

# Introducción

La visualización de datos comprende al conjunto de herramientas necesarias para conseguir una representación gráfica o mediante elementos visuales de la información de 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. Una visualización clara y precisa tiene el poder de contar una historia, en tanto permite el reconocimiento e identificación de patrones, tendencias y asociaciones entre variables, así como la detección de valores atípicos. Por el contrario, una mala elección en esta etapa puede generar confusión y resultados engañosos que conduzcan a conclusiones incorrectas. Si bien la importancia de la visualización de datos para la investigación científica es incuestionable, la literatura científica desborda de representaciones gráficas deficientes y pocos investigadores centran su atención en las visualizaciones en la misma manera en que lo hacen con la generación de datos o la escritura acerca de ellos. Con todo lo anterior en mente, en este trabajo se propuso evaluar el impacto que tienen las decisiones que se toman respecto a la visualización de datos para la comunicación de resultados en casos basados en ejemplos de investigación en ciencias bioquímicas.

# Visualización de pequeños conjuntos de datos

Para evaluar distintas herramientas gráficas, se modelaron datos teniendo como ejemplo un trabajo cuyo objetivo sea evaluar el efecto de la dosis de una droga sobre el **nivel de expresión génica** determinado por RT-qPCR de un gen de interés, para lo cual se realizan **5 determinaciones** en cada una de las condiciones bajo estudio (*dosis 0, dosis 1, dosis 2*). En la Tabla 1 se presentan los datos simulados que corresponderían a la variable **expresión génica relativa**.

Dosis 0	Dosis 1	Dosis 2
1.87	1.75	2.06
1.24	0.30	2.34
0.68	1.70	1.93
1.96	1.30	2.14
0.87	1.54	2.15

En la Figura 1 se presentan dos gráficos para los datos modelo. El **gráfico dinamita** (izquierda), ampliamente usado para estas situaciones, es poco informativo (muestra solo media y SD de cada conjunto) y puede llevar a conclusiones incorrectas.

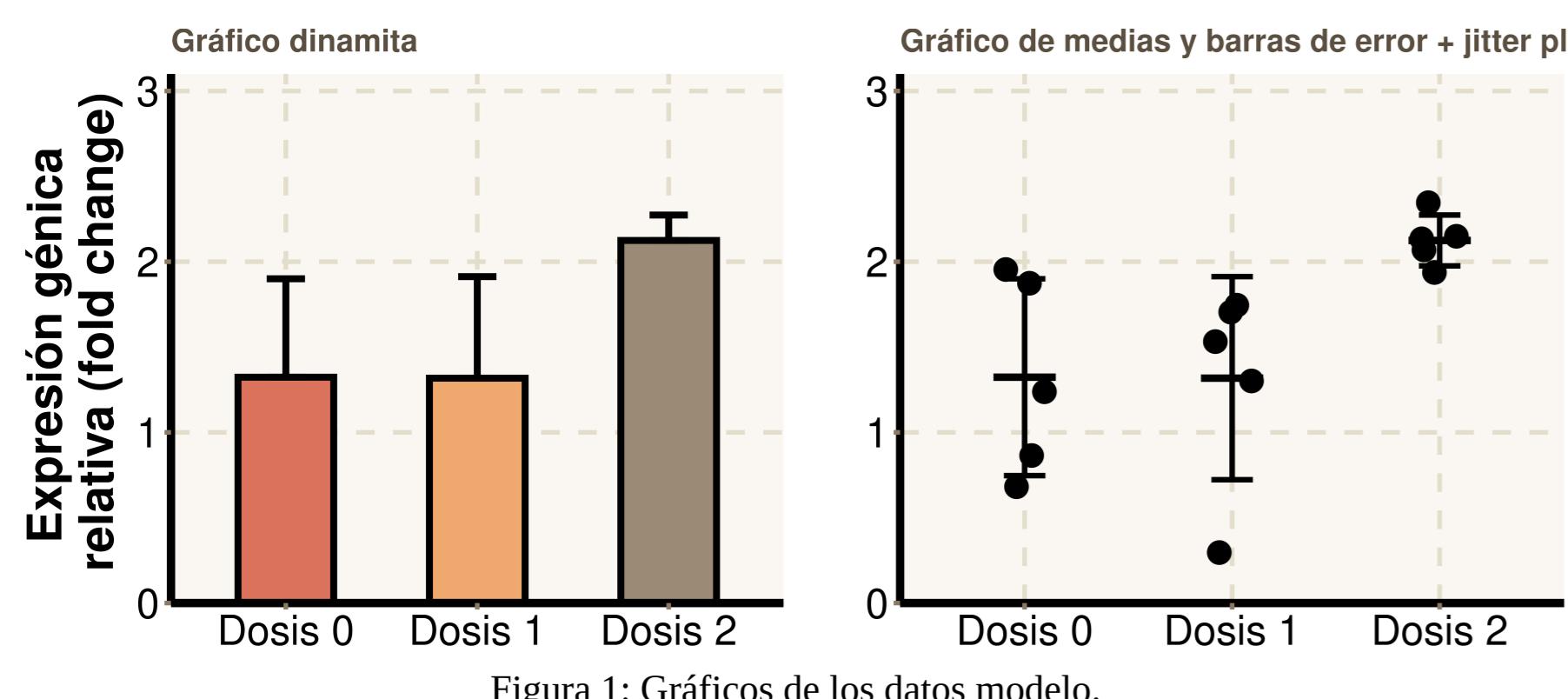


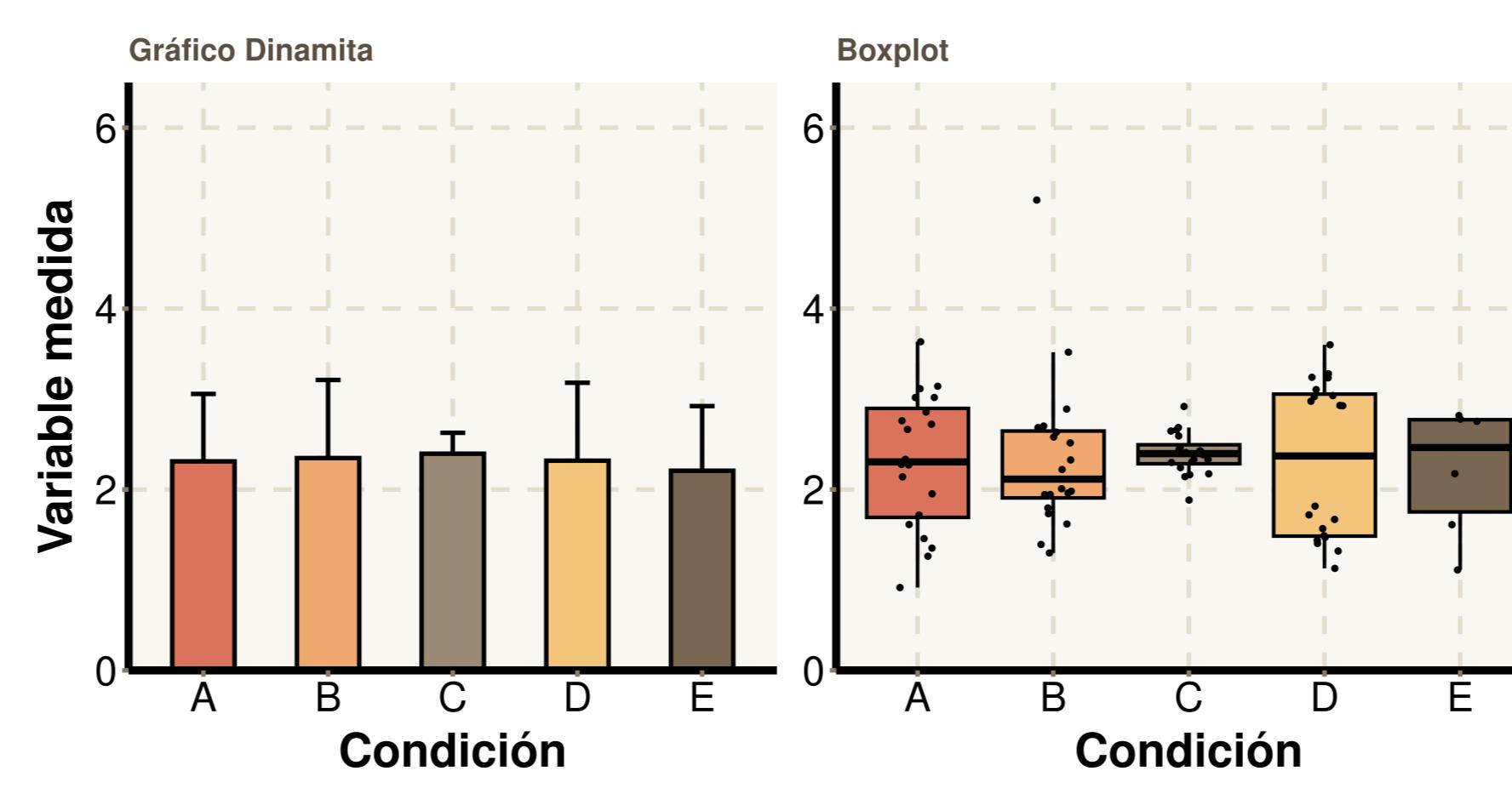
Figura 1: Gráficos de los datos modelo

# Métodos

Con base en ejemplos provenientes de las áreas de biofísica y biología molecular, y haciendo uso de herramientas de simulación en R, se evaluó la implementación de distintas herramientas gráficas durante el análisis exploratorio, examinando sus potencialidades y limitaciones según el tipo de dato a analizar, así como errores de construcción comunes. 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.

# Visualización de datasets de mayor extensión

A continuación se analiza la aplicación de herramientas de visualización al trabajar con un dataset simulado con mayor extensión muestral, considerando 5 niveles de un factor de interés. En la Figura 2 se encuentran representados gráficamente los datos. Los 5 conjuntos presentan importantes diferencias en su distribución.



The graph displays the following approximate data points:

Condición	30 min	60 min
1	0.73	0.75
2	0.64	0.73
3	0.63	0.72
4	0.61	0.69
5	0.62	0.70
6	0.61	0.67
7	0.54	0.61
8	0.44	0.50

# Conclusiones

El análisis realizado en este trabajo refuerza la importancia que tiene la elaboración de elementos de comunicación visual que resulten claros y precisos como medio para generar conocimiento a partir de los datos recolectados y comunicar resultados. Además proporciona herramientas valiosas para repensar esta práctica en el marco de investigación científica.

**Gráfico Dinamita.** Se pierde toda información acerca de la existencia de potenciales valores atípicos o extremos, sobre las características de forma de la distribución (por ejemplo: simetría o asimetría, bimodalidad), tamaño muestral, entre otros.

**Boxplot.** Construido a partir de cuartilos y valores observados mínimo y máximo. Permite visualizar de forma muy simple características de las distribuciones. La versión modificada permite detectar potenciales *outliers*. Agregando los valores observados se pueden identificar distribuciones bimodales, diferencias en los tamaños muestrales, etc.  
*No se puede utilizar cuando tenemos menos de 5 observaciones*

En la Figura 3 se muestra una alternativa a los gráficos anteriores: el **Diagrama de Violín**. Estos son gráficos de densidad de probabilidad espejados. Permiten visualizar fácilmente la distribución de conjuntos de muchos datos. Se les pueden agregar los valores observados, un boxplot, media y barras de error, etc. *No son útiles cuando tenemos pocos datos.*

**Tanto el Boxplot como el Diagrama de Violín resultan muy informativos y de gran utilidad. Qué gráfico resulta más adecuado dependerá de cada caso ya que no hay un solo gráfico superador que sirva para todo.**

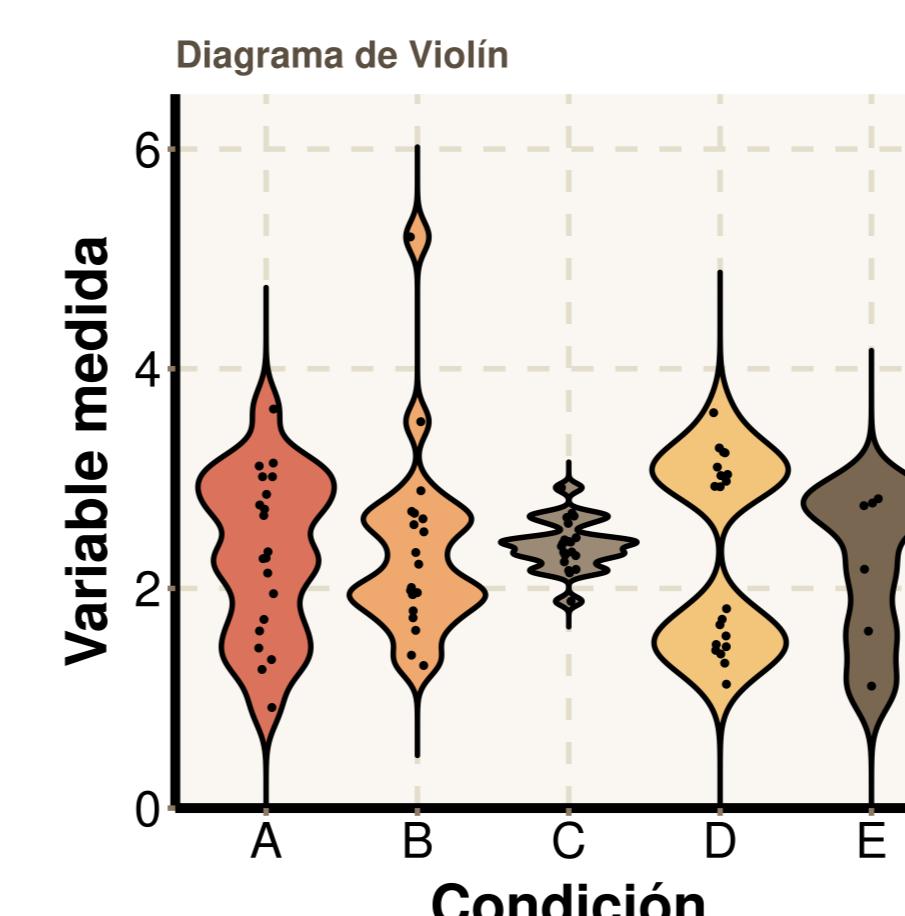


Figura 3: Diagrama de Violín para set de datos

# Visualización de datos correlacionados

Para analizar el impacto de las herramientas gráficas usadas con datos provenientes de muestras pareadas, se modelaron datos basados en un experimento en el cual se mide actividad enzimática por fluorescencia luego de dos tiempos de reacción de formación de aducto. Para evaluar si el tiempo de reacción da lugar a una diferencia en la actividad medida, se mide la actividad en cada una de 10 muestras luego de 30 y 60 minutos de reacción.

En la Figura 4 se muestra un boxplot múltiple para los datos simulados. Este gráfico nos llevaría a pensar en datos provenientes de **muestras independientes**. A simple vista, las diferencias entre ambos conjuntos no parecen ser grandes. Para evaluar si existen diferencias estadísticamente significativas en la actividad enzimática promedio entre los dos tiempos de reacción de interés, utilizariamos un **test-t para la comparación de dos promedios en base a muestras independientes**. Este test arroja un p-value de 0.1131 y concluiríamos que **las diferencias no son estadísticamente significativas**.

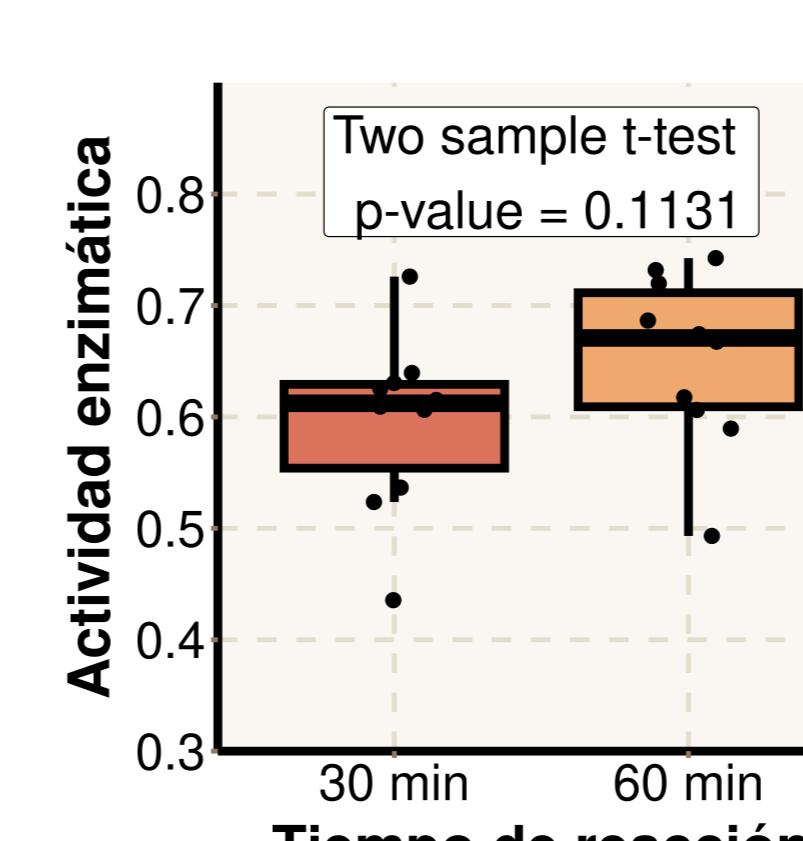


Figura 4: Boxplot para datos simulados