


FBIOyF

UNR

# ¿QUÉ HISTORIAS TIENEN NUESTROS DATOS PARA CONTAR?

## La visualización como una instancia clave en el análisis de datos

Lic. Joaquín Ferreyra - Lic. Natalia Labadie

Departamento de Matemática y Estadística. Facultad de Ciencias Bioquímicas y Farmacéuticas (Universidad Nacional de Rosario) - Instituto de Química Rosario (UNR - CONICET)

# Datos: los grandes protagonistas

**¿Qué son los datos?** Piezas de información que reunimos con el objetivo de conocer un poco más sobre nuestro campo de investigación (aunque tener datos a disposición no implica automáticamente el acceso esa información).

La **estadística** nos brinda herramientas para organizar, resumir, visualizar y procesar los datos que recolectamos para obtener la información que éstos contienen.

# Empecemos con un caso de estudio

El objetivo de un trabajo fue 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 realizaron **5 determinaciones** en cada una de las condiciones bajo estudio *(dosis 0, dosis 1, dosis 2)*.

A continuación se presentan las observaciones correspondientes a la variable **expresión génica relativa (fold change)**:

| 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 |

# Medidas descriptivas

Podríamos comenzar el análisis de nuestros datos calculando algunas medidas descriptivas de **posición** (media aritmética, mediana) y de **dispersión** (desviación estándar, rango intercuartil) para cada una de las condiciones bajo estudio por separado:

| | Posición | | Dispersión | |
|---|---|---|---|---|
| Condición | Media | Mediana | Desvío Estándar | Rango Intercuartil |
| Dosis 0 | 1.3 | 1.2 | 0.58 | 1.00 |
| Dosis 1 | 1.3 | 1.5 | 0.60 | 0.40 |
| Dosis 2 | 2.1 | 2.1 | 0.15 | 0.08 |

**¿Nos alcanza esta información para responder a nuestro objetivo?**

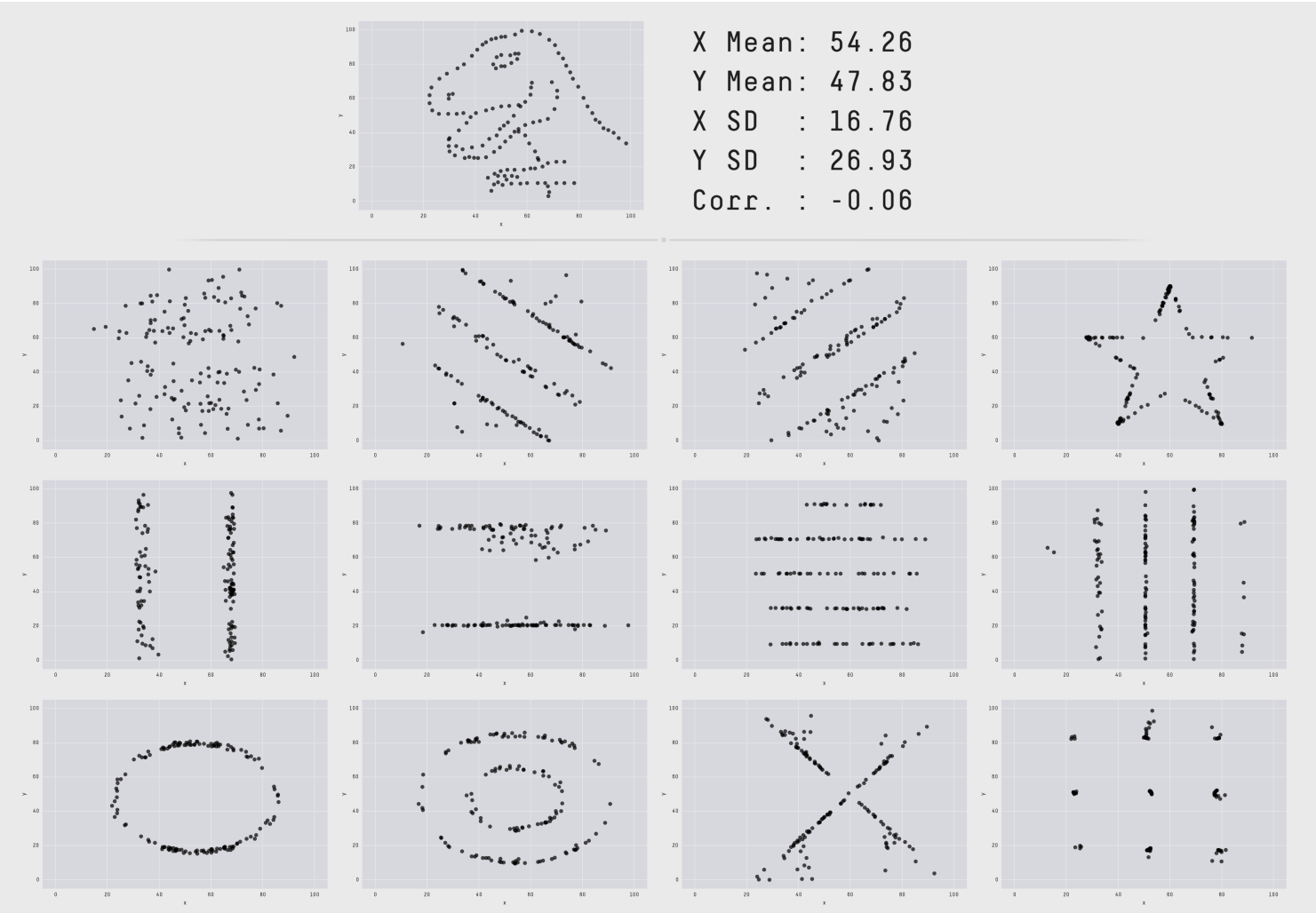
# La visualización de datos

**Las estadísticas no nos cuentan la historia completa.**

## Visualización de datos

Es una forma importante y poderosa de relacionar las ideas, experiencias e historias contenidas en esos datos, que facilita la presentación y la comunicación de información en diversos contextos.

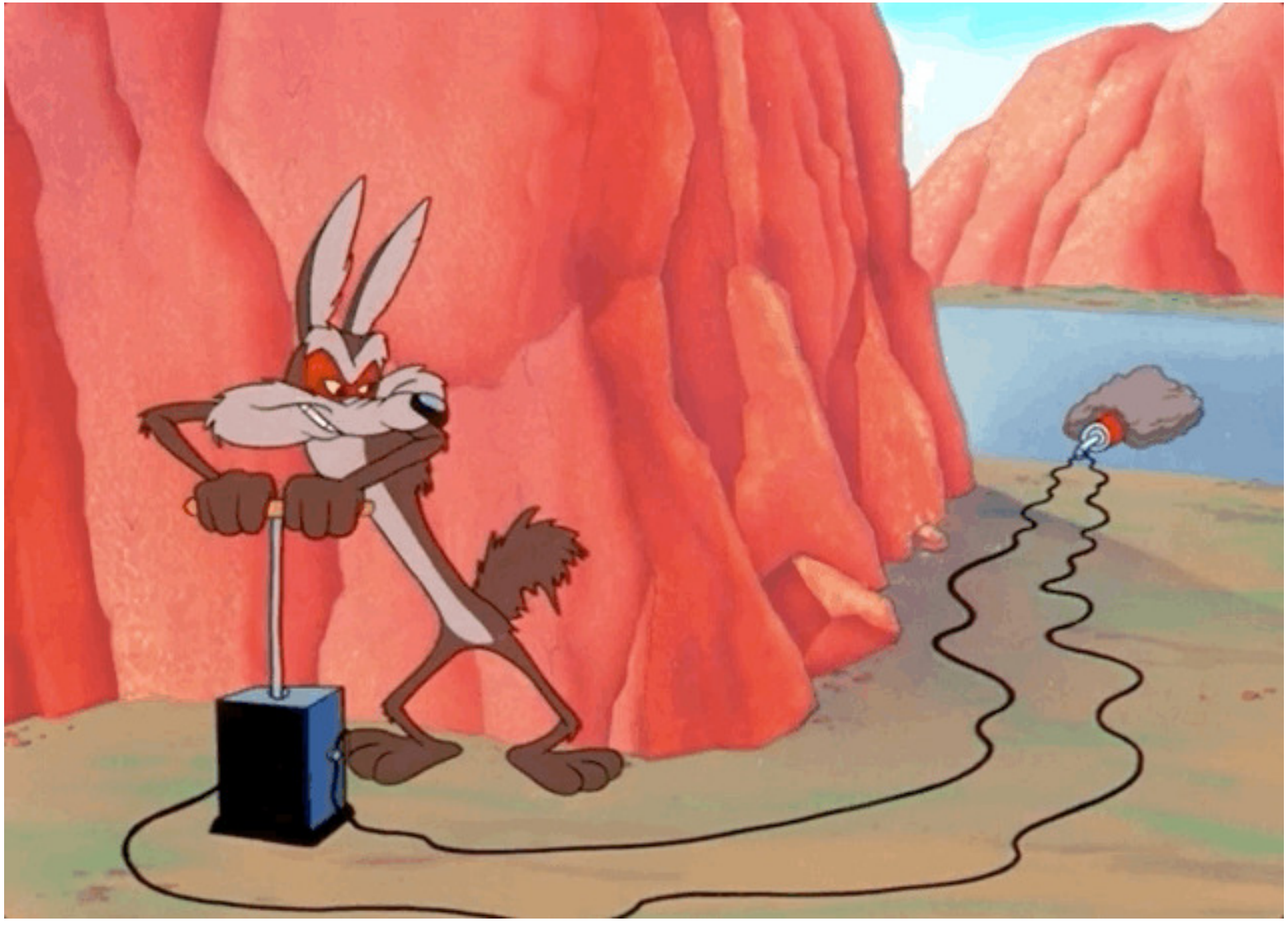
# Datasaurus Dozen[1]




[1] Matejka, J.; Fitzmaurice, G. (2017). **Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.**

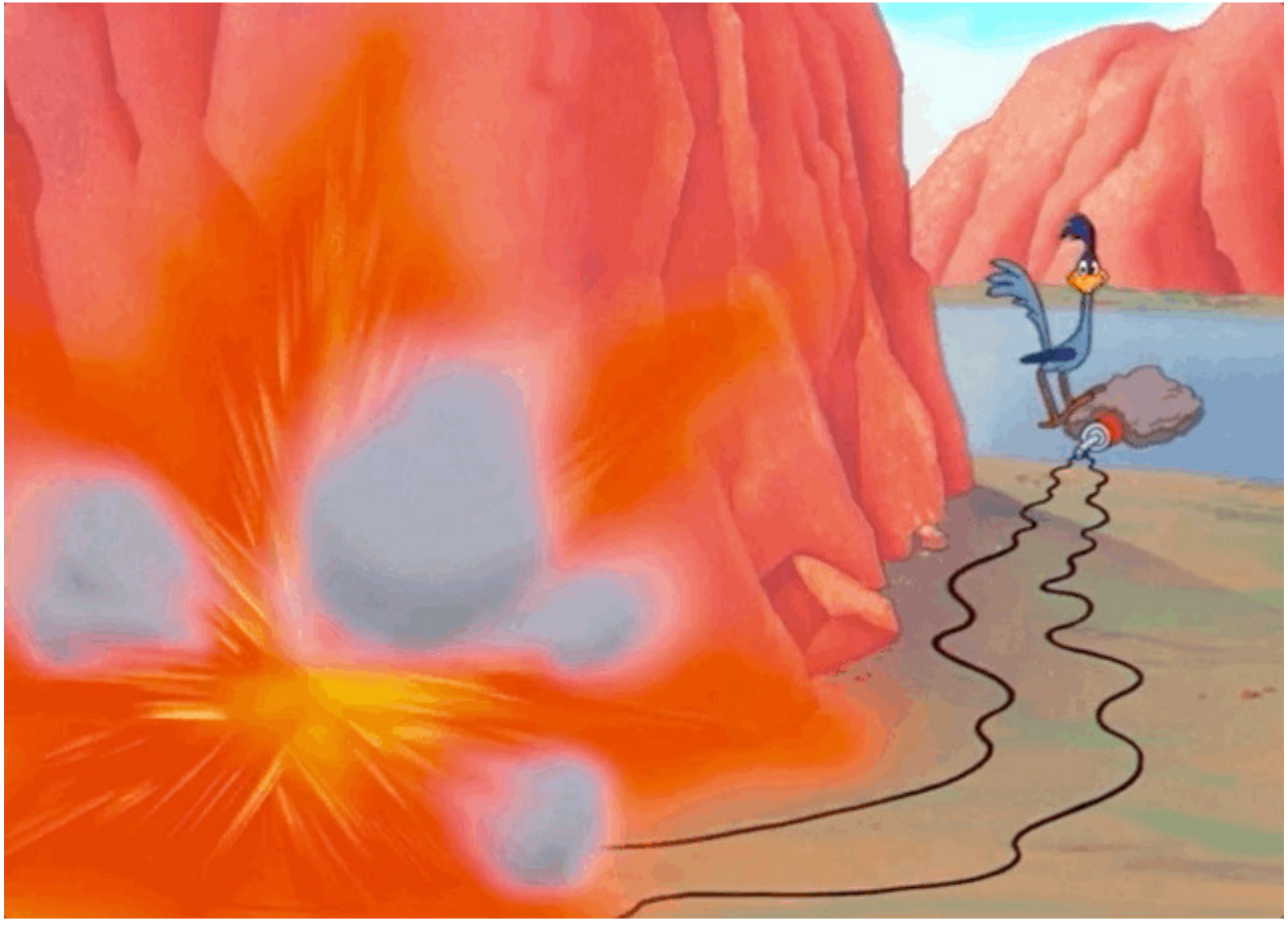
# Barplot with error bars - a.k.a. gráfico de dinamita

*Gráfico de... ¿dinamita?*



# Barplot with error bars - a.k.a. gráfico de dinamita

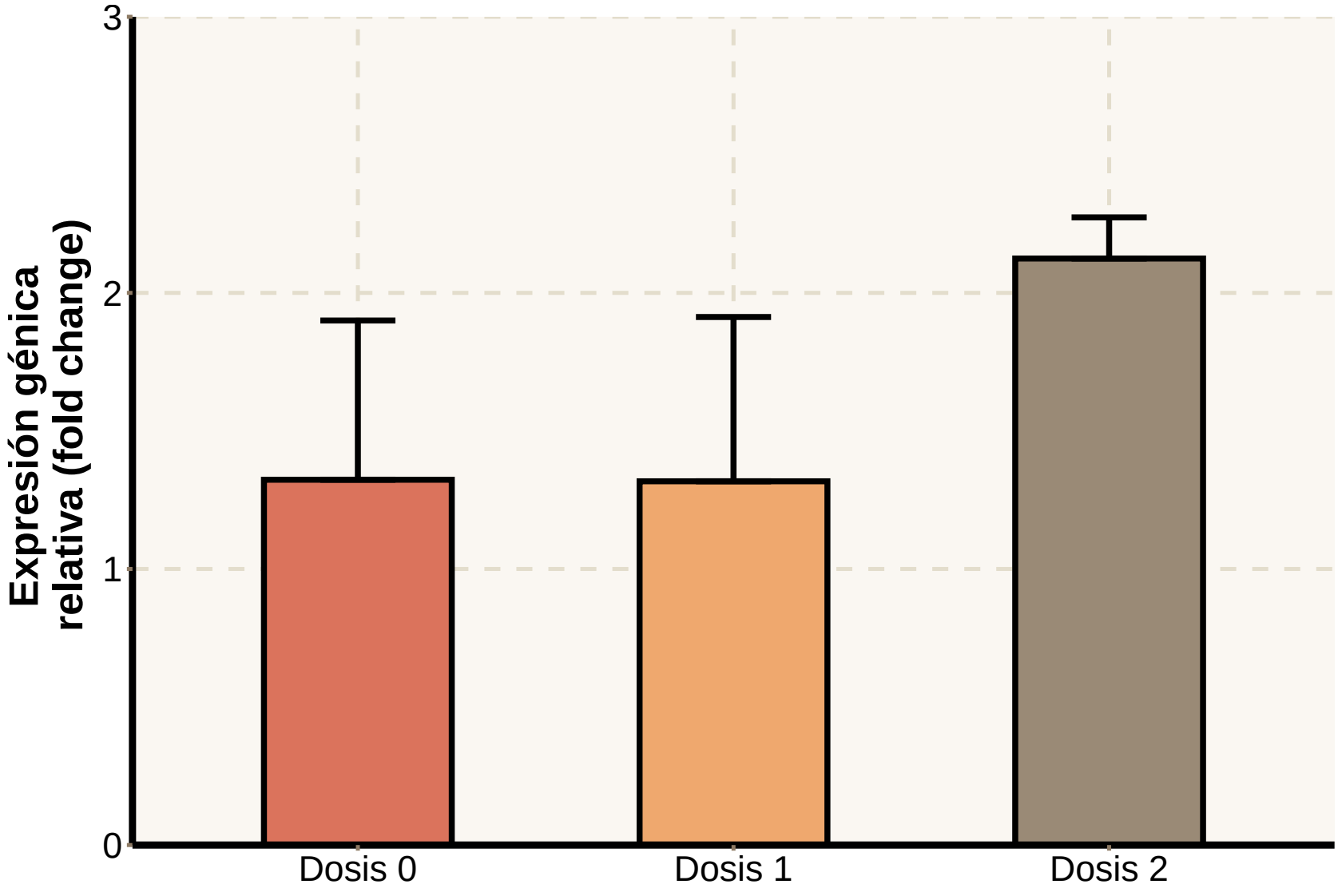
*Gráfico de... ¿dinamita?*



# Barplot with error bars - a.k.a. gráfico de dinamita

*Gráfico de... ¿dinamita?*

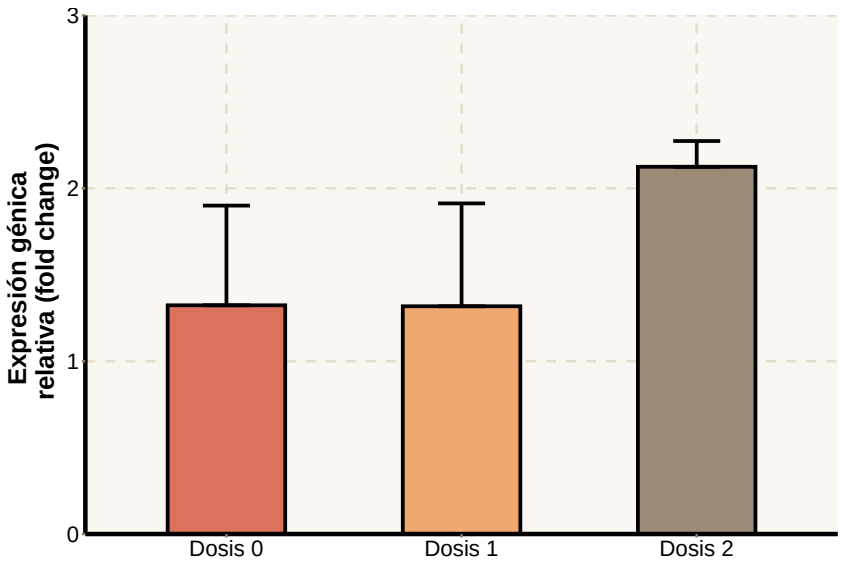
# Barplot with error bars - a.k.a. gráfico de dinamita



# Barplot with error bars - a.k.a. gráfico de dinamita

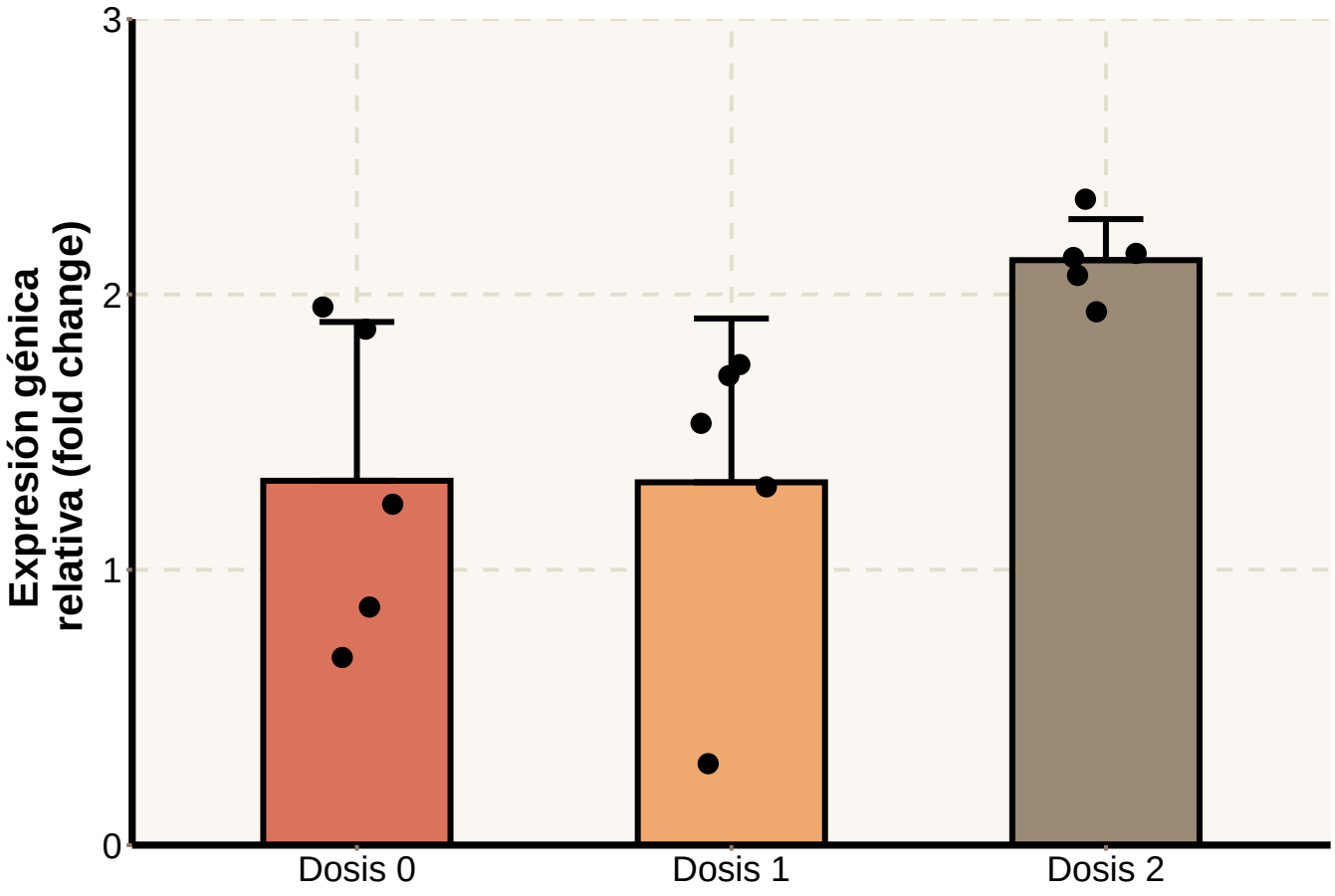
Al elegir este tipo de gráfico, no estamos mostrando más información que la que podríamos haber encontrado en nuestra tabla de estadísticas descriptivas.

| Condición | Media | DE |
|---|---|---|
| Dosis 0 | 1.3 | 0.58 |
| Dosis 1 | 1.3 | 0.60 |
| Dosis 2 | 2.1 | 0.15 |



# Barplot with error bars - a.k.a. gráfico de dinamita

Incorporar los valores observados nos brinda más información acerca de la **distribución** de los datos.



# ¿Qué son y qué hacemos con los *outliers*?

Los **outliers** son observaciones que no son típicas del conjunto de datos que estamos analizando.

**¿Por qué es importante detectarlos?** -> Porque pueden influir sobre los resultados de un análisis estadístico clásico, en tanto gran parte de las técnicas utilizadas habitualmente son muy sensibles a la presencia de estos valores.

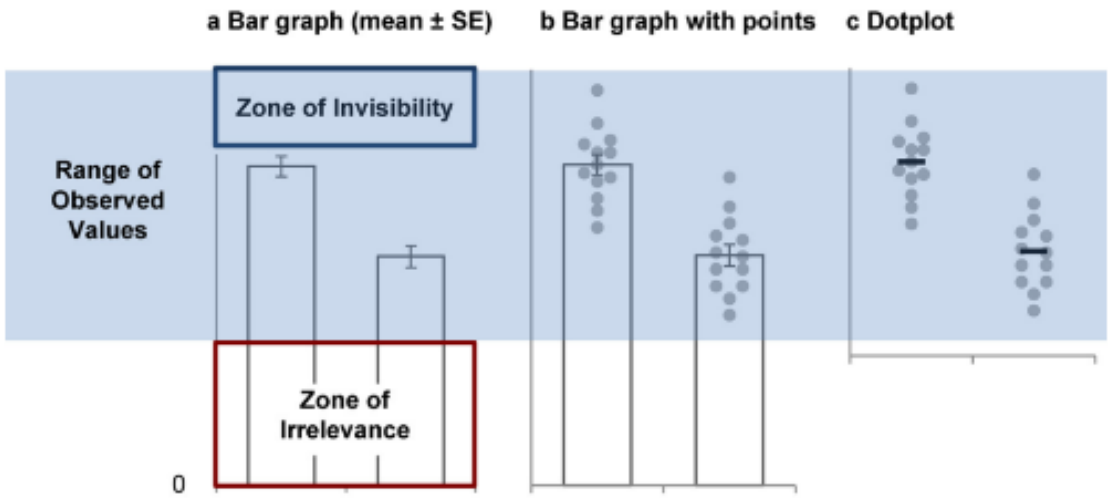
Los outliers deben ser cuidadosamente inspeccionados. Si no hay evidencia de error y su valor es posible, no deberían ser eliminados.

# Redflags del gráfico de dinamita

**Redflag #1** Los gráficos de barras están diseñados para representar **conteos y proporciones** relacionados al trabajo con **variables categóricas**. No obstante, continúan siendo una estrategia ampliamente aceptada para presentar información sobre variables cuantitativas.

# Redflags del gráfico de dinamita

**Redflag #2** En un gráfico de dinamita se reconocen **dos regiones** que, potencialmente, pueden generar confusión y malas interpretaciones: una **región de irrelevancia** y una **región de invisibilidad**.




Adaptado de Weissberger *et al.* (2017)[2].

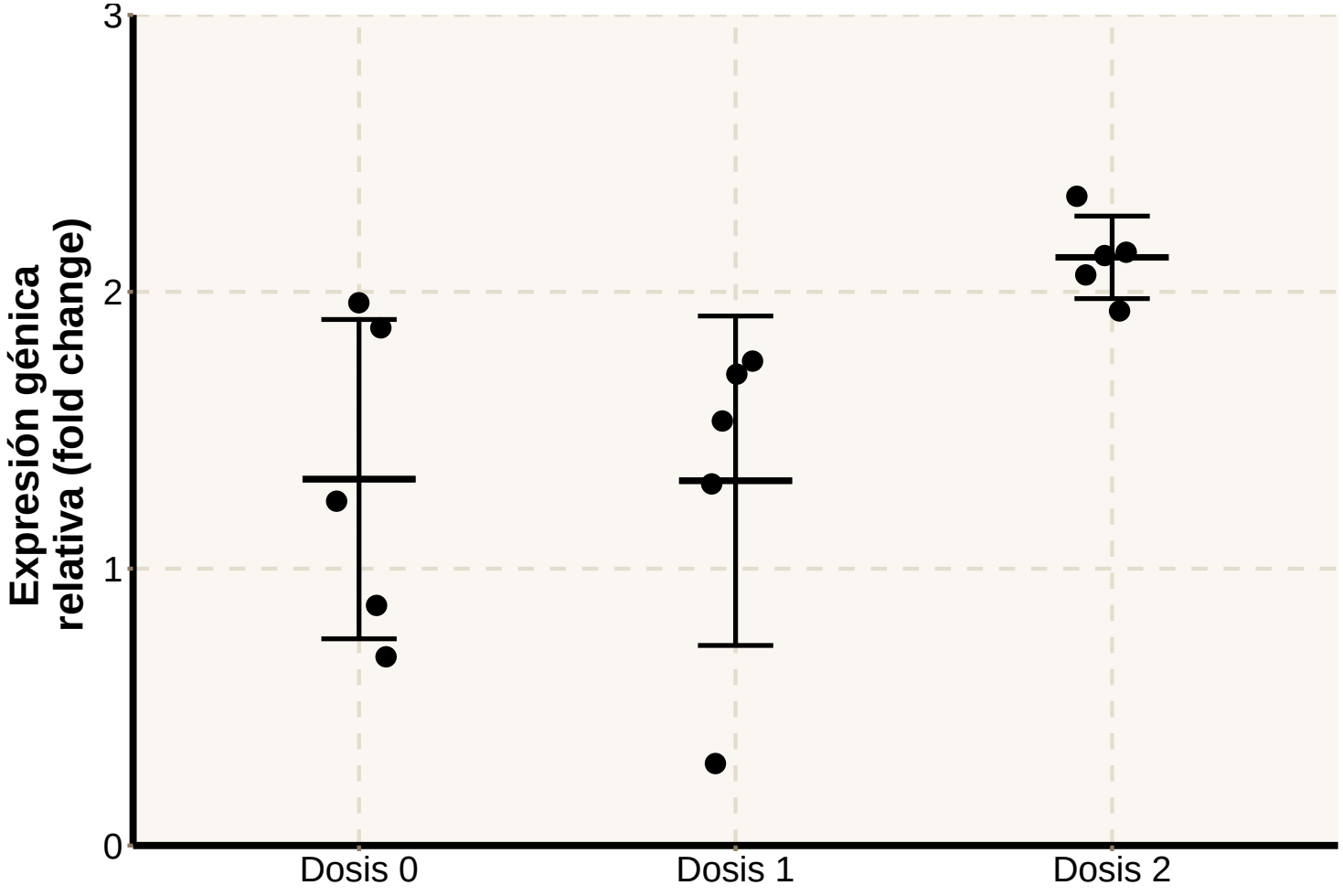
[2]Weissberger, T. L., Savic, M. *et al.* (2017) **Data visualization, bar naked: A free tool for creating interactive graphics.** *J. Biol. Chem.* 292(50) 20592-20598.

# Redflags del gráfico de dinamita

**Redflag #3** Si nos quedamos únicamente con una representación gráfica de la media y el desvío estándar de cada conjunto, nos estamos perdiendo **mucha** información valiosa sobre nuestros datos y la posibilidad de evaluarlos críticamente.

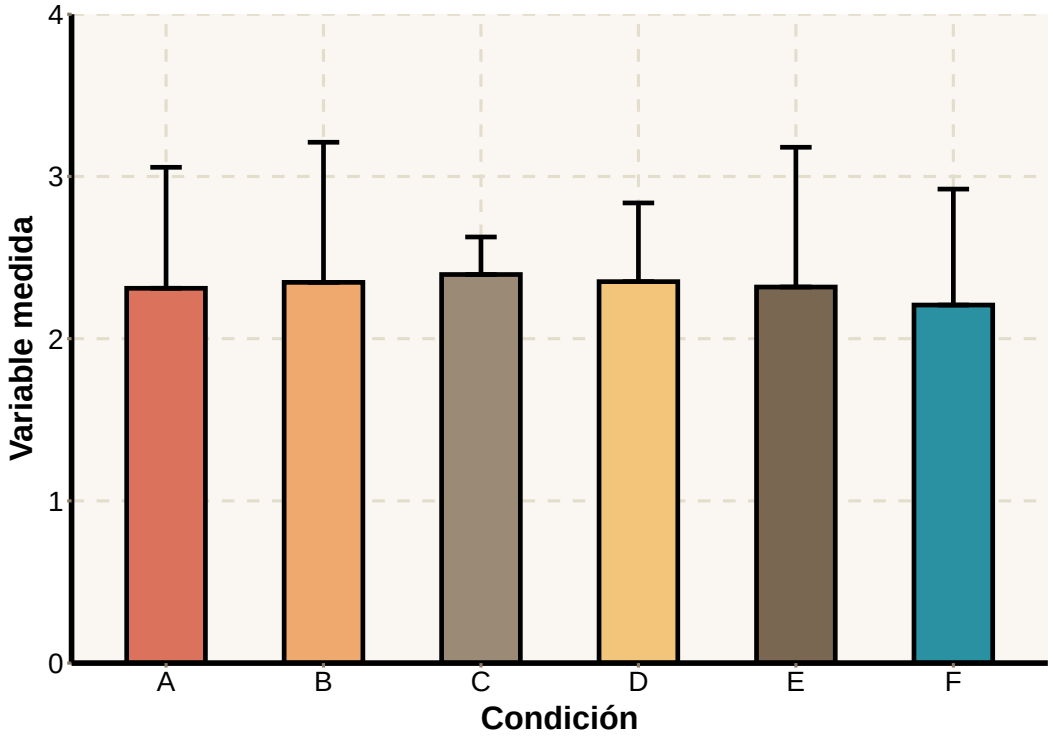
# Una alternativa al gráfico de dinamita (pocos datos)

Representamos media $\pm$ DE como intervalo y superponemos las observaciones individuales con un jitter plot:



# Más datos... y más gráficos de dinamita

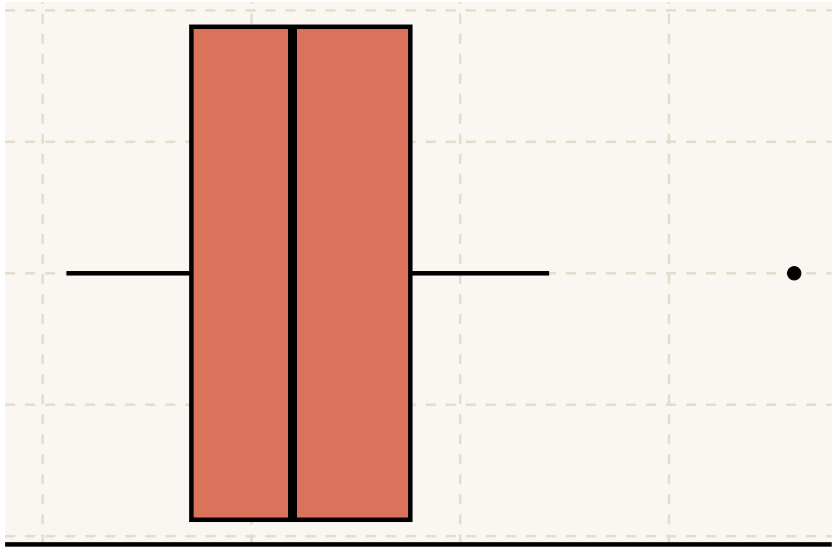
Supongamos que a partir de otros experimentos obtenemos un dataset mayor, y que al graficar nuestros resultados **utilizando un gráfico de dinamita** observamos lo siguiente:




¿Qué podemos concluir a partir de esta representación visual?

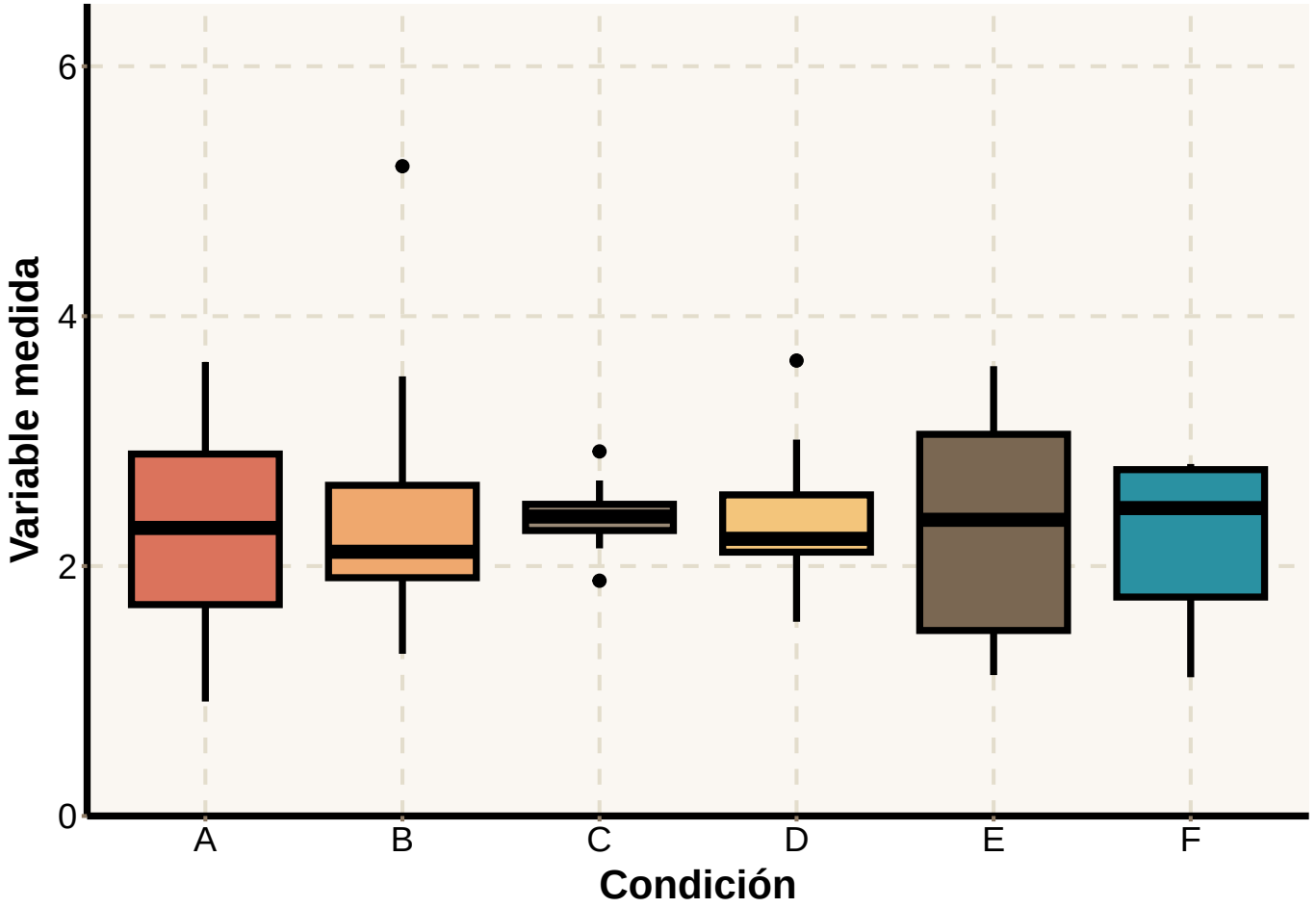
# Un gráfico estrella: el boxplot

Los cuartilos y los valores observados mínimo y máximo conforman un conjunto de cinco números que brindan un buen resumen de nuestros datos y con los cuales podemos construir un gráfico llamado **boxplot o diagrama de caja.**



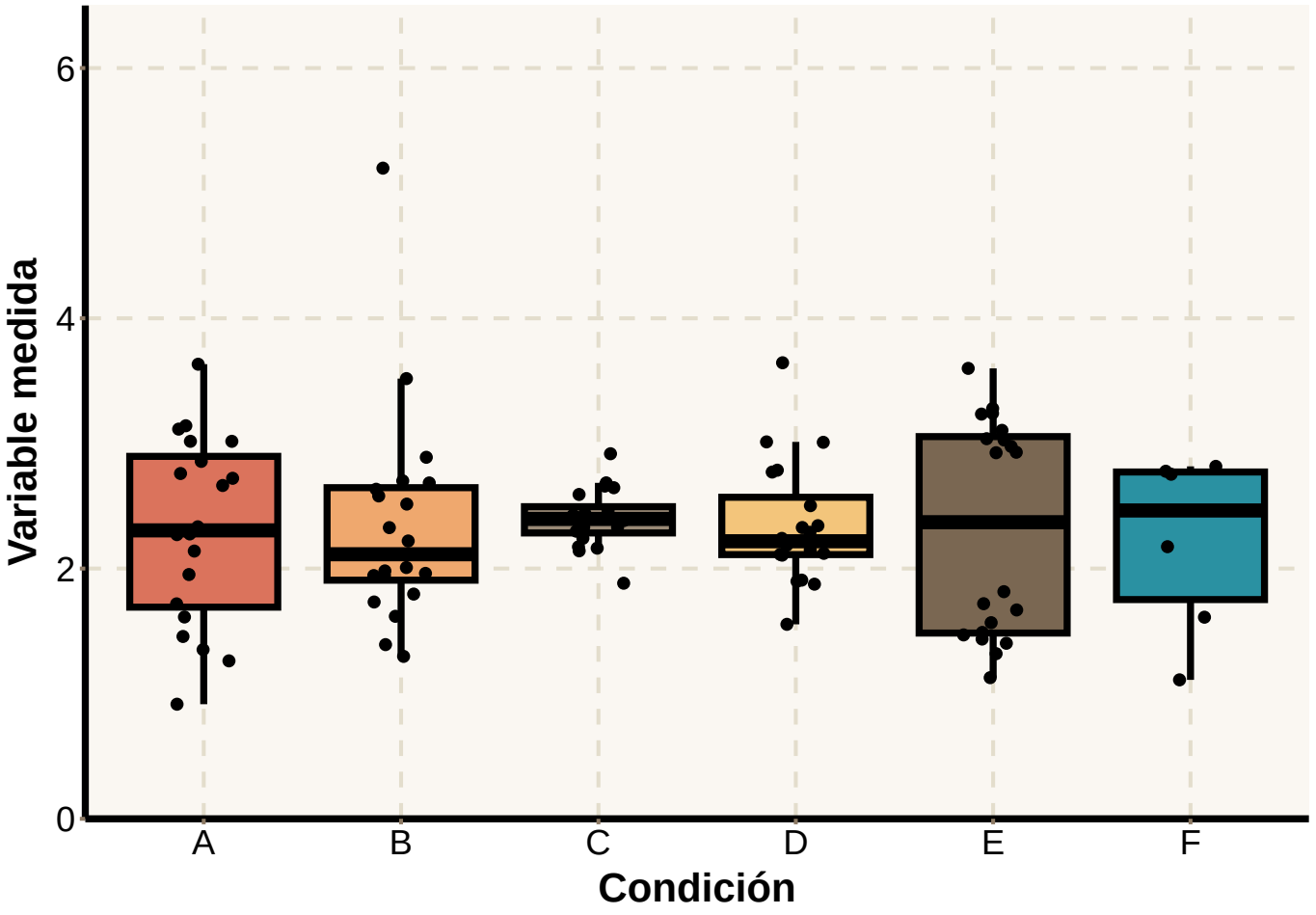
Una versión modificada del boxplot permite detectar potenciales *outliers*.

# Un boxplot múltiple para nuestros datos



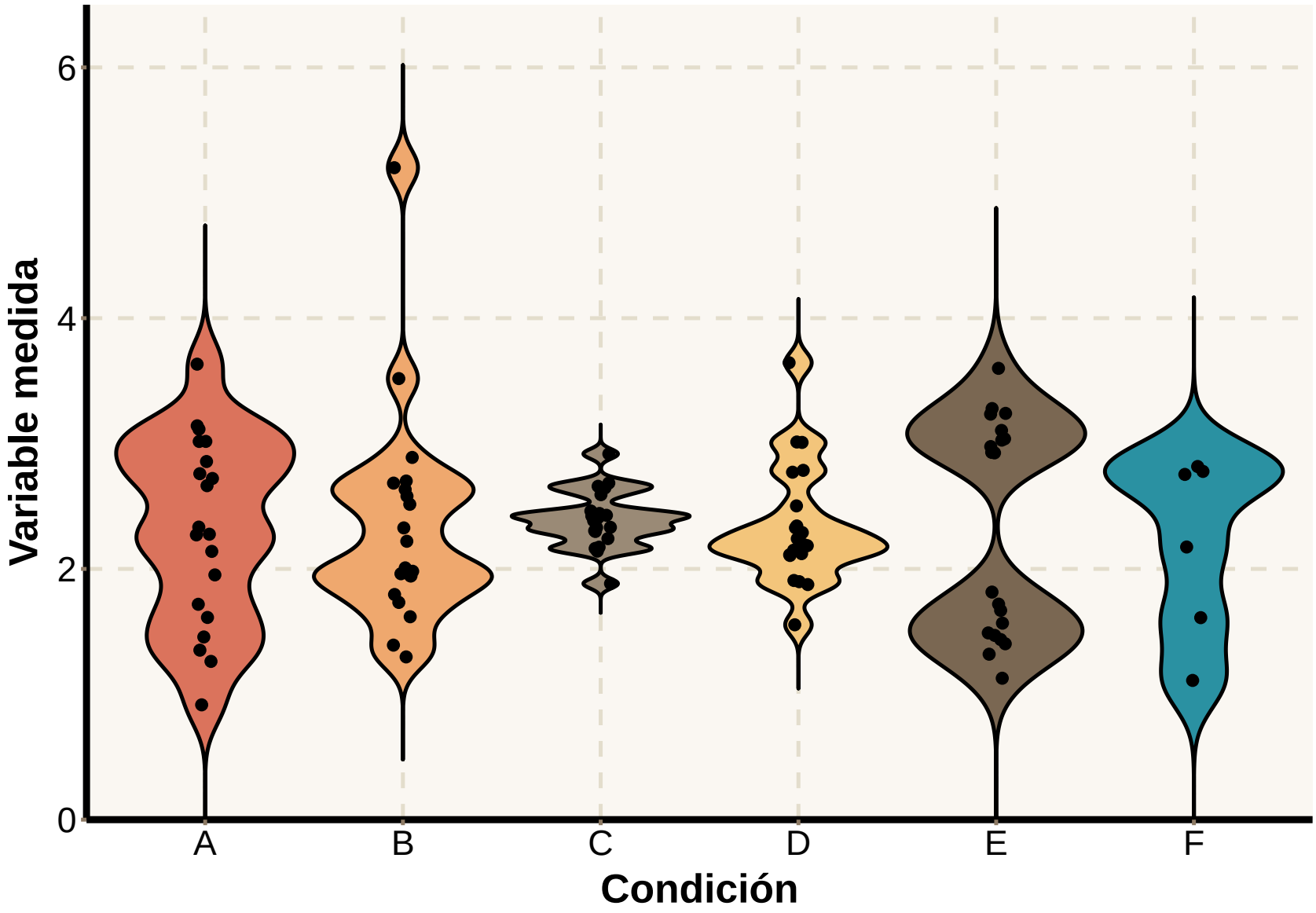

¿Qué información tenemos ahora acerca de los datos recolectados?
¿Qué preguntas seguimos sin poder responder?

# Superponemos el jitter plot...



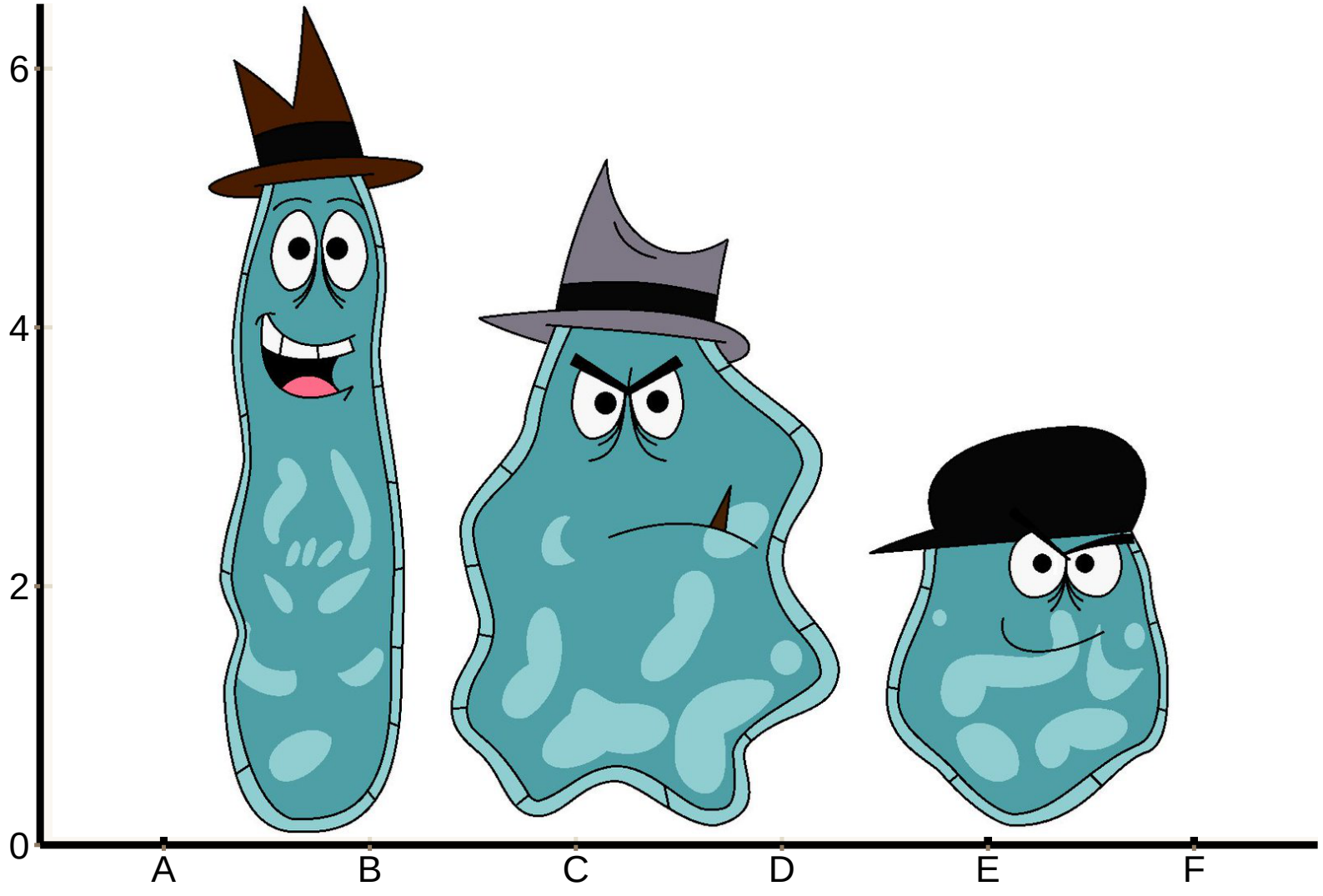

Agregando los valores observados podemos identificar distribuciones bimodales, diferencias en los tamaños muestrales, etc.

# Violin plots: una alternativa a los boxplots



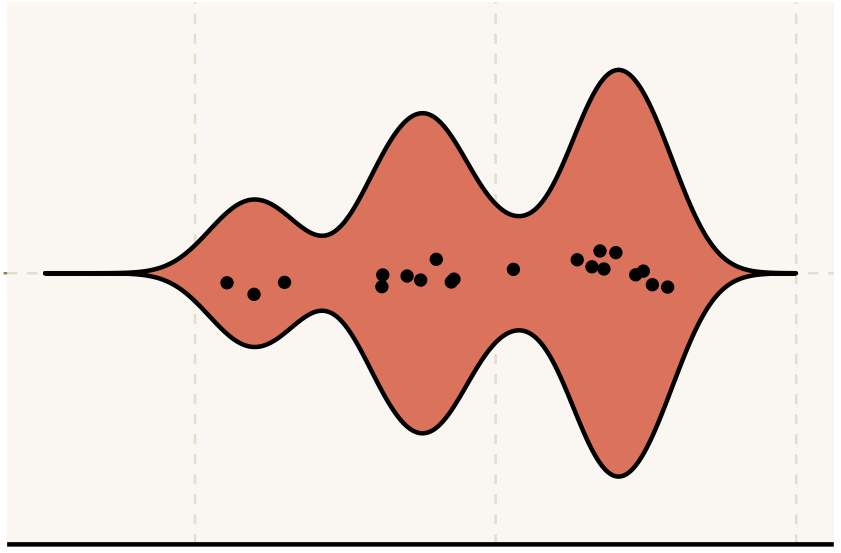
# Violin plots: una alternativa a los boxplots

*¿Ameba plots?*



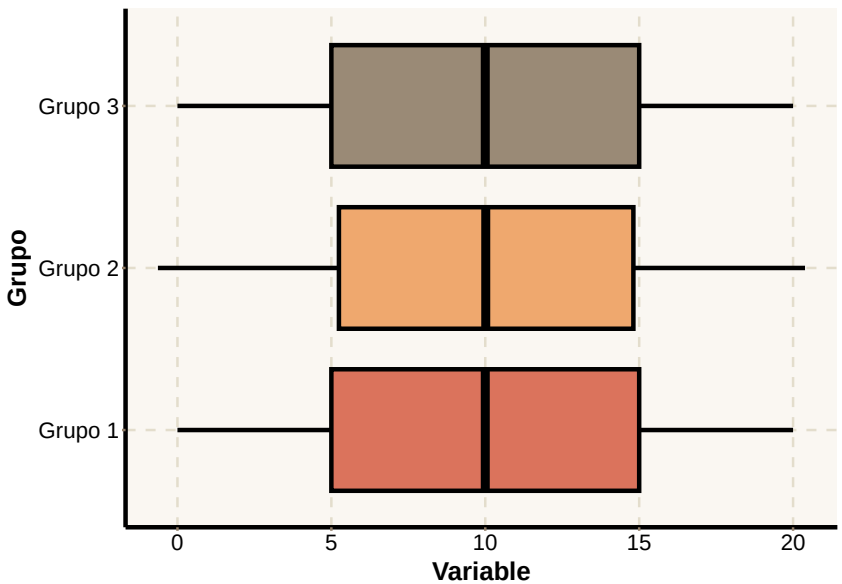
# Violin plots: una alternativa a los boxplots

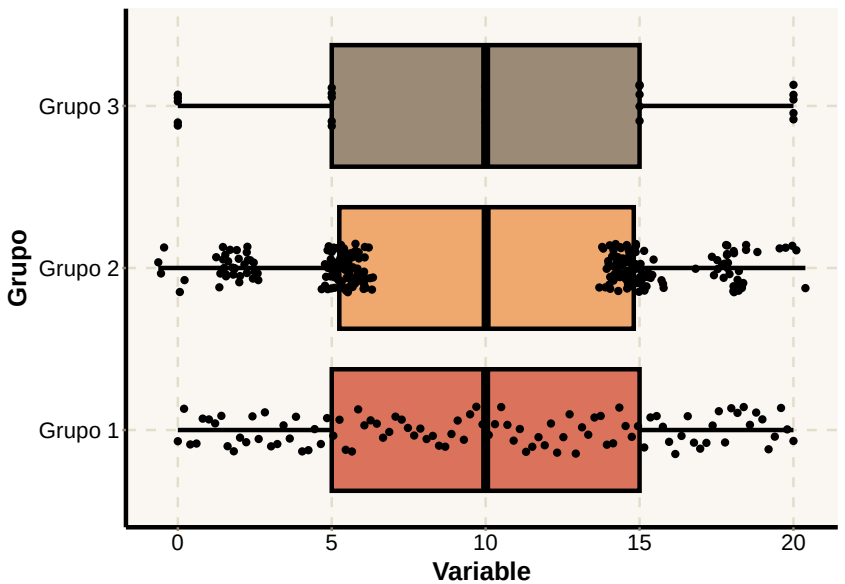
Los **violin plot o diagramas de violín** son gráficos de densidad de probabilidad espejados. Permiten visualizar fácilmente la distribución de conjuntos de muchos datos. Se les pueden agregar los valores observados, un boxplot, media y barras de error, etc. No son útiles cuando tenemos pocos datos.



# No hay gráfico que cuente la historia completa[3]

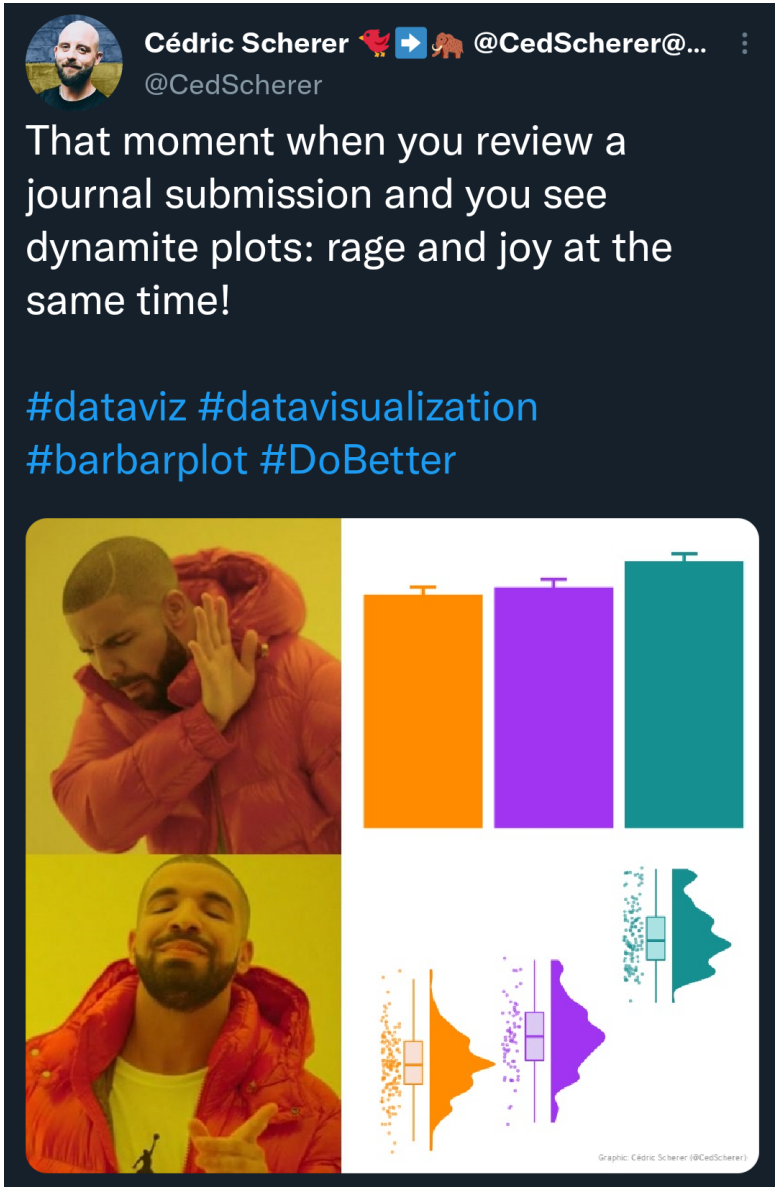
A pesar de que los boxplots para los tres grupos lucen idénticos, superponer un Jitter plot con las observaciones individuales nos revela la existencia de tres patrones radicalmente diferentes en los datos que los originan:




[3]Scherer, C. (2021). **Visualizing Distributions with Raincloud Plots (and How to Create Them with ggplot2)**. Recuperado de: https://www.cedricscherer.com/2021/06/06/visualizing-distributions-with-raincloud-plots-and-how-to-create-them-with-ggplot2/

# Hay muchas alternativas. . .



# Apliquemos lo aprendido

**Ejemplo**

Pusimos a punto un método para determinar la actividad enzimática de una enzima de interés en extractos celulares por fluoresencia. El último paso del ensayo consiste en formar un aducto fluorescente entre el producto de reacción y un reactivo adecuado por un tiempo suficiente. Deseamos evaluar si hay diferencia en la actividad medida al formar el aducto durante 30 o 60 minutos. Para eso, trabajamos con 10 muestras y medimos la actividad en cada una de ellas luego de 30 y 60 minutos de reacción de formación del aducto.

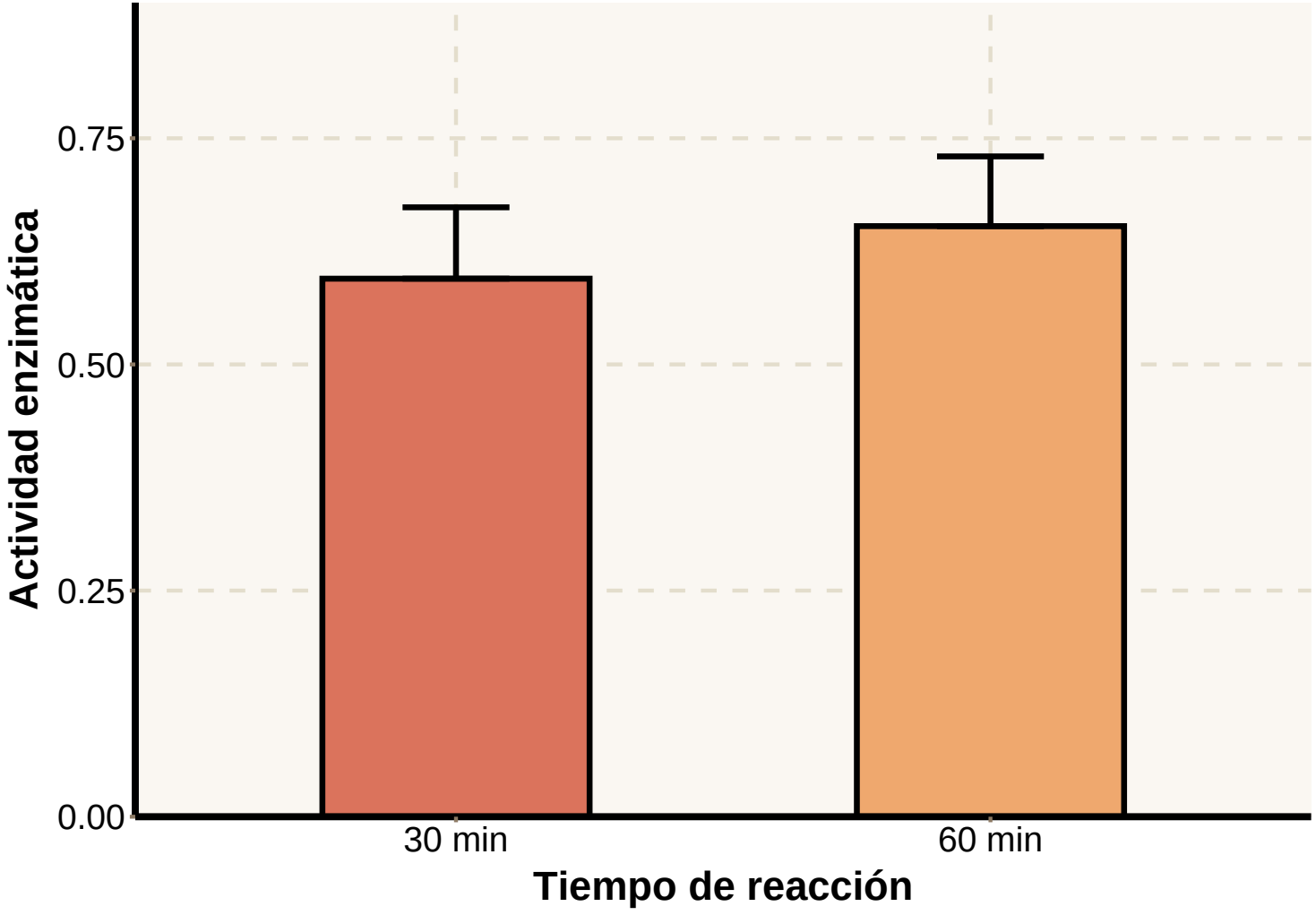
# Apliquemos lo aprendido

Estos son los resultados obtenidos (en $\mu$moles sustrato/mg proteína):

| | Actividad enzimática | |
|---|---|---|
| Muestra | 30 min | 60 min |
| 1 | 0.63 | 0.67 |
| 2 | 0.61 | 0.69 |
| 3 | 0.52 | 0.61 |
| 4 | 0.54 | 0.59 |
| 5 | 0.43 | 0.49 |
| 6 | 0.63 | 0.72 |
| 7 | 0.64 | 0.73 |
| 8 | 0.61 | 0.62 |
| 9 | 0.73 | 0.74 |
| 10 | 0.61 | 0.67 |

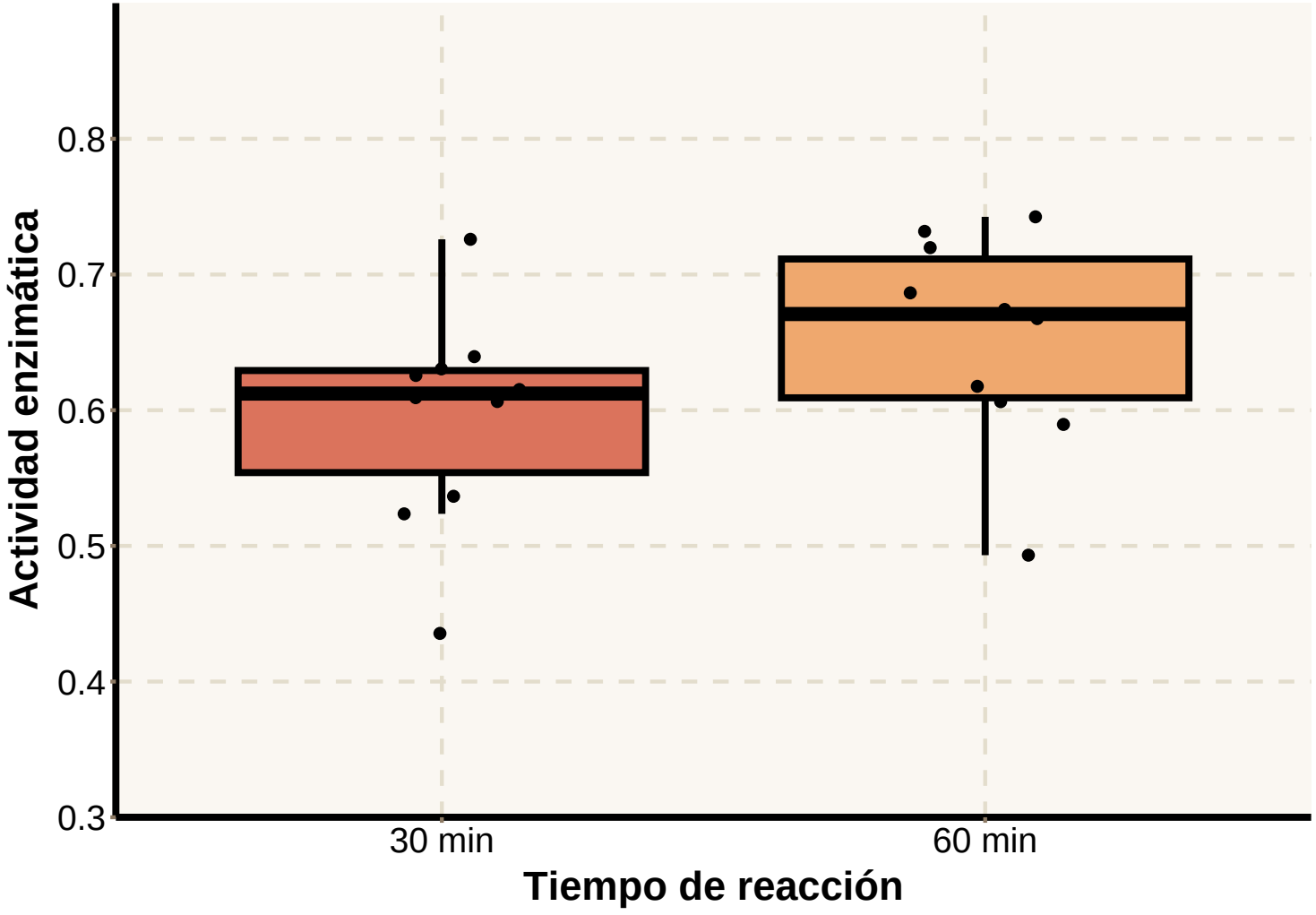
# Apliquemos lo aprendido

Insistimos un poquito más con los gráficos de dinamita (¿por qué no?)

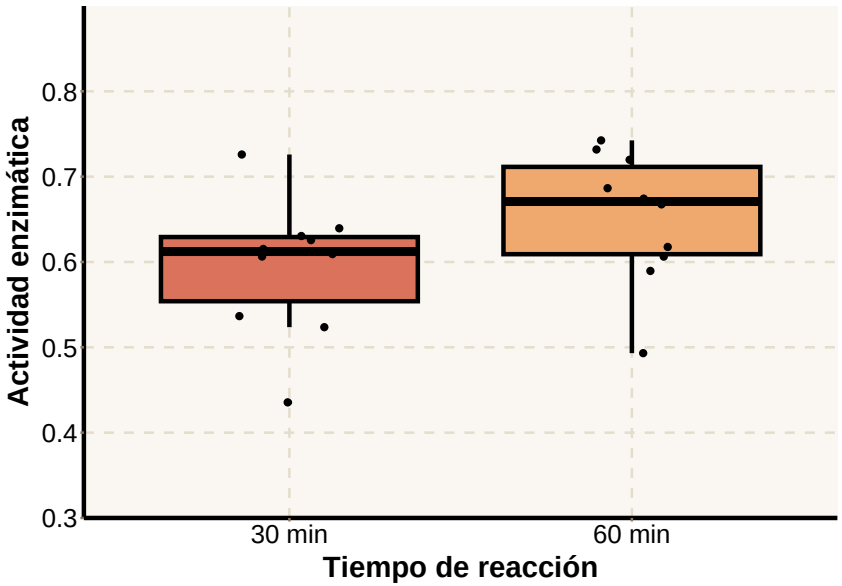


# Apliquemos lo aprendido

¿Por qué no probamos mejor con un boxplot múltiple + jitter plot?



# Apliquemos lo aprendido




```
## Two sample t-test

## p-value = 0.11
```

En línea con el gráfico construido, si quisiéramos evaluar si existen diferencias estadísticamente significativas en la actividad enzimática promedio entre los dos tiempos de reacción de interés, utilizaríamos un **test-t para la comparación de dos promedios en base a muestras independientes.**

¿Cuál sería nuestra conclusión en este caso?

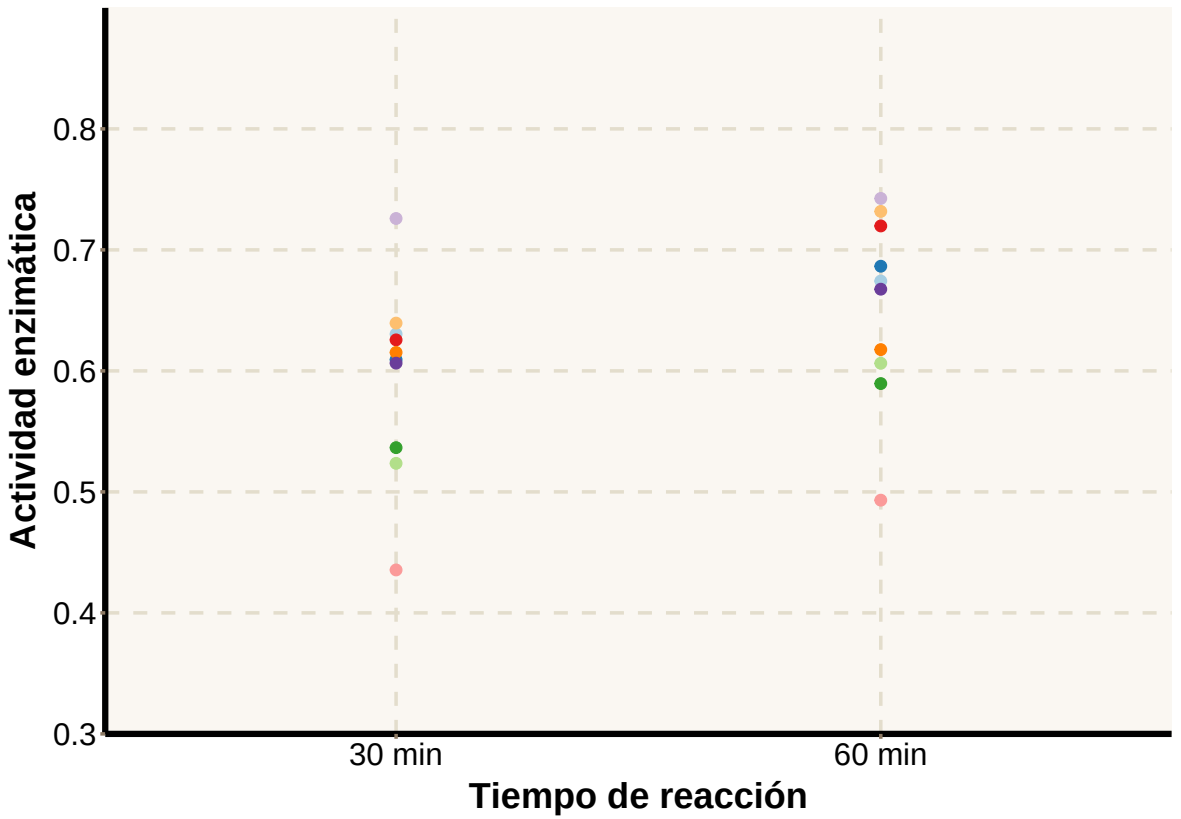
# La importancia del diseño experimental

¿De qué nos estamos olvidando? El gráfico debe contar también la historia del diseño experimental utilizado, es decir, *cómo fueron obtenidos esos datos*.

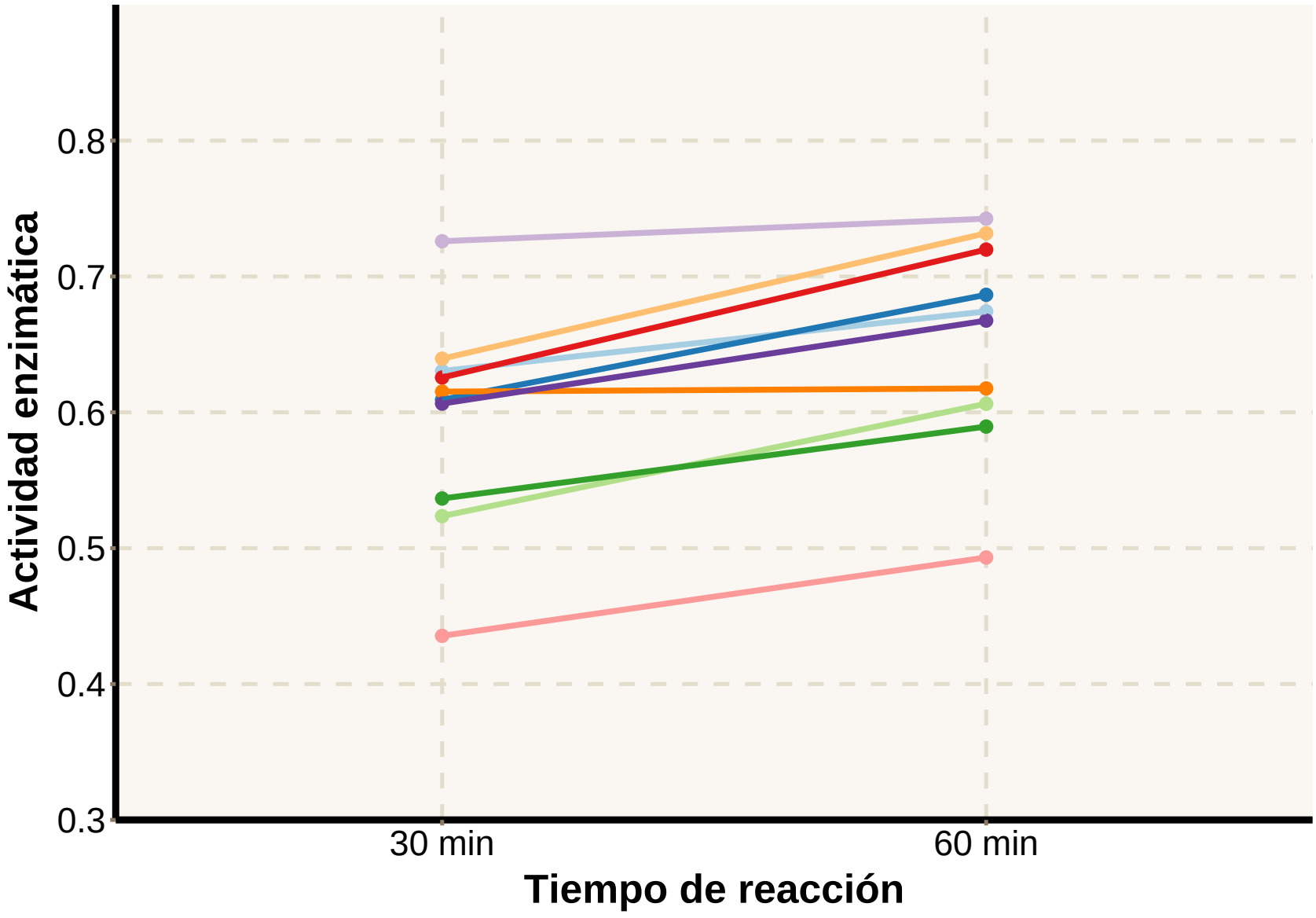
Recordando el ejemplo, las muestras estaban **pareadas**: a la misma muestra se le medía la actividad luego de 30 y 60 minutos de reacción. Si existe un *efecto muestra*, lo estamos obviando tanto en el gráfico como en el test.

# La importancia del diseño experimental

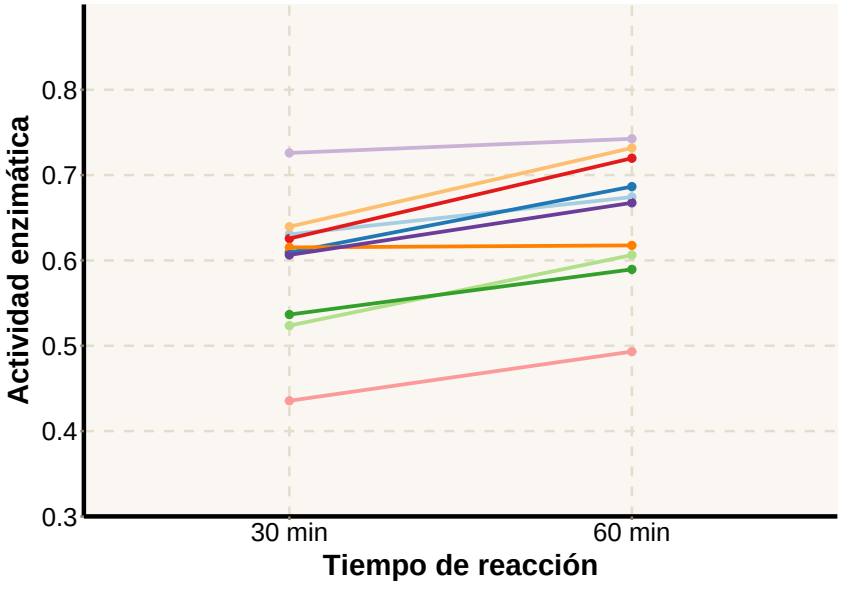
¿Qué pasa si agregamos la información de la muestra a un gráfico adecuado? (*Los distintos colores representan a las distintas muestras*)



# La importancia del diseño experimental



# La importancia del diseño experimental




```
## Two sample t-test

## p-value = 0.00021
```

Teniendo en cuenta ahora que trabajamos con **muestras pareadas**, el test adecuado para este caso es **test-t para la comparación de dos promedios en base a muestras dependientes.**

¿Cuál sería nuestra conclusión en este caso? ¿Es la misma a la que habíamos llegado anteriormente?

# ¿Qué aprendimos hasta acá?

- La importancia de visualizar.
- La elección de un gráfico adecuado.
- Cada gráfico tiene sus potencialidades y limitaciones.

# ¿Qué aprendimos hasta acá?

- En toda figura intervienen **cuestiones estéticas** y de **percepción visual**.

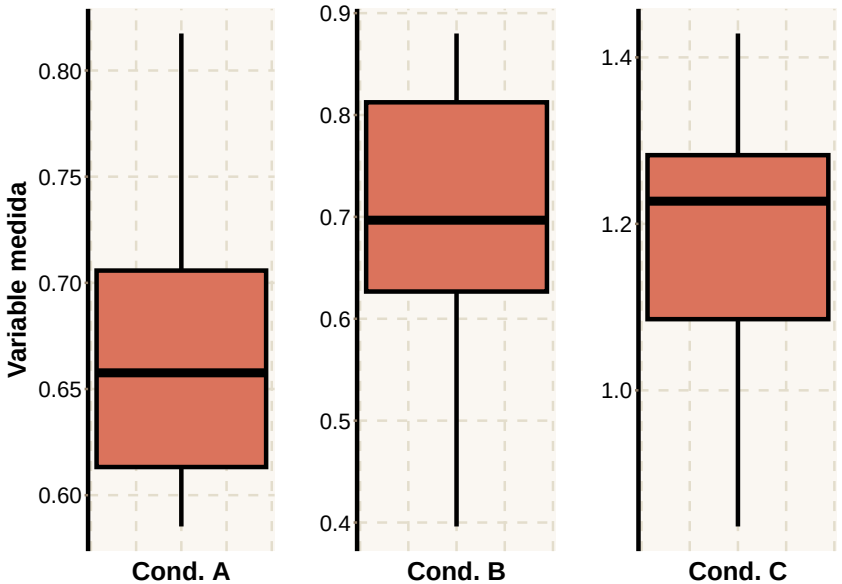
Un gráfico bastante, bastante feíto[4]:

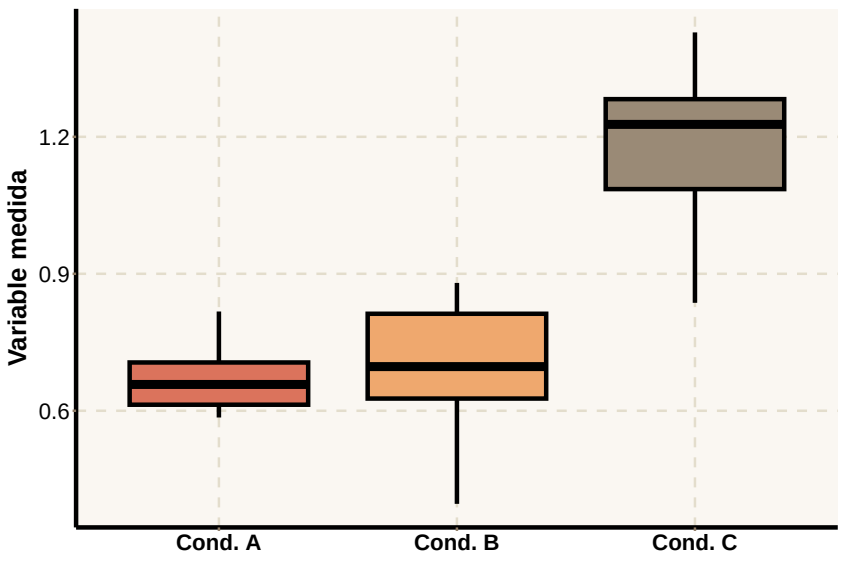

[4]Extraído de: Healy, K. (2019). **Data Visualization. A practical introduction.** Princeton University Press.

# ¿Qué aprendimos hasta acá?

- En toda figura intervienen **cuestiones estéticas** y de **percepción visual**.

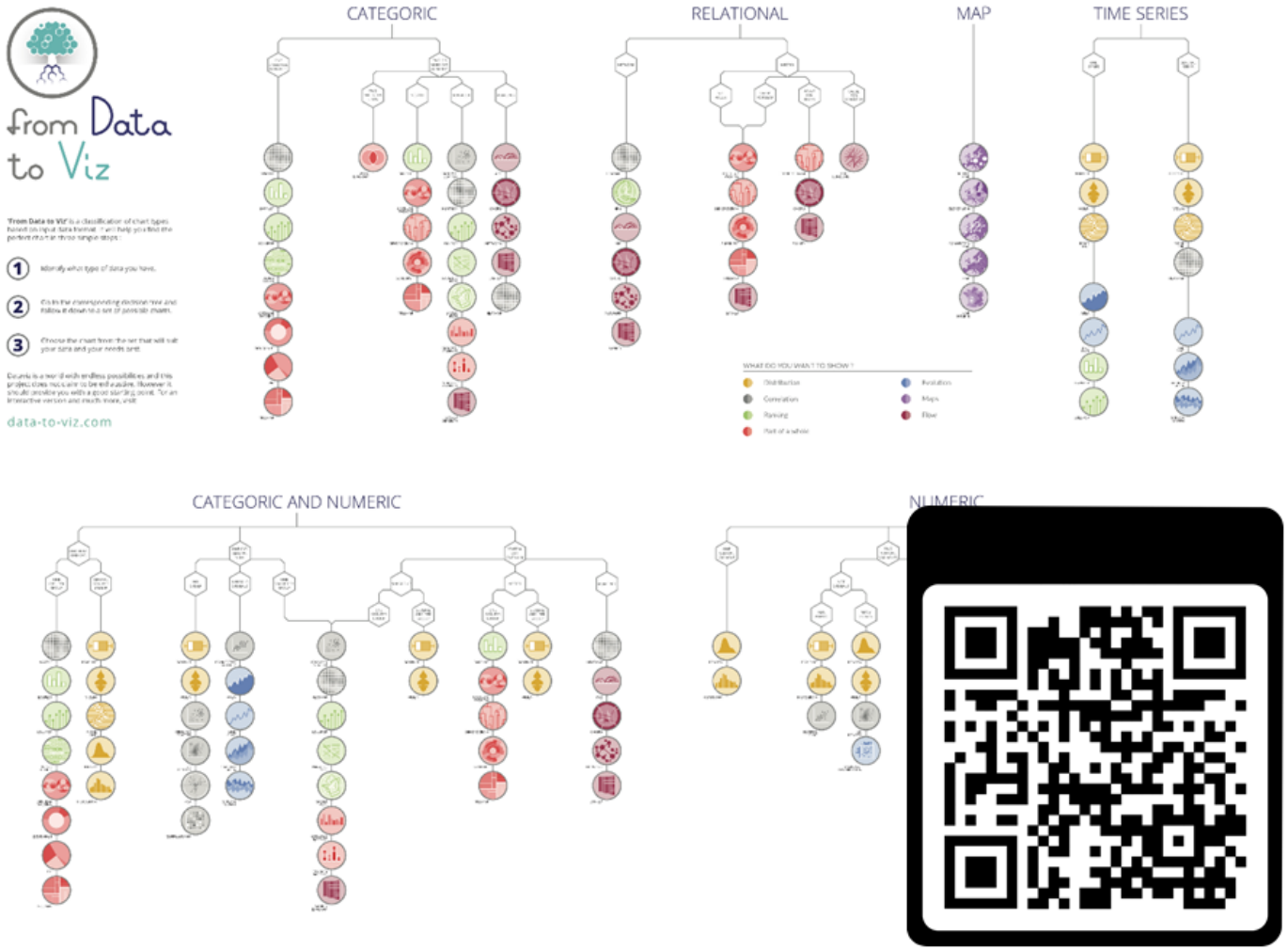
A pesar de estar construidos con los mismos datos, estos dos gráficos no parecen mostrar lo mismo:



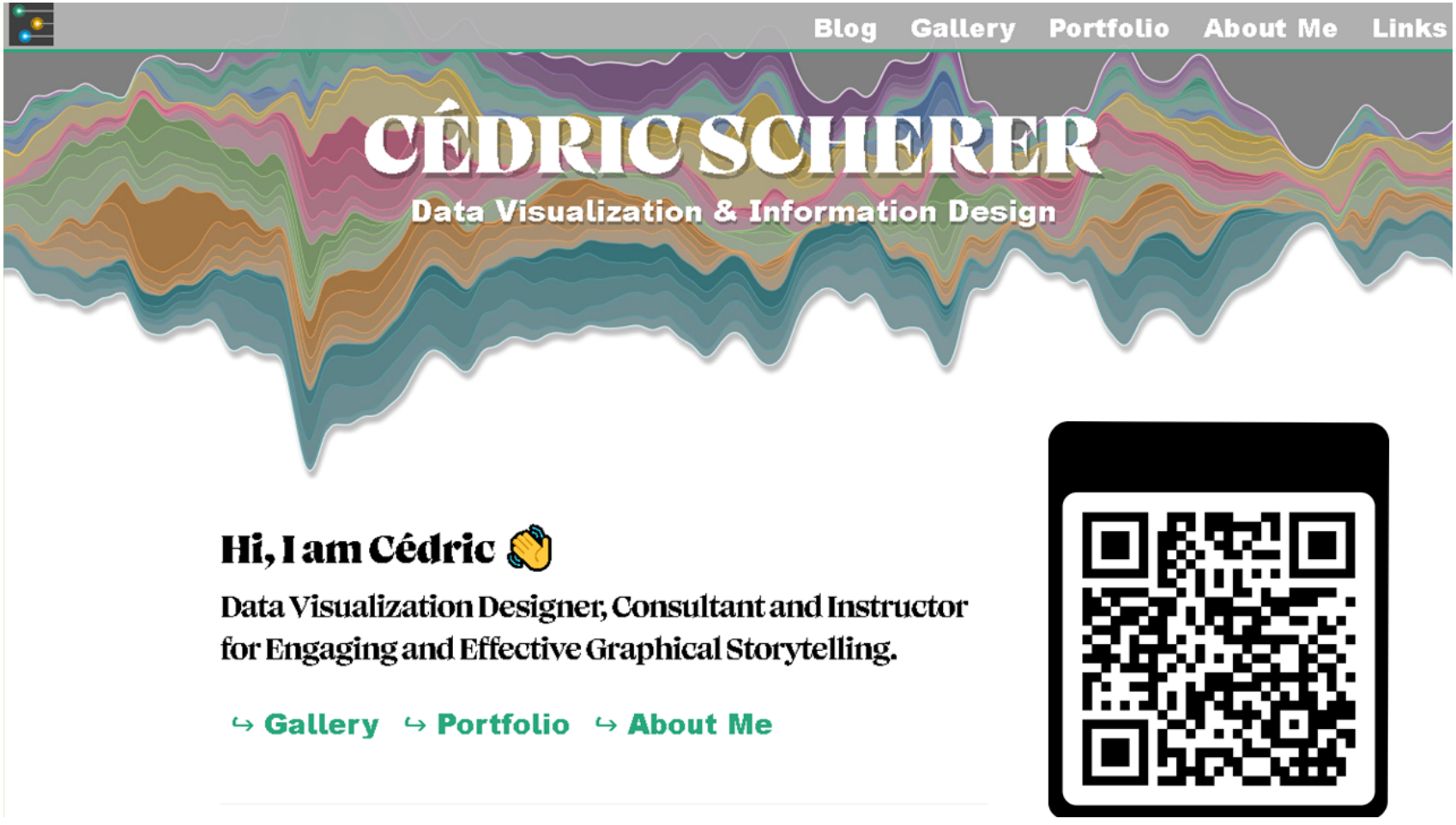
# Recursos interesantes

- **data-to-viz.com**: árboles de decisión que conducen desde un tipo particular de datos a un conjunto de gráficos posibles.



# Recursos interesantes

- **cedricscherer.com**: blog de Cédric Scherer, *data visualization designer.*



# Recursos interesantes

- Healy, K. (2019). **Data Visualization. A practical introduction.** Princeton University Press.
- Weissberger, T. L., Savic, M. *et al.* (2017) **Data visualization, bar naked: A free tool for creating interactive graphics.** *J. Biol. Chem.* 292(50) 20592-20598.
- Wong, B. (2012). **Visualizing biological data.** *Nat. Methods* 9, 1131.

# Agradecimientos


ÁREA ESTADÍSTICA Y PROCESAMIENTO DE DATOS
Facultad de Ciencias Bioquímicas y Farmacéuticas
Universidad Nacional de Rosario

**E-mails:** ferreyra@iquir-conicet.gov.ar - nlabadie@iquir-conicet.gov.ar