Diagramas de cajas (boxplots), de violín (Violin plots) y sinaplots

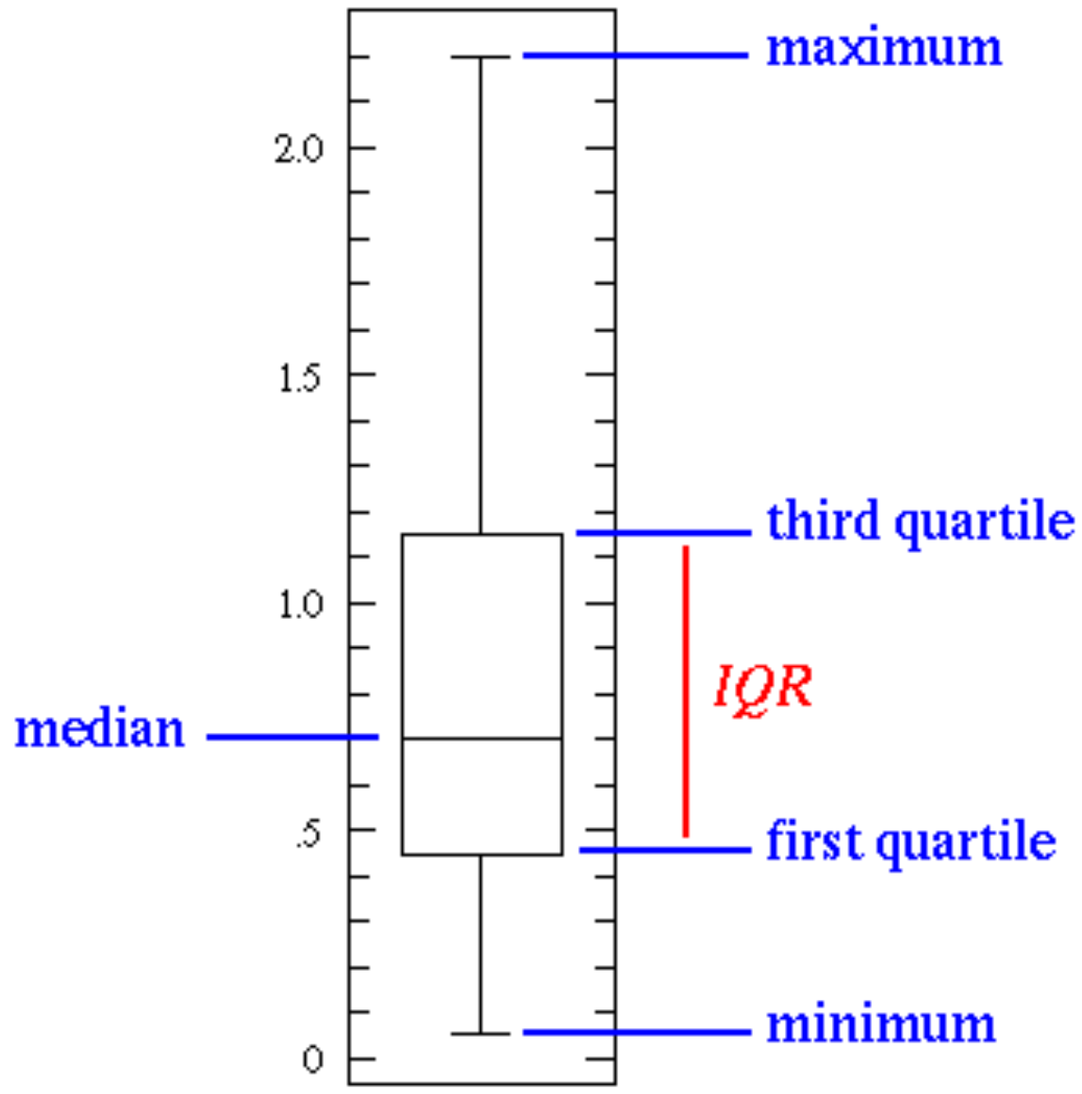
Visualización Científica
4 de octubre de 2016

Eduardo Castro-Nallar, PhD
Center for Bioinformatics and Integrative Biology
www.cbib.cl
www.castrolab.org

Boxplots o diagramas de caja

- Otra forma de visualizar una distribución

- Percentiles, quintiles, cuartiles, etc.



¿Qué es un percentil?

- El valor por debajo del cual un porcentaje de observaciones cae

- Percentil 20, es el valor donde el 20% de las observaciones están ubicadas

Table 1. Test Scores.

Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

¿Qué es un percentil?

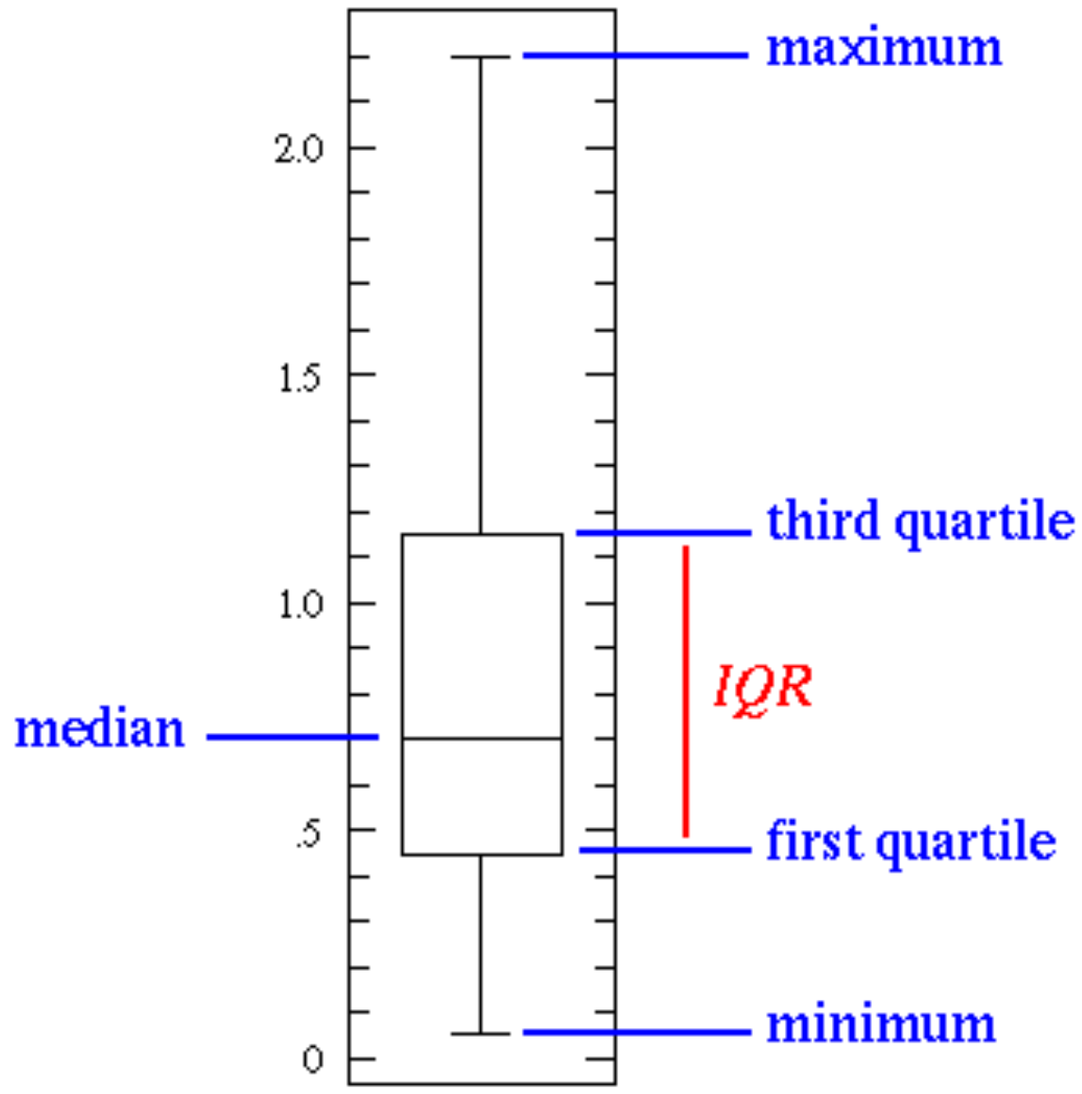
- Ejemplo con percentil 25
- R = P/100 x (N + 1)
- R = 25/100 x (8 + 1) = 9/4 = 2.25
- Entre 5 y 7
- 0.25*(7-5) + 5 = 5.5

Table 1. Test Scores.

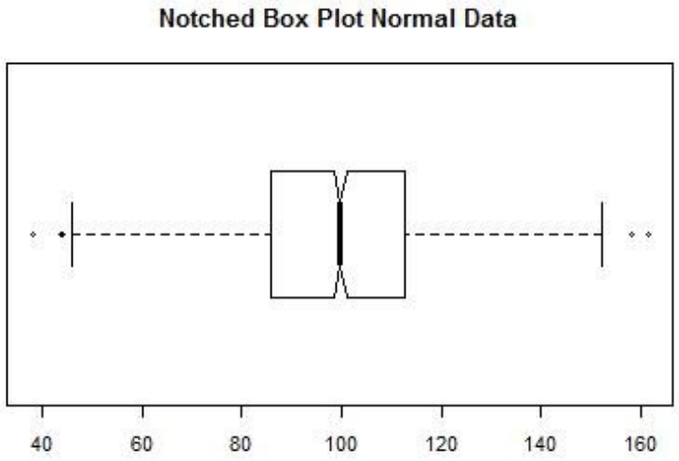
Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

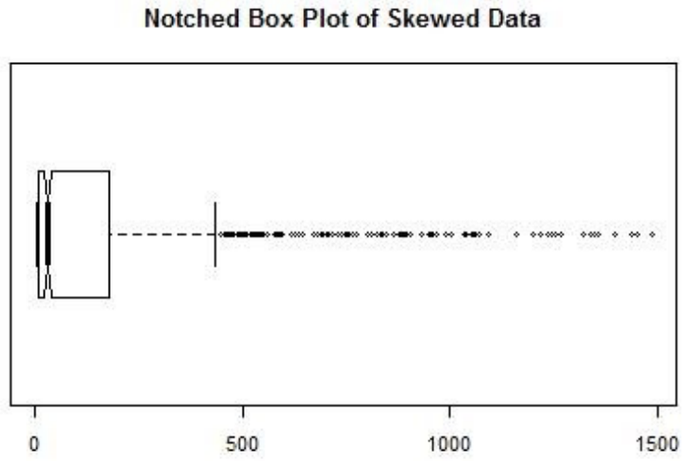
Boxplots o diagramas de caja

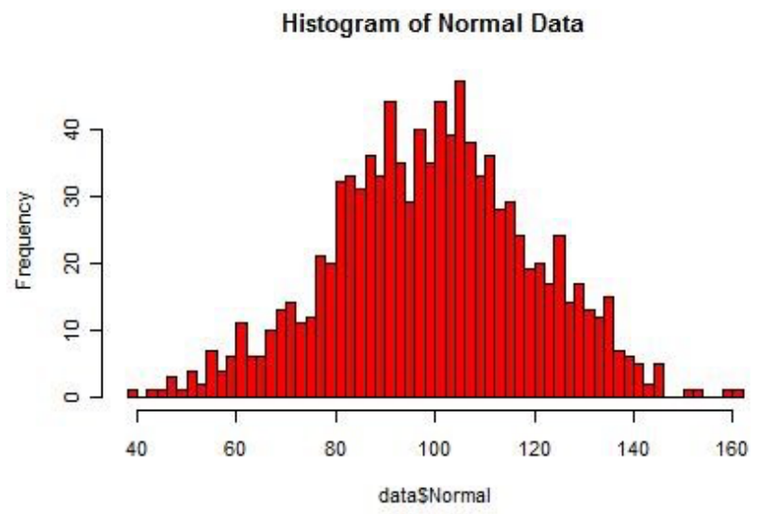
- Cuartil = 25avo de la distribución
- Quintil = un quinto de la distribución

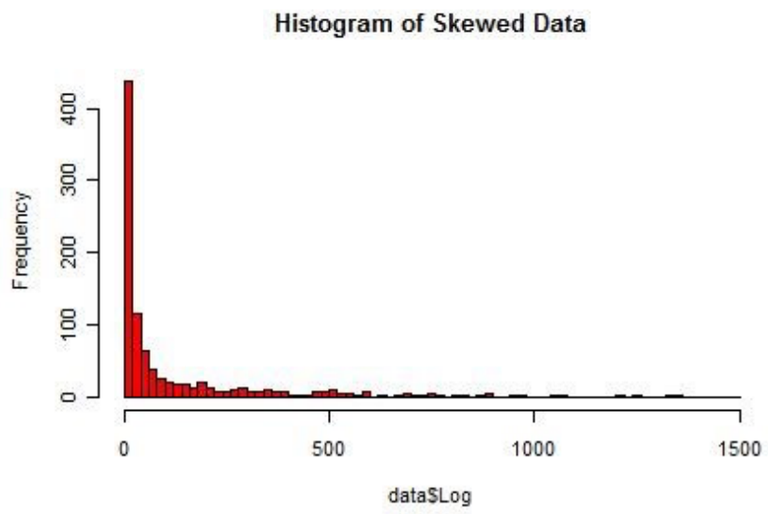


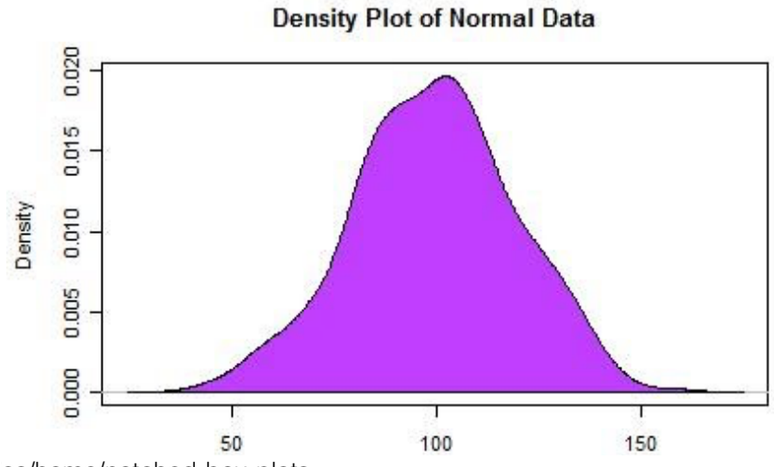
Boxplots - parientes del histograma y del gráfico de densidad

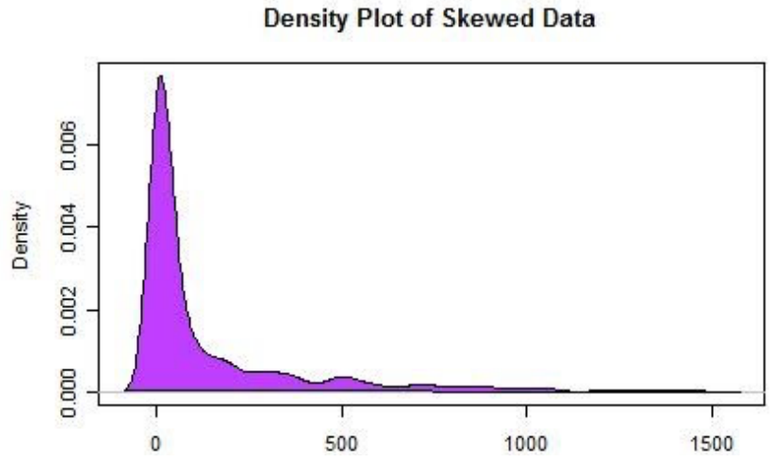






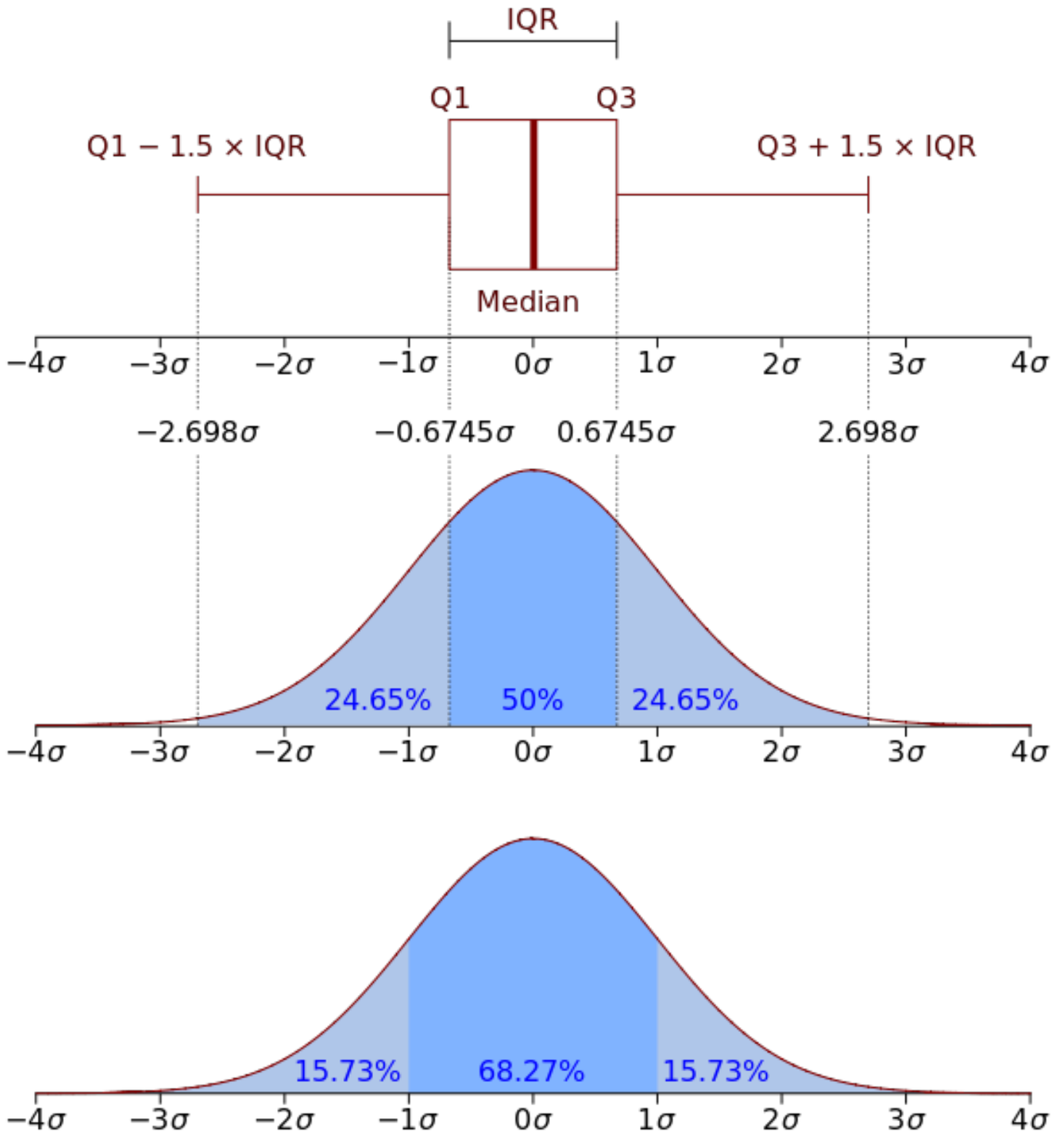






https://sites.google.com/site/davidsstatistics/home/notched-box-plots

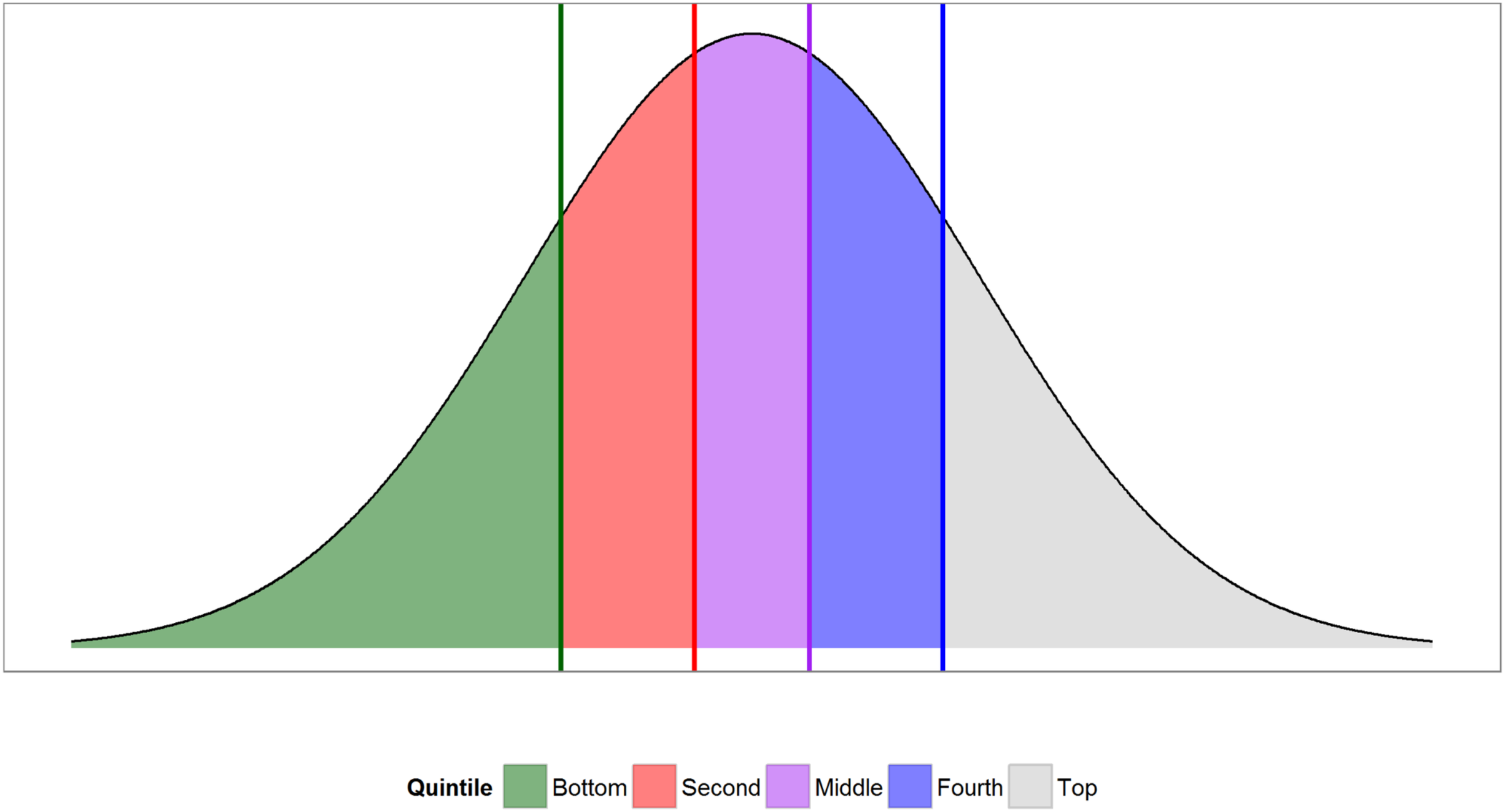
Boxplots


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

Gratuidad

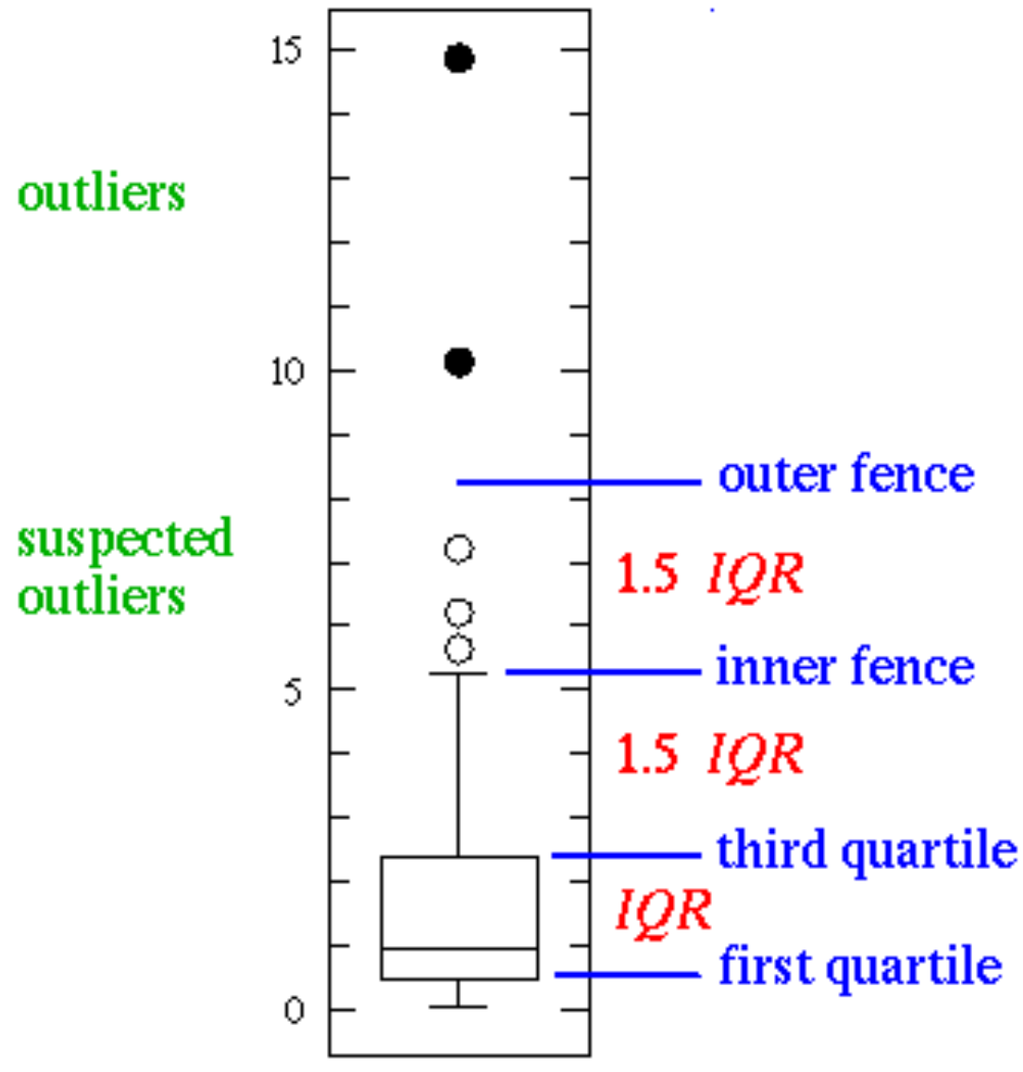
- Los requisitos para poder acceder a este beneficio es ser estudiante que ingresa a primer año de Educación Superior y pertenecer al primer, segundo o tercer **quintil** de ingreso per cápita

¿Qué es un quintil?

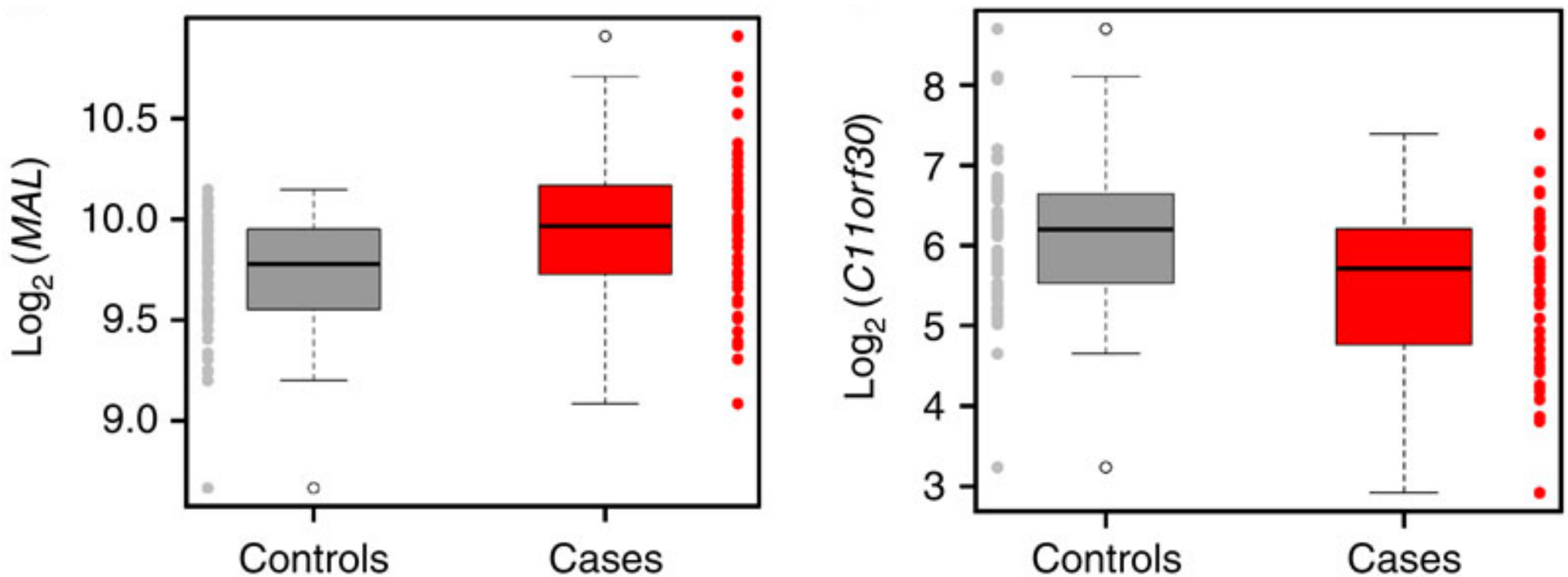


Utilidad de los boxplots

- Valores atípicos probables - 1.5 IQR
- Valores atípicos - 3 IQR

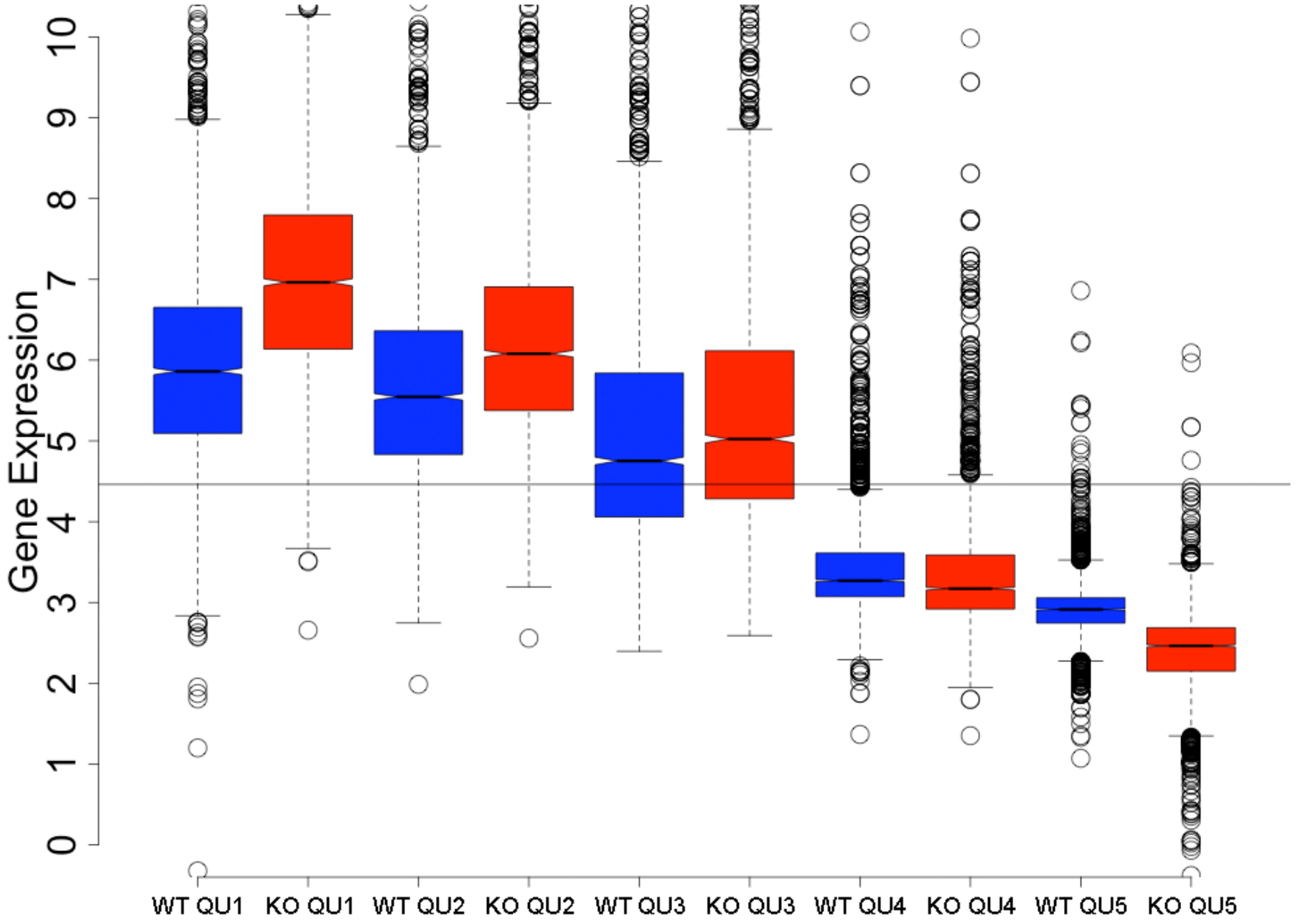


Boxplots sirven para comparar distribuciones y detectar valores atípicos



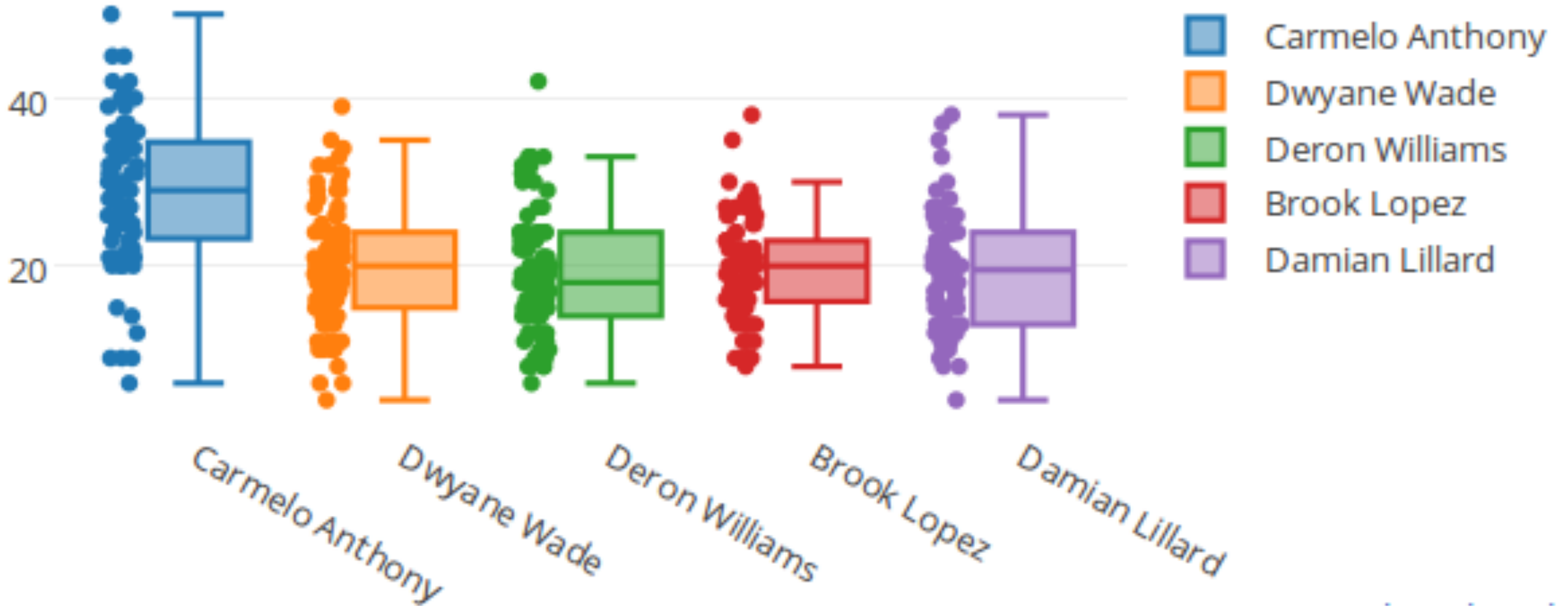
http://www.nature.com/articles/ncomms6748?ncid=txtlnkusaolp00000618

Boxplots sirven para comparar distribuciones y detectar valores atípicos



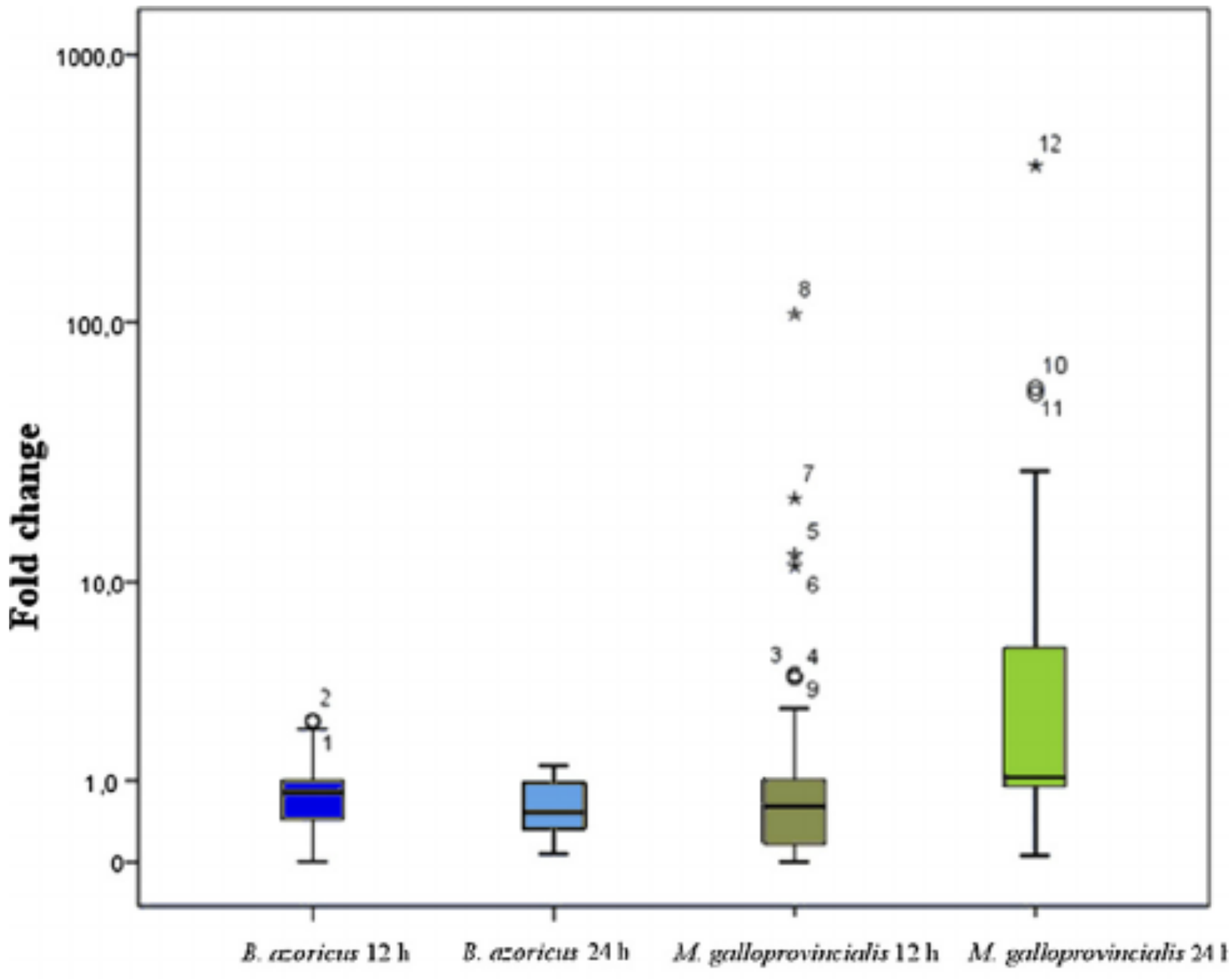
http://link.springer.com/article/10.1186/1471-2105-11-396

Buenas prácticas - mostrar los puntos del boxplot



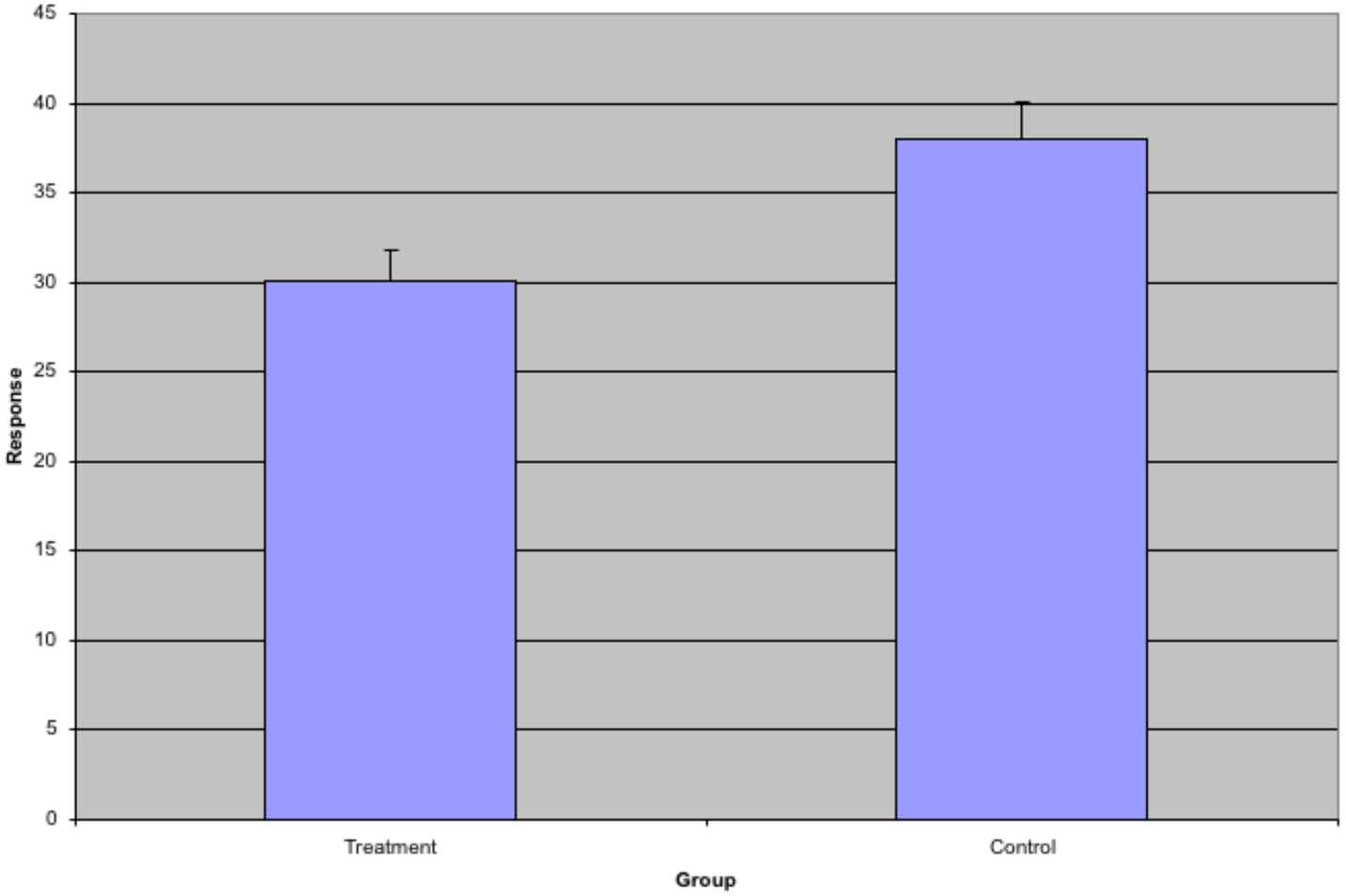
http://help.plot.ly/excel/box-plot

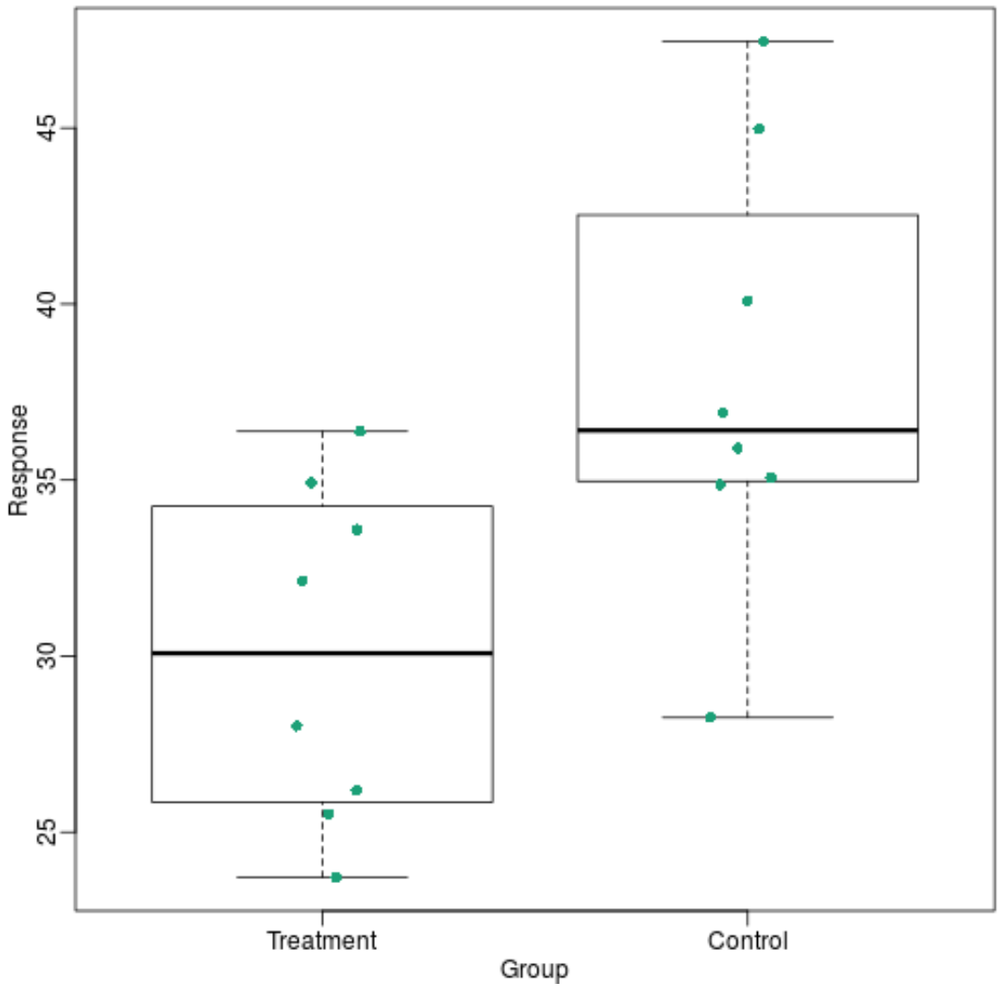
Buenas prácticas - identificar outliers



http://www.sciencedirect.com/science/article/pii/S1050464814002599

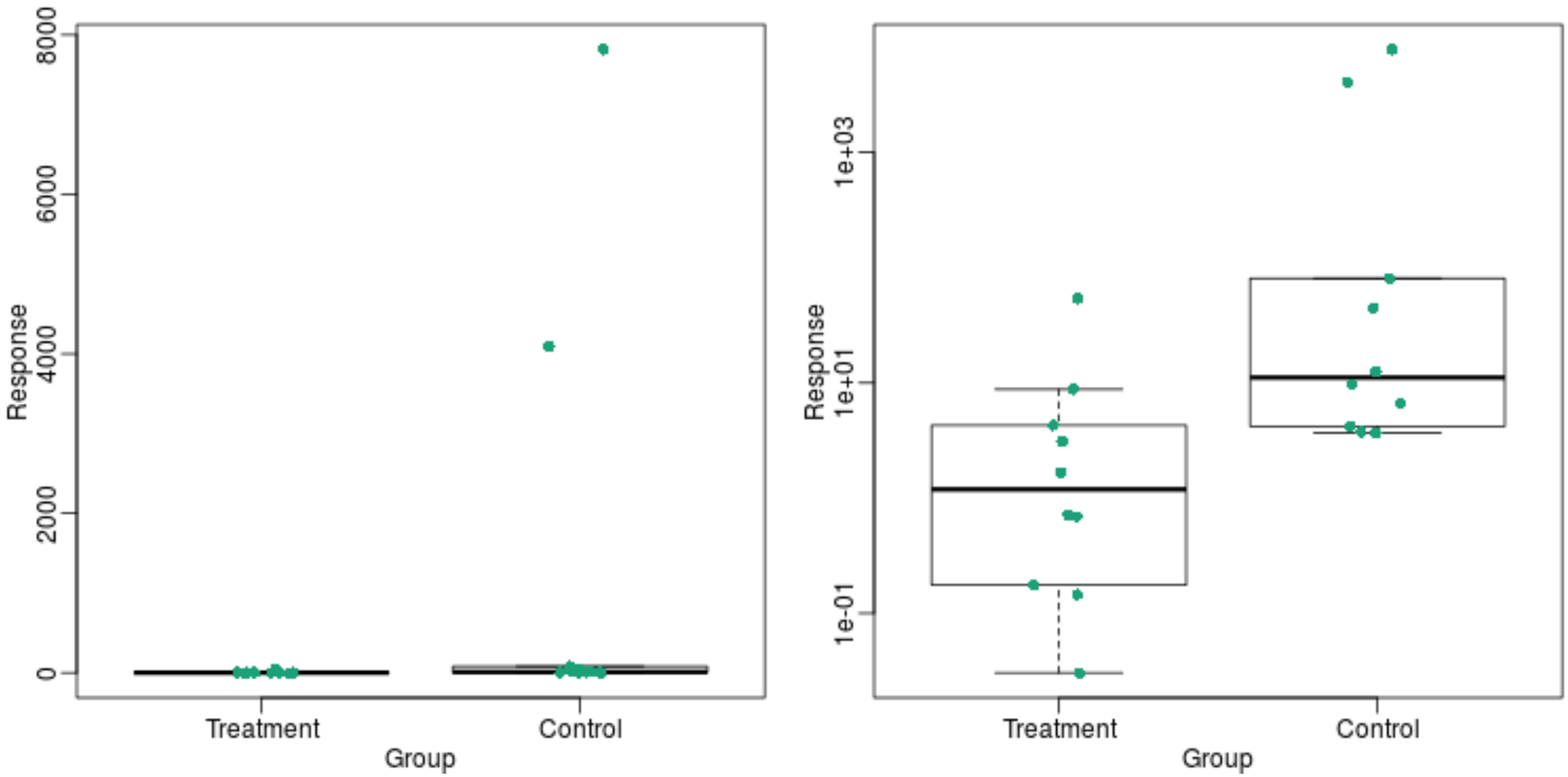
Malas prácticas - ocultar la distribución de datos con un gráfico de barras





http://genomicsclass.github.io/book/pages/plots_to_avoid.html

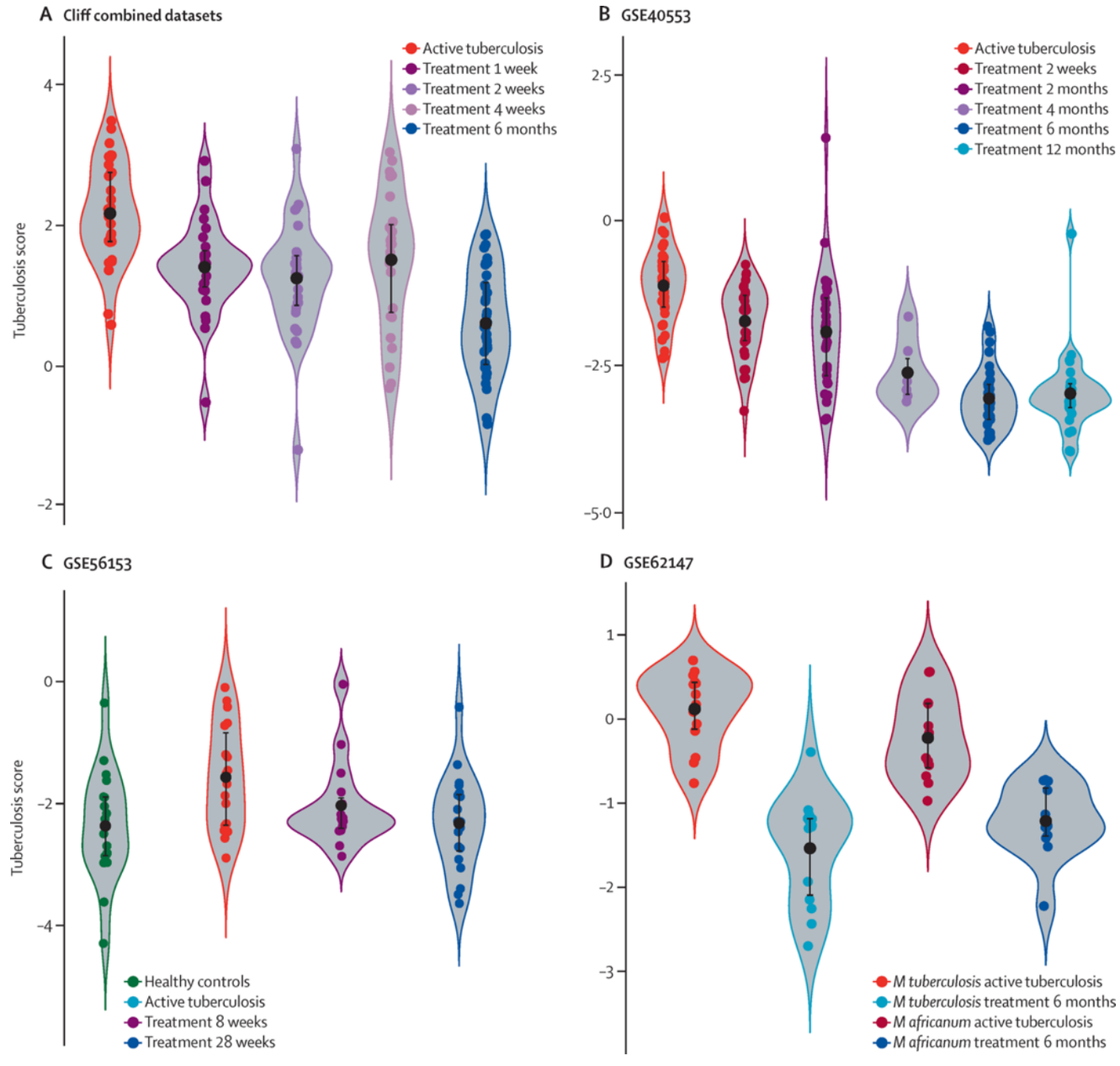
Malas prácticas - ocultar la distribución de datos al usar una escala inadecuada



http://genomicsclass.github.io/book/pages/plots_to_avoid.html

Violin plot

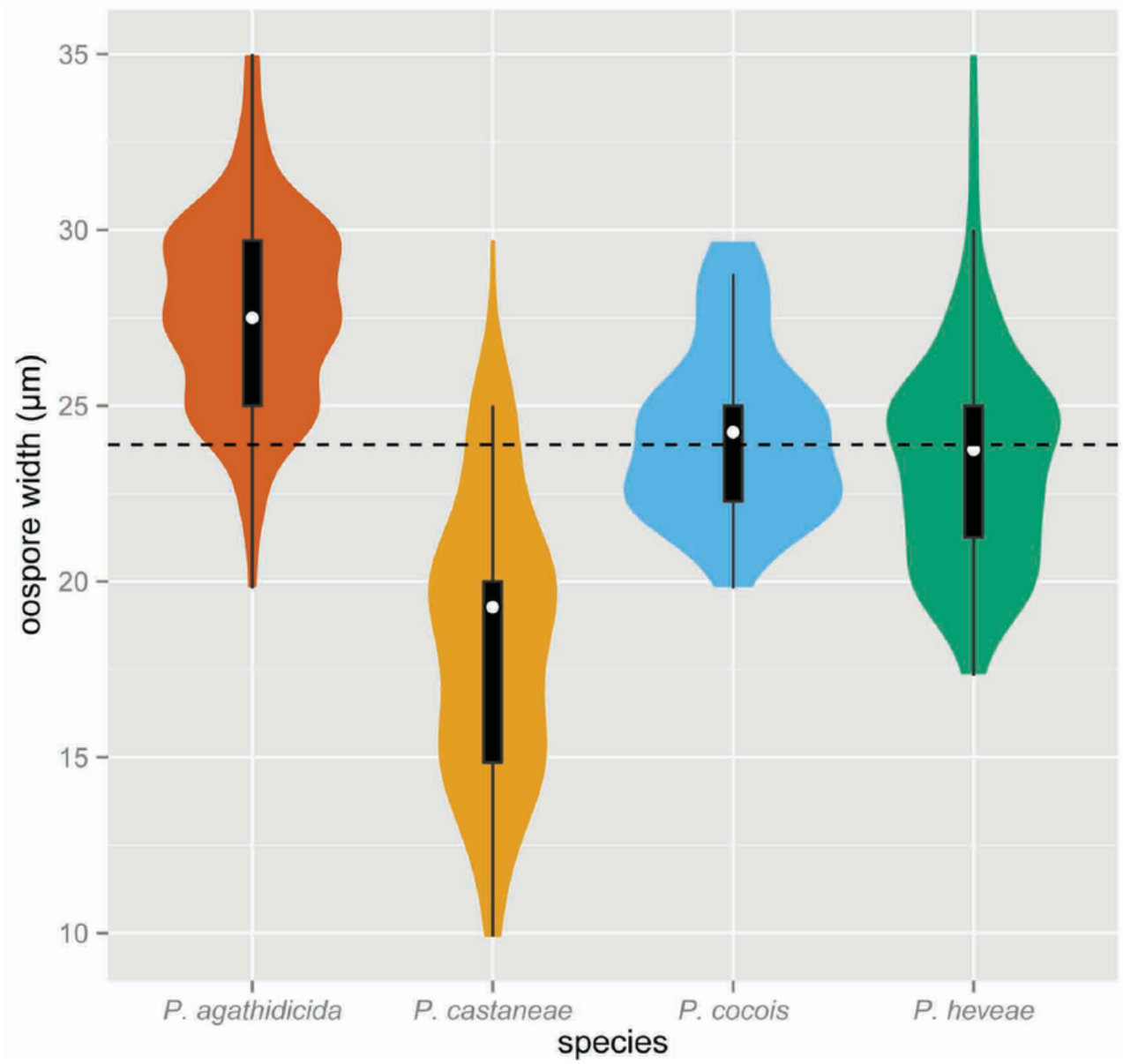
- Análogo a histogramas y gráficos de densidad
- Es un boxplot con un gráfico de densidad a cada lado



http://thelancet.com/cms/attachment/2049442930/2058662335/gr6_lrg.jpg

Violin plot

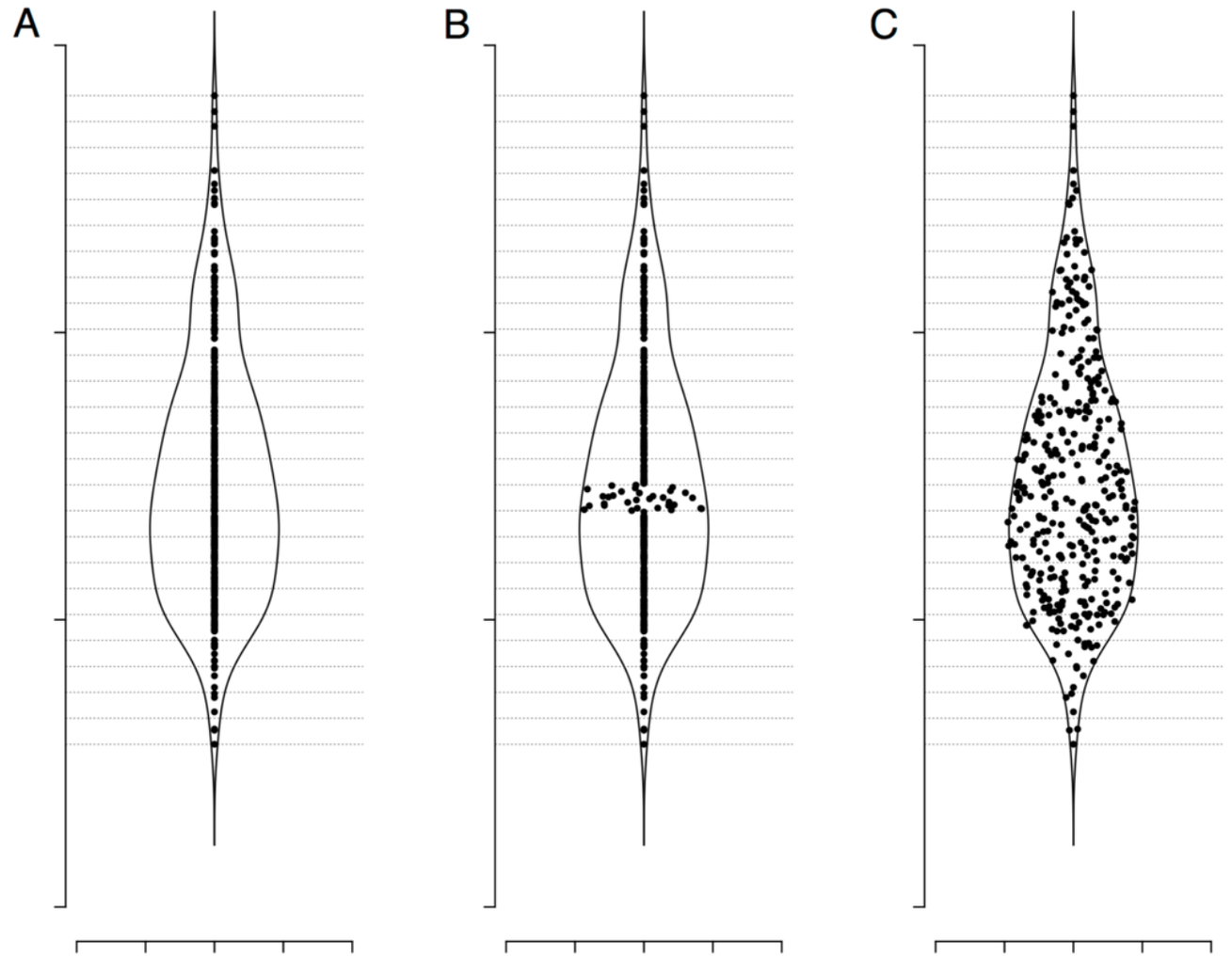
- También puede mostrar el IQR y la mediana



http://biotaxa.org/Phytotaxa/article/view/phytotaxa.205.1.2

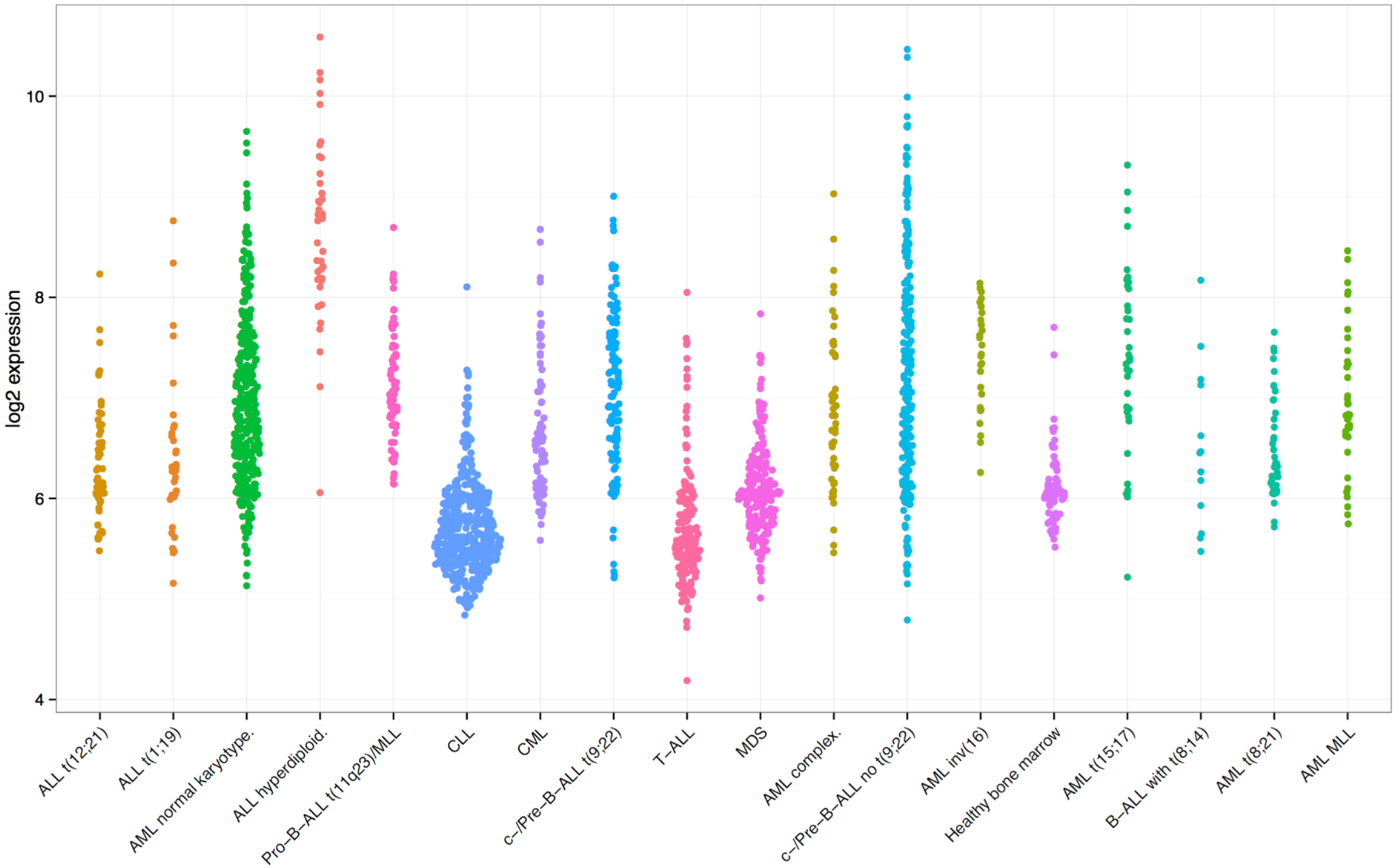
Sina plot

- Muestra el número de puntos, la distribución, valores atípicos, y la dispersión de los datos



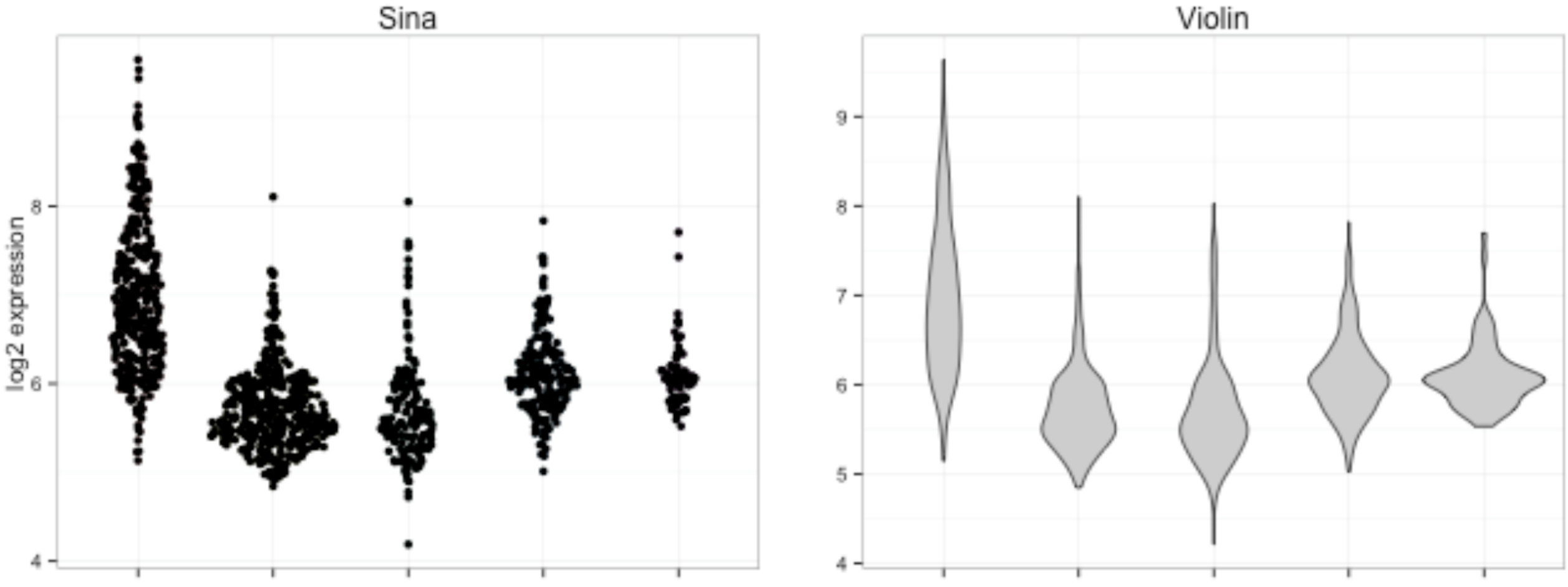
http://biorxiv.org/content/biorxiv/early/2015/10/02/028191.full.pdf

Sina plot



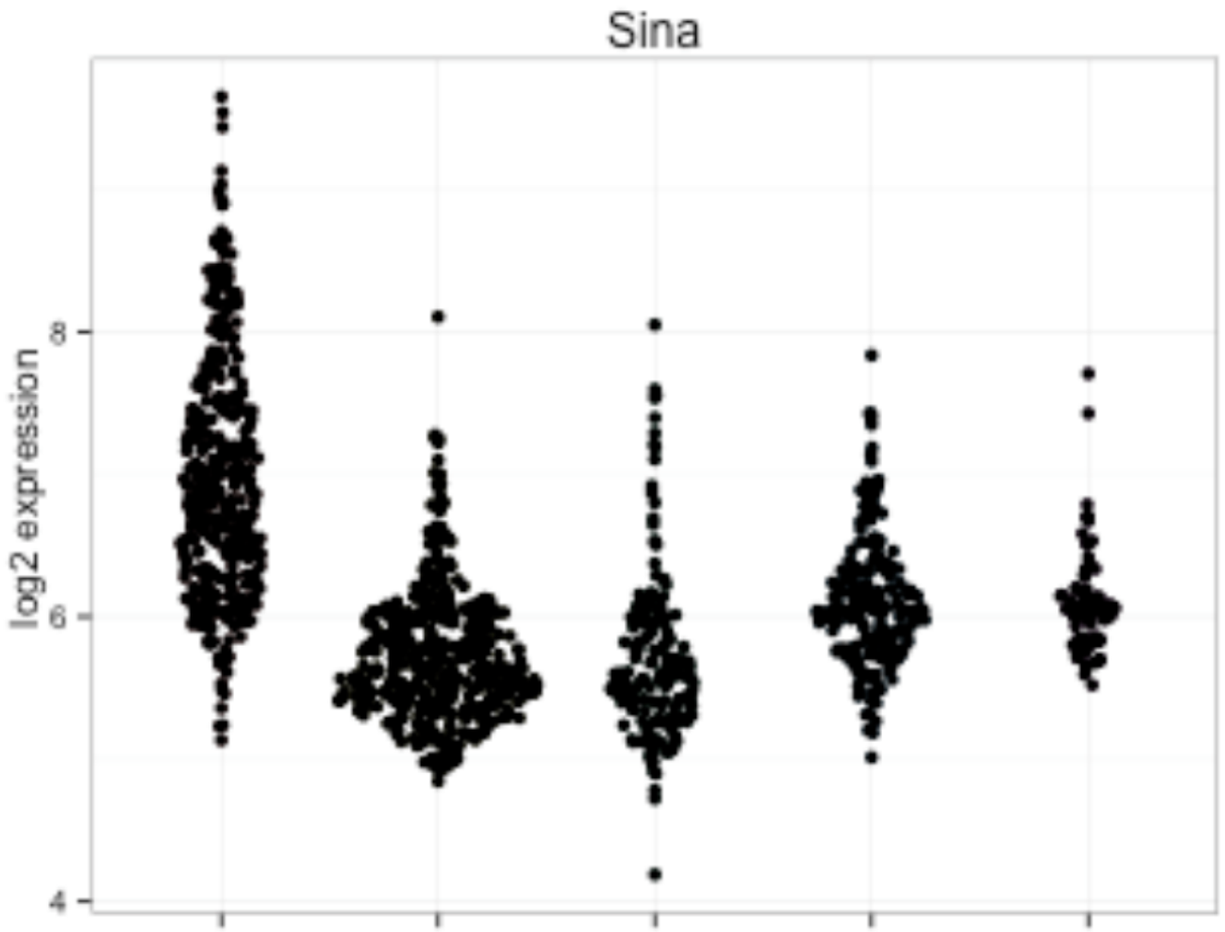
http://biorxiv.org/content/biorxiv/early/2015/10/02/028191.full.pdf

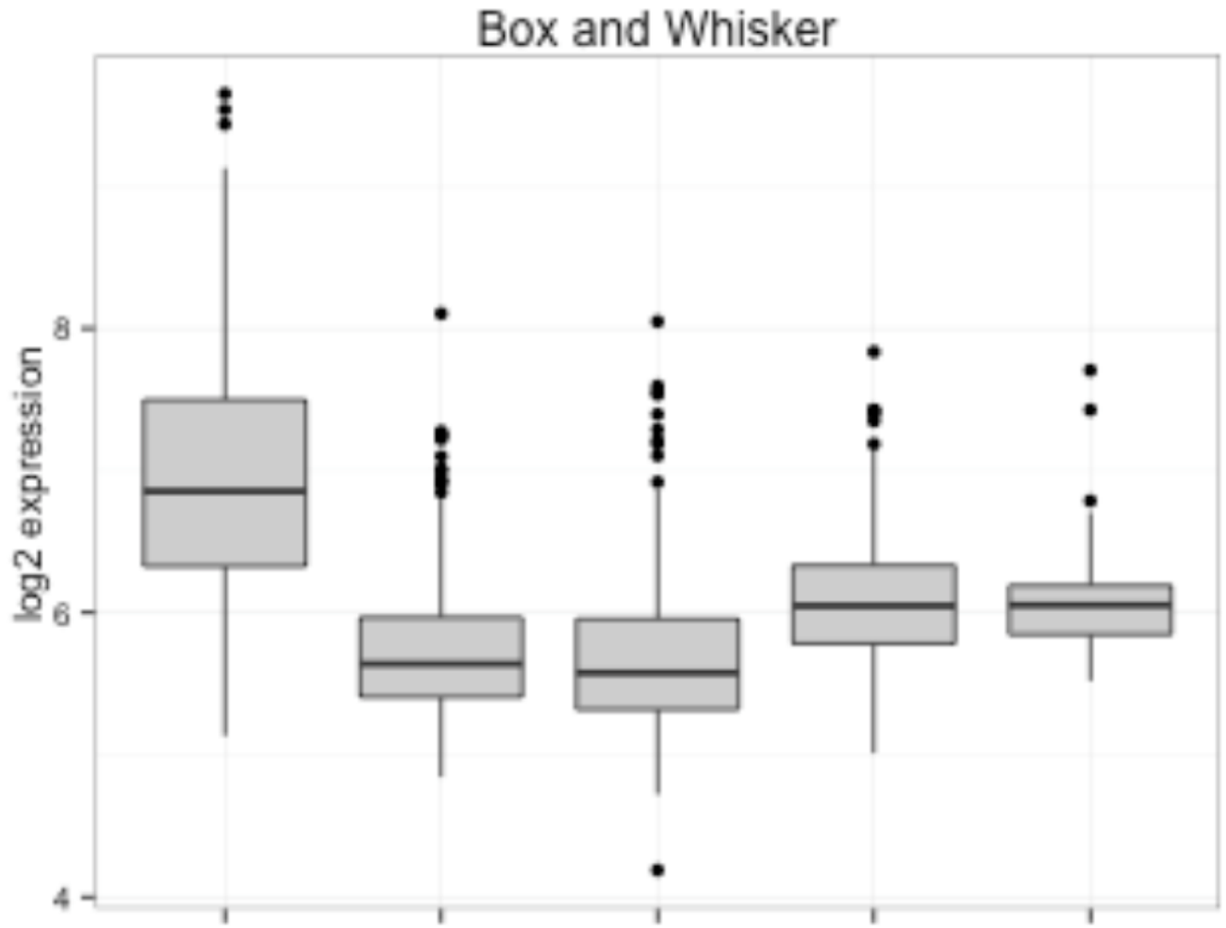
Sina plot



http://biorxiv.org/content/biorxiv/early/2015/10/02/028191.full.pdf

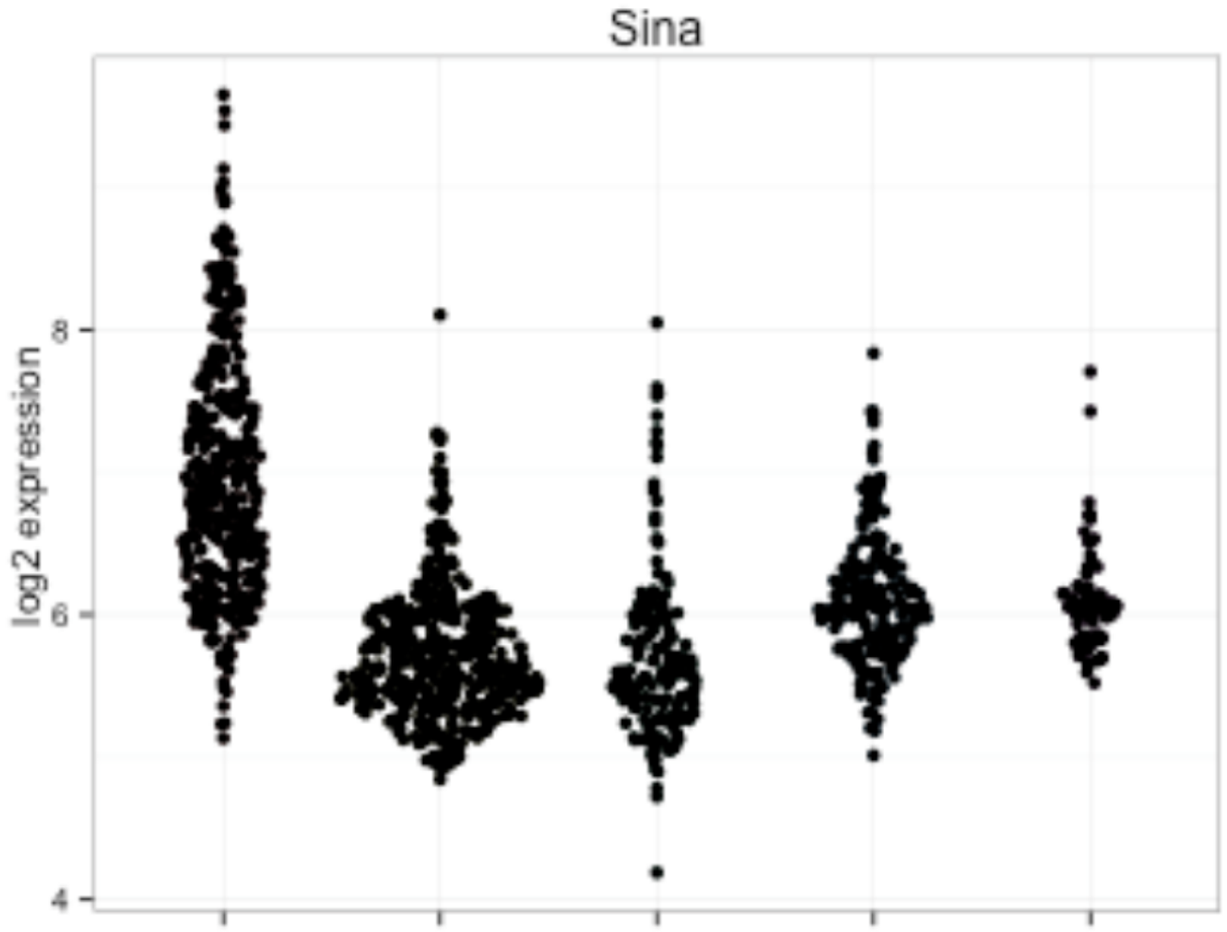
Sina plot

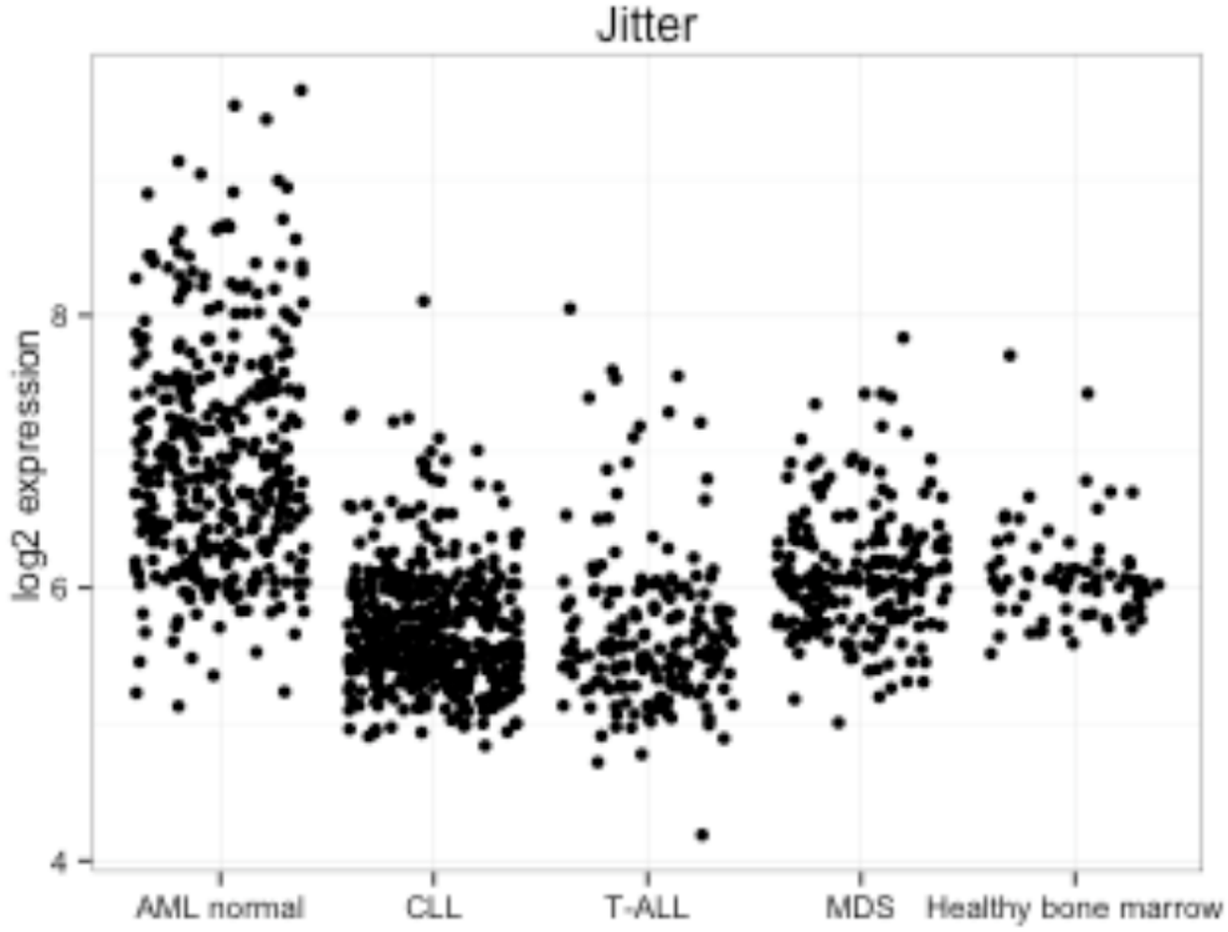




http://biorxiv.org/content/biorxiv/early/2015/10/02/028191.full.pdf

Sina plot

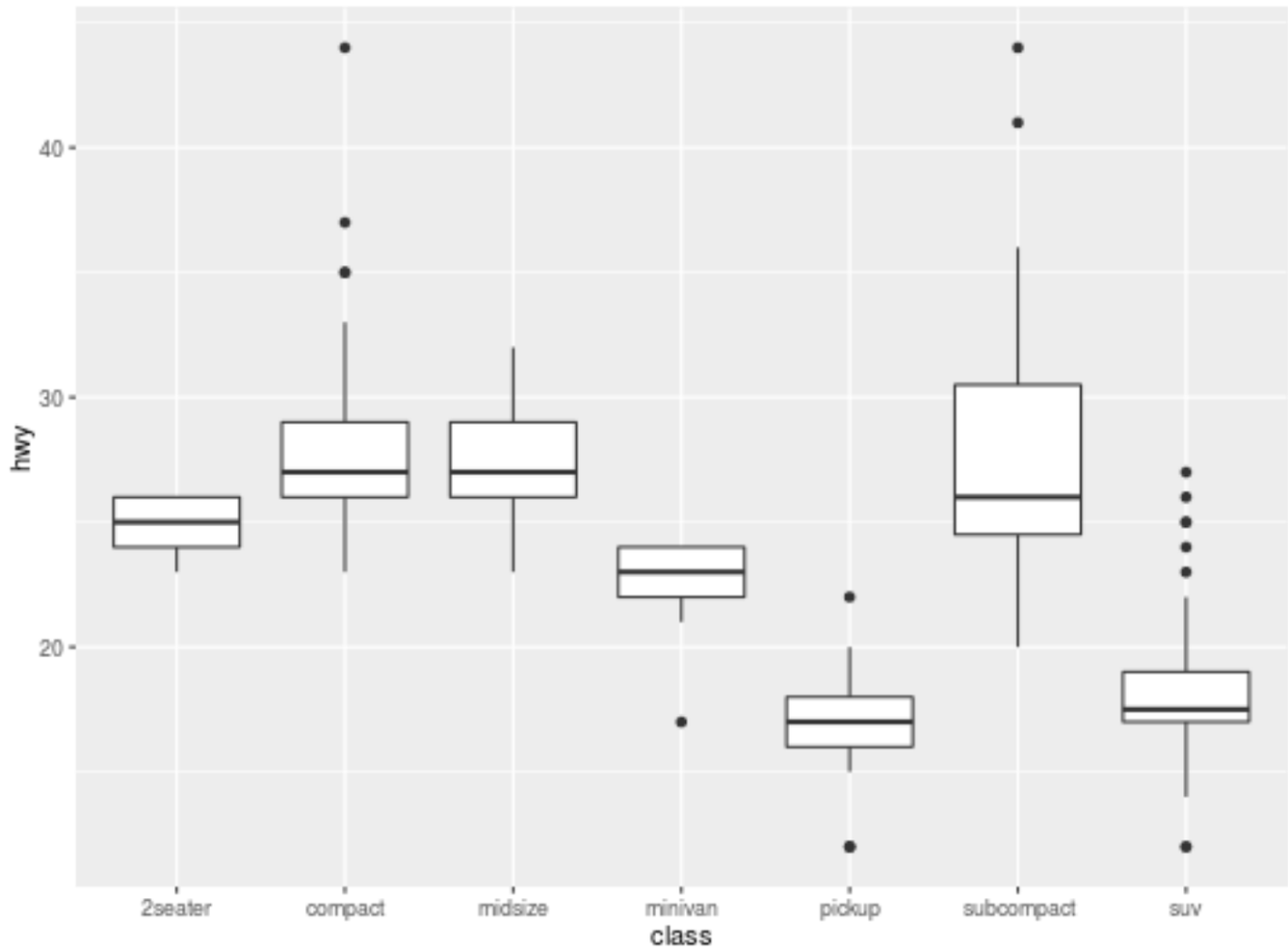




http://biorxiv.org/content/biorxiv/early/2015/10/02/028191.full.pdf

Un poco de ggplot2

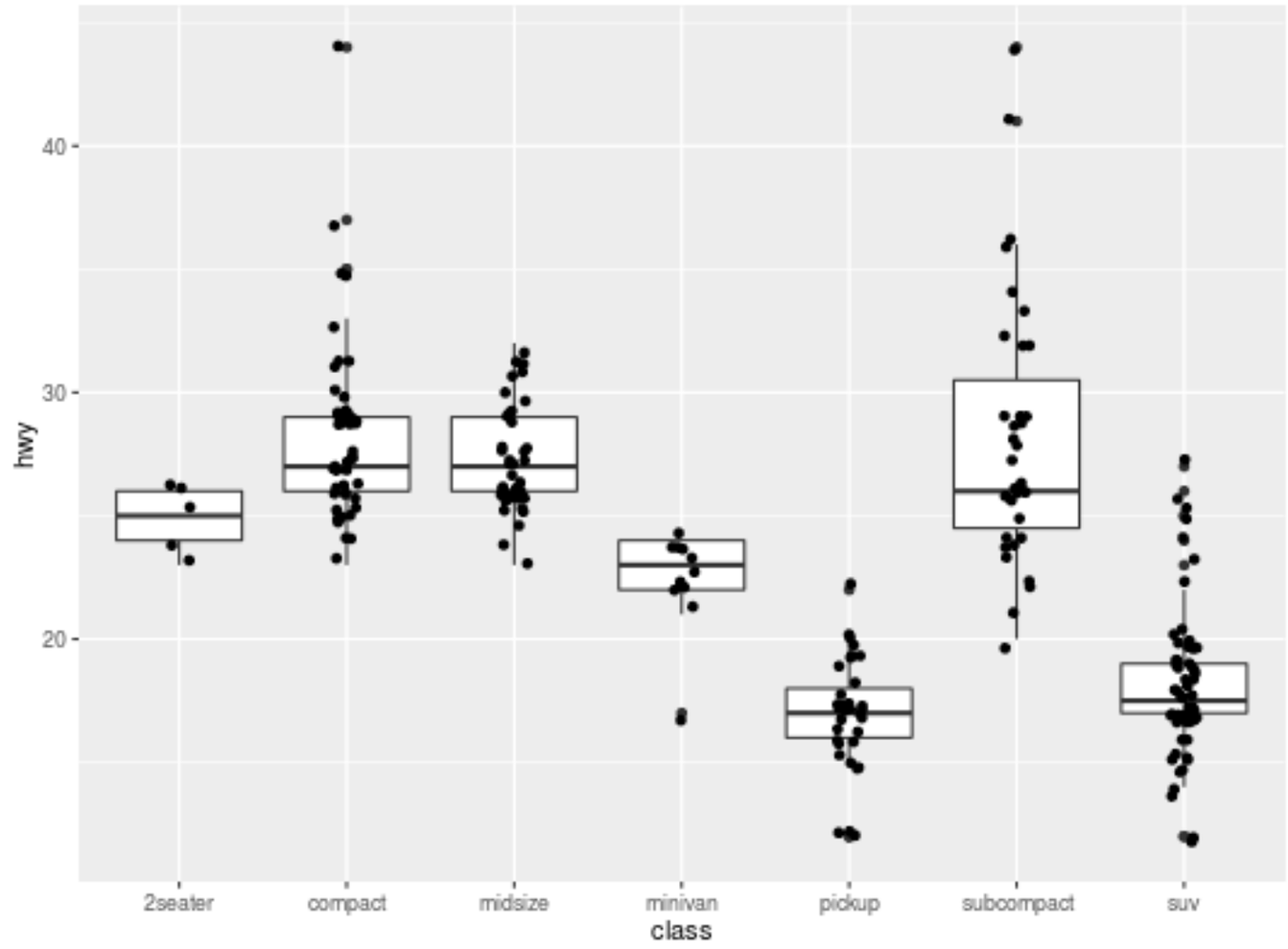
- probemos geom_boxplot()
- p <- ggplot(mpg, aes(class, hwy))
- p + geom_boxplot()



Un poco de ggplot2

- Podemos decirle a ggplot que nos muestre los puntos

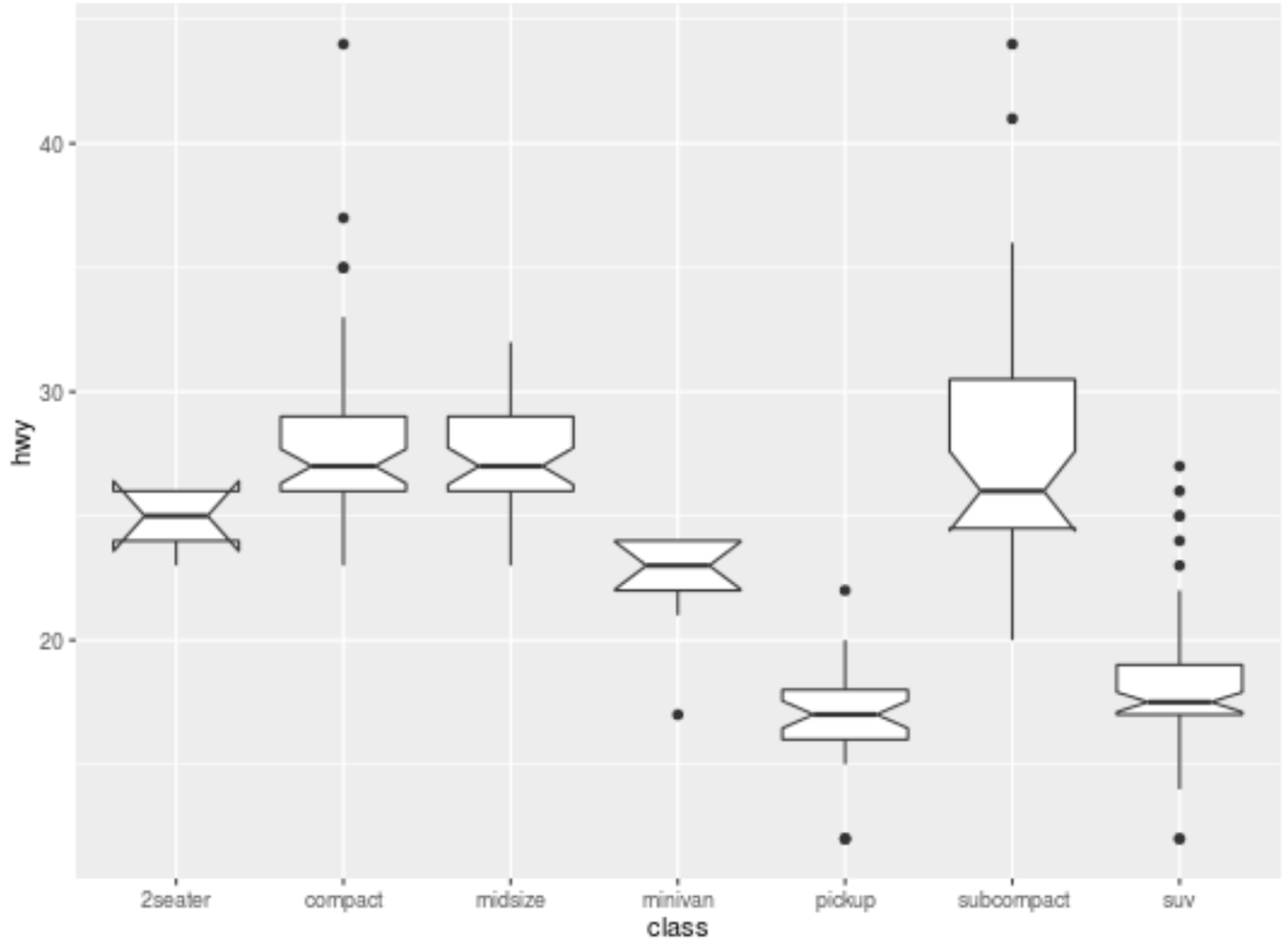
- p + geom_boxplot() + geom_jitter(width = 0.2)



Un poco de ggplot2

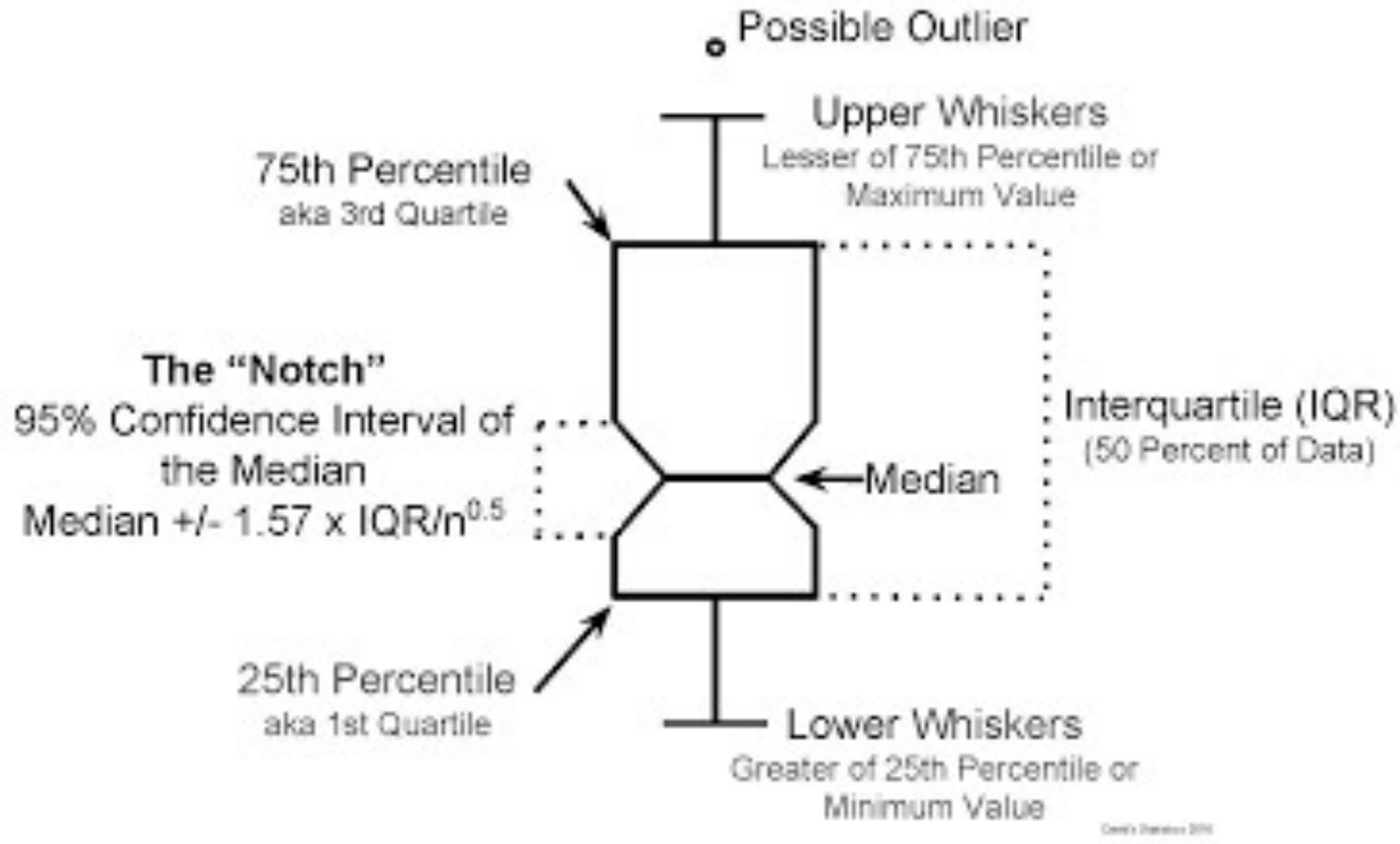
- Podemos decirle a ggplot que queremos ver intervalos de confianza en nuestro boxplot

- p + geom_boxplot(notch = TRUE)



Agregar un “notch” o muesca

- Si las muescas no se superponen, hay evidencia fuerte de que las medianas son significativamente diferentes



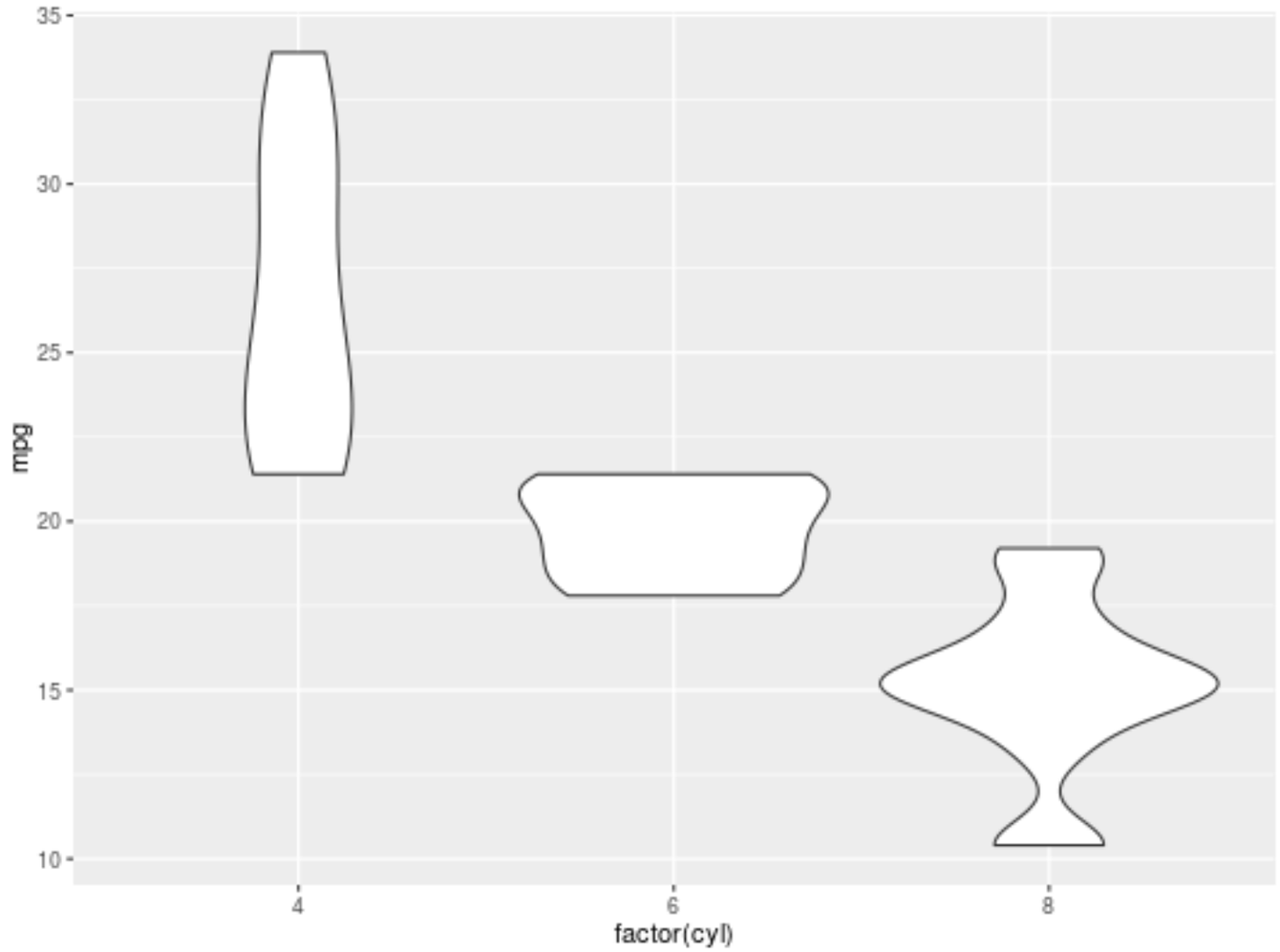
https://sites.google.com/site/davidsstatistics/home/notched-box-plots

Un poco de ggplot2

- De manera similar podemos construir un violin plot

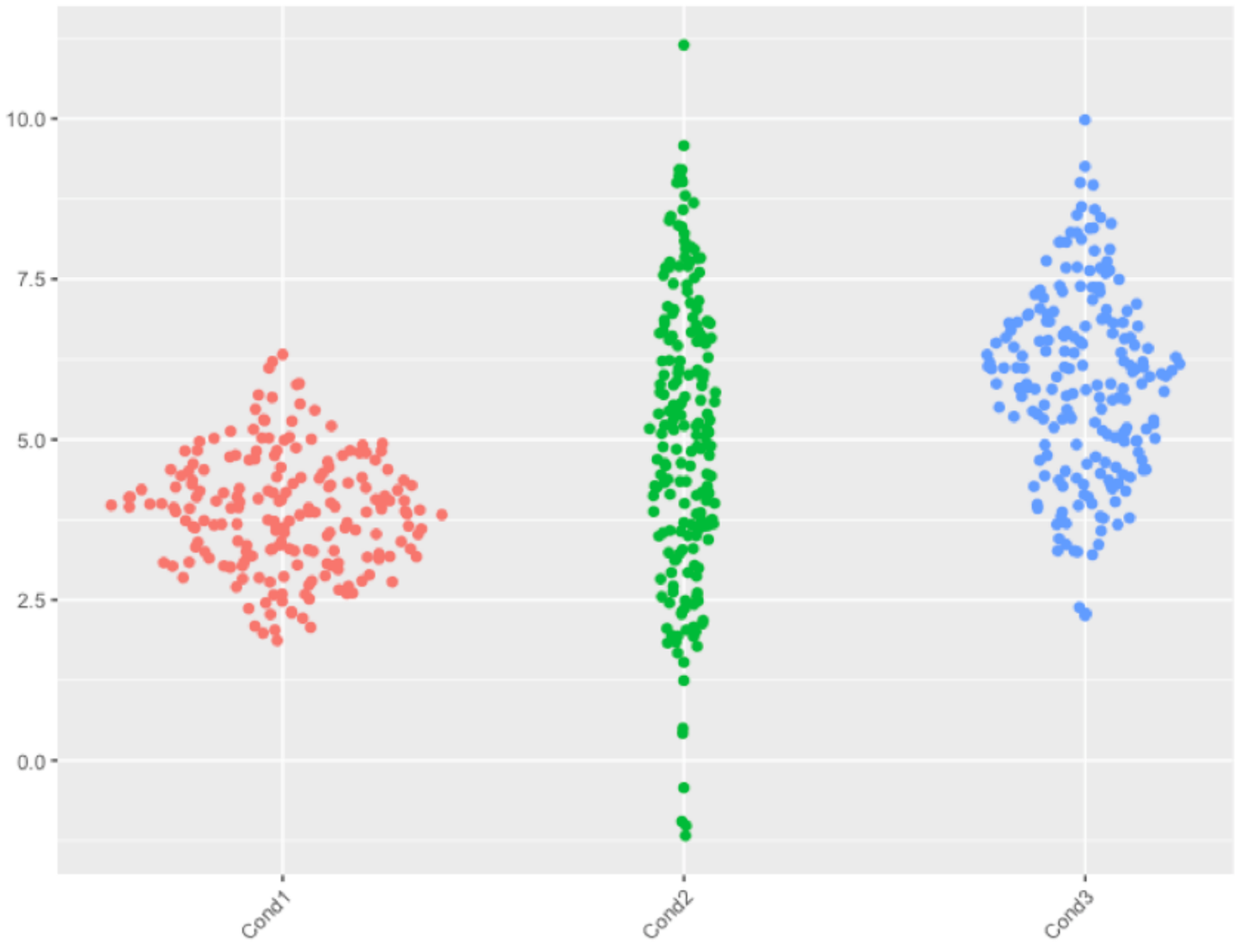
- p <- ggplot(mtcars, aes(factor(cyl), mpg))

- p + geom_violin()



Dejemos instalado el paquete sinaplot

- install.packages(“sinaplot”)
- library(“sinaplot”)
- x <- c(rnorm(200, 4, 1), rnorm(200, 5, 2), rnorm(200, 6, 1.5))
- groups <- c(rep("Cond1", 200), rep("Cond2", 200), rep("Cond3", 200))
- sinaplot(x, groups)



“They are different, but not different enough to matter […]”

– Roald Hoffmann, Premio Nobel de Química 1981