramientas gráficas y de simulación en R, que se presentan en este trabajo.

Metodología y Abordaje

Para poder transmitir de manera clara y precisa conceptos claves de la visualización de datos, se buscó simular datos basados en ejemplos provenientes de las áreas de biofísica y biología molecular que permitan ilustrar los problemas que surgen del uso incorrecto de herramientas gráficas y que, a su vez, resulten de fácil comprensión para los asistentes. Se simularon conjuntos de datos de distintos tamaños (n entre 3 y 20) y distribución, los cuales fueron utilizados posteriormente para exponer, en el marco de cada uno de los casos de interés, el impacto de la implementación de distintas herramientas gráficas. Asimismo, se llevó adelante una revisión minuciosa de la literatura científica de las áreas de interés con el propósito de identificar el tipo de gráficos más frecuentemente utilizados para comunicar resultados de investigación y sus principales limitaciones.

Con el objetivo de mostrar la insuficiencia de los gráficos "dinamita", empleados con frecuencia en investigación biológica y bioquímica, se simularon conjuntos de datos caracterizados por medidas de centralidad y dispersión similares pero que presentaran diferencias importantes en la forma de sus distribuciones, incluyendo la presencia de valores atípicos. A través de la visualización de dichos conjuntos, se dejó en evidencia una de las serias limitaciones de este tipo de gráficos: el ocultamiento de información relevante en relación a la existencia de potenciales outliers, las características de forma de la distribución, el tamaño muestral, entre otros. Se propuso a los asistentes una alternativa fácil de lograr y útil en el caso de muestras con pocos datos (n=3-5): la superposición de las observaciones individuales (*jitter plot*) sobre el gráfico de media y barras de error, recuperando así la información perdida en el gráfico "dinamita".

Para el trabajo con muestras de mayor tamaño (n = 10-20) se mostraron las potencialidades y limitaciones de los *boxplots* en conjunto con otras alternativas, como los gráficos de violín o los gráficos de densidad de Ridge.

Por último, se simularon dos conjuntos de datos correlacionados para el trabajo con más de una muestra de datos. Se mostró, en este caso, la importancia que reviste la elección de un tipo de representación gráfica que

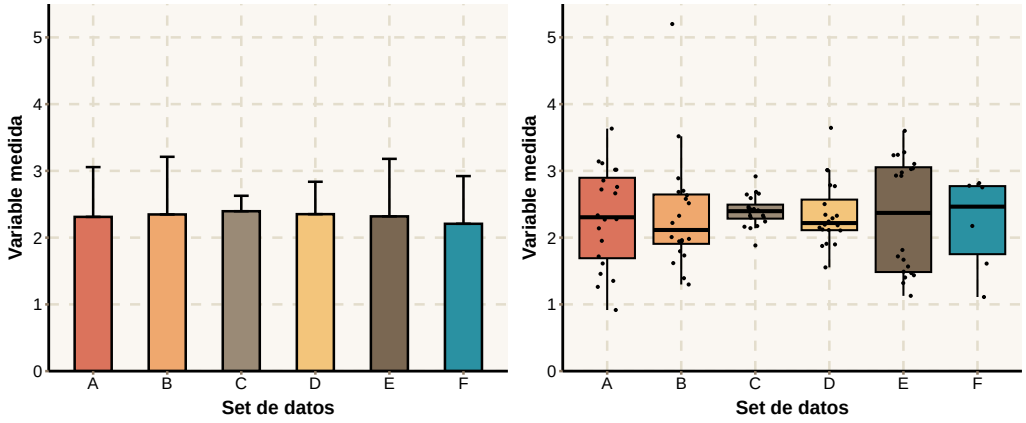

Figura 1: Una de las figuras mostradas en la presentación: Comparación del gráfico "dinamita"(izquierda) y boxplot + jitter plot (derecha) para representar una serie de conjuntos de datos con distribuciones muy diferentes.

se corresponda con el diseño experimental utilizado, tanto para el análisis exploratorio como para el posterior análisis inferencial.

Conclusiones y Perspectivas

El uso de herramientas de R facilitó la construcción de ejemplos fáciles de comprender para los asistentes y la demostración visual de las características, ventajas y desventajas asociadas a cada gráfico. De esta forma, se pudieron transmitir de manera sencilla y clara conceptos claves de la visualización de datos. Los participantes mostraron buena recepción e interés sobre la presentación y manifestaron que la misma los hizo cuestionar y modificar su metodología de trabajo para la aplicación de herramientas gráficas. Motivados por esto, decidimos continuar acercando propuestas de contenido similar en otras reuniones científicas y desarrollar herramientas didácticas que puedan ser compartidas, a futuro, en otras charlas, congresos y capacitaciones de investigadores. Con este objetivo, nos encontramos en este momento desarrollando una Shiny app para mostrar de manera interactiva estos contenidos.

Paquetes utilizados

Se usaron los paquetes `sn` (Azzalini 2023) y `truncnorm` (Mersmann et al. 2023) para simular datos con distribuciones normales sesgadas y truncadas; `faux` (DeBruine, Krystalli, and Heiss 2023) para simular datos con una estructura de correlación determinada y `ggplot2` (Wickham et al. 2023) para la elaboración de herramientas de visualización.

Referencias

10 Azzalini, A. 2023. *The Skew-Normal and Related Distributions Such as the Skew-t and the SUN*. https://cran.r-project.org/web/packages/sn/index.html.

DeBruine, L., A. Krystalli, and A. Heiss. 2023. *Simulation for Factorial Designs*. https://cran.r-project.org/web/packages/faux/index.html.

Mersmann, O., H. Trautmann, D. Steuer, and B. Bornkamp. 2023. *Truncated Normal Distribution*. https://cran.r-project.org/web/packages/truncnorm/index.html.

Wickham, H., W. Chang, L. Henry, T. Lin Pedersen, K. Takahashi, C. Wilke, K. Woo, Yutani H., and D. Dunnington. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://cran.r-project.org/web/packages/ggplot2/index.html.