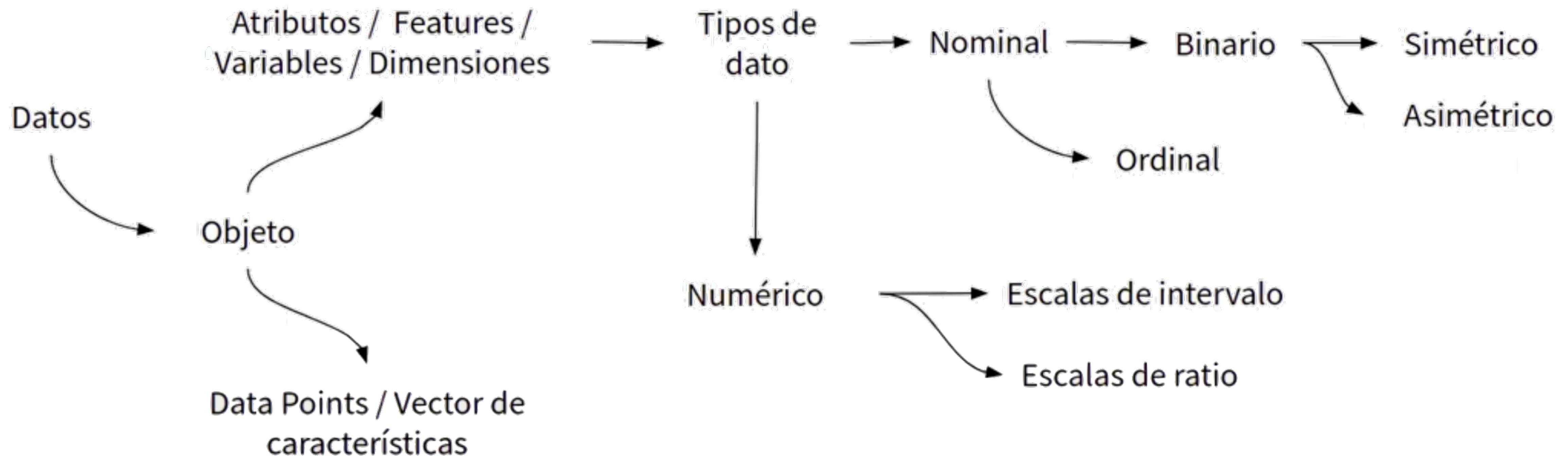


Introducción a la Minería de Datos

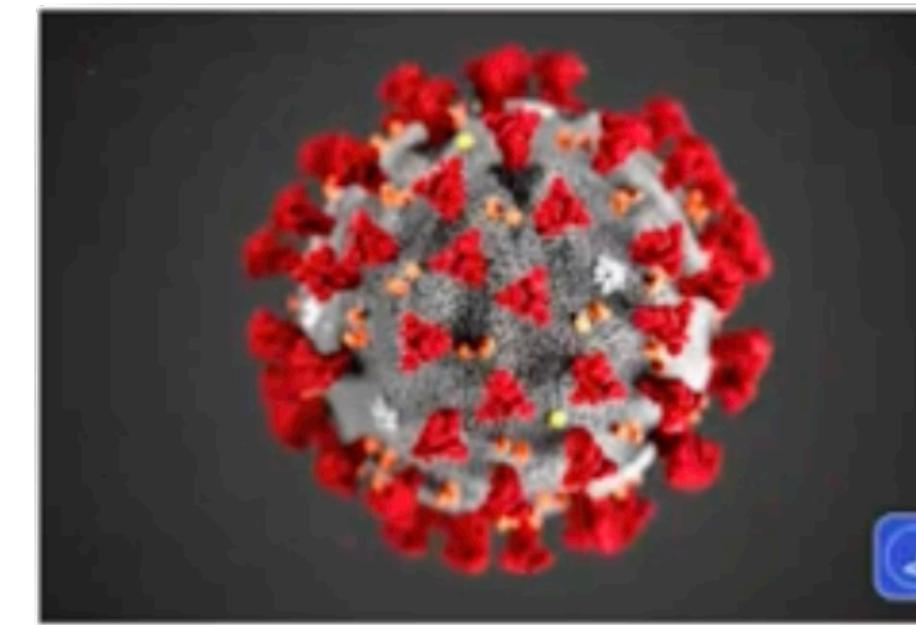
Verónica Guarín Escudero
Escuela de Estadística

Correo: jvguarine@unal.edu.co

Tipos de Datos



Tipos de Datos



Tipo	Nombre	Síntomas Habituales	Síntomas No Habituales
------	--------	---------------------	------------------------

Tipos de Atributos

- **Nominal:** categorías, estados o "nombres de cosas"
 - Color de pelo = {castaño, negro, rubio, marrón, gris, rojo, blanco}
 - estado civil, ocupación, números de identificación, códigos postales
- **Binario**
 - Atributo nominal con solo 2 estados (0 y 1)
 - **Binario simétrico:** ambos resultados son igualmente importantes
 - por ejemplo, género
 - **Binario asimétrico:** los resultados no son igualmente importantes.
 - por ejemplo, examen médico (positivo vs. negativo)
 - Convención: asigne 1 al resultado más importante (p. Ej., VIH positivo)
- **Ordinal**
 - Los valores tienen un orden que tiene un sentido (*ranking*), pero se desconoce la magnitud entre los valores sucesivos.
Tamaño = {pequeño, mediano, grande}, calificaciones, rango militar

Tipos de Atributos

- **Cuantitativos:** valores enteros o reales
- **Escalas de Intervalos**
 - Mide una escala de iguales unidades
 - Los valores tienen un orden:
 - temperatura en C° o F°
 - El cero es relativo (y arbitrario)
- **Escalas de razón (*ratio*)**
 - Hay un punto cero
 - Podemos hablar de un valor como un múltiplo (o razón) de otro valor.
 - 10 cm es dos veces más alto que 5 cm.
 - Ejemplos:
 - temperatura en Kelvin, longitud/latitud, cantidades, montos de dinero

Descripciones estadísticas básicas

- Motivación:**
 - Tener una vista general de sus datos.
- Medidas de tendencia central:**
 - Dado un atributo ¿Donde caen la mayoría de sus valores?
 - Media, mediana, moda, rango medio
- Medidas de dispersión:**
 - ¿Cómo se distribuyen los datos?
 - Rango, cuartiles y rango intercuartil.
 - Desviación estándar y varianza
 - 5 números mágicos, boxplots...
- Análisis gráfico**
 - Histogramas, gráficos de dispersión, QQ, etc

Medidas de Tendencia Central

Media aritmética:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\mu = \frac{\sum x}{N}$$

Media aritmética pesada:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Los pesos reflejan la importancia o frecuencia de ocurrencia de sus respectivos valores.

Media recortada: Se quitan valores extremos (no más de un 2%)

Medidas de Tendencia Central

Mediana:

- ❑ Es el valor que divide en 50% y 50%. Se utiliza el valor central si la cantidad de observaciones es impar y sino el promedio de los dos centrales
- ❑ Puede ser estimada también a partir de datos agrupados.

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq})l}{\text{freq}_{\text{median}}} \right) \text{width}$$

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

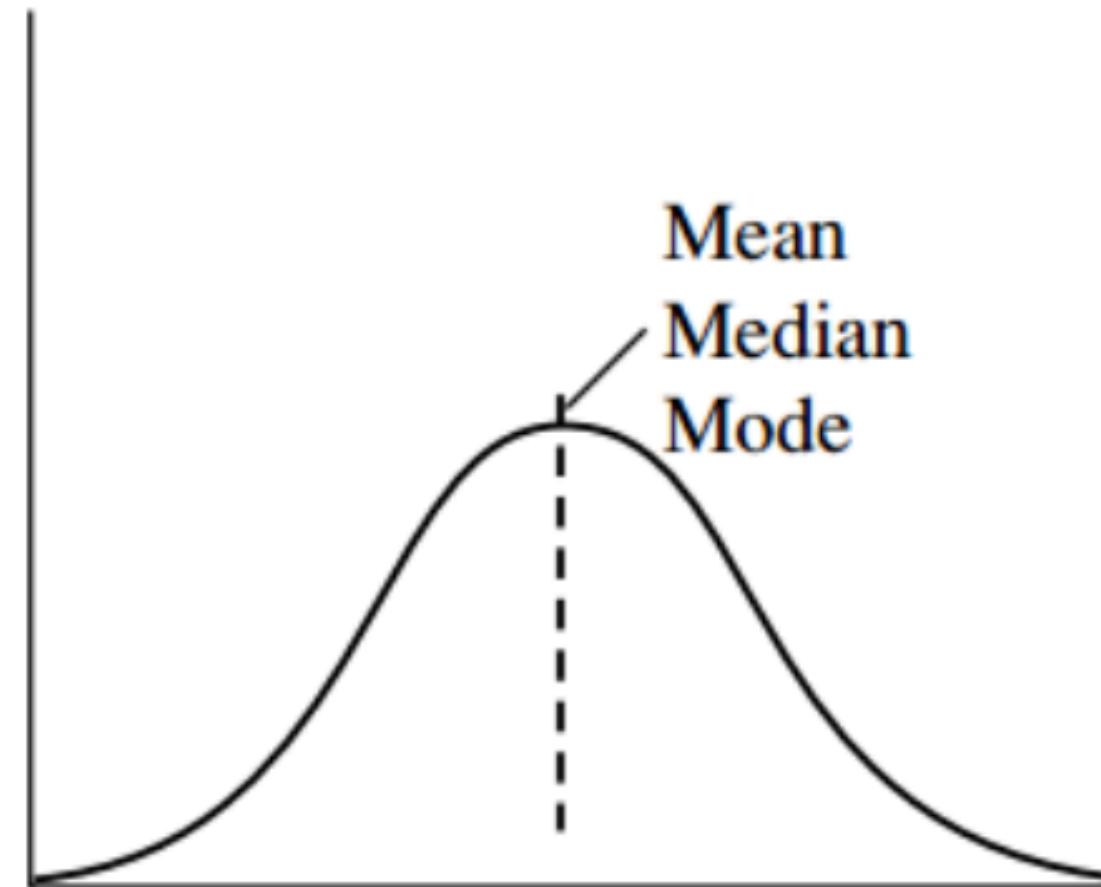
Moda:

- ❑ El valor que ocurre con mayor frecuencia en una variable
- ❑ Unimodal, **bimodal, trimodal** Multimodal

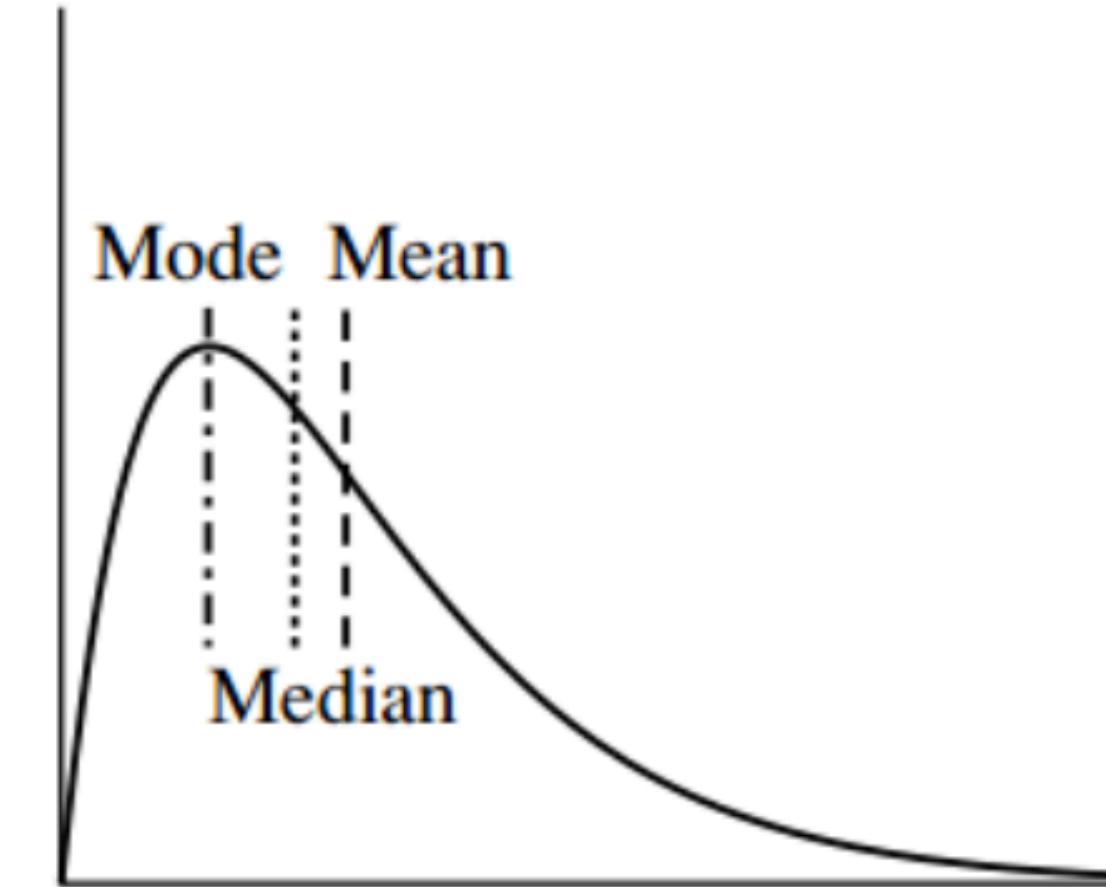
Simetría y datos sesgados

Los datos en la mayoría de las aplicaciones reales no son simétricos.

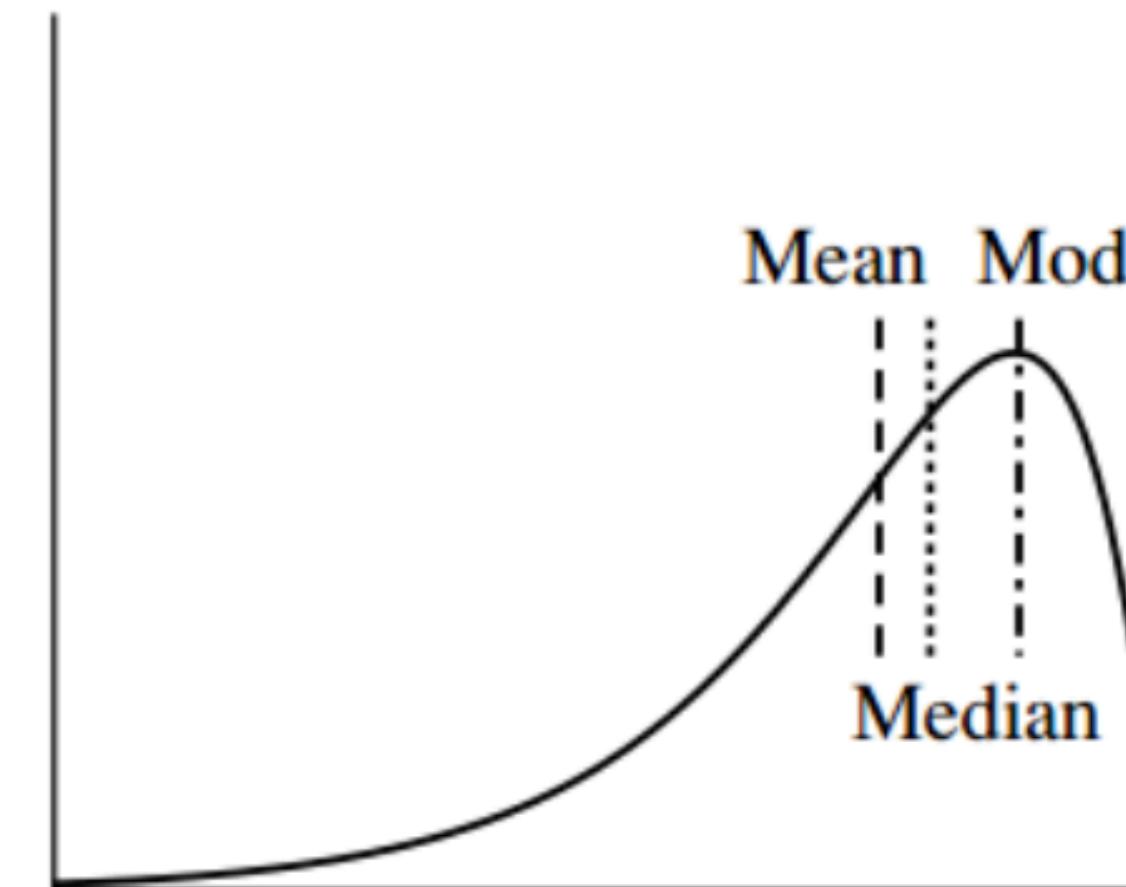
- Sesgo positivo, la moda es menor que la mediana
- Sesgo negativo, la moda es mayor que la mediana



(a) Symmetric data



(b) Positively skewed data



(c) Negatively skewed data

Medidas de Dispersion

❑ Cuartiles, outliers y boxplots

- ❑ Cuartiles: Q1 (percentil 25), Q3 (percentil 75)
- ❑ Rango intercuartil: $IQR = Q3 - Q1$
- ❑ Resumen de 5 números: min, Q1, median, Q3, max
- ❑ Boxplot:
 - ❑ los extremos de la caja son los cuartiles;
 - ❑ la marca de la caja es la mediana; los bigotes están a $1.5 * IQR$
- ❑ Outlier: generalmente son valores por encima o por debajo de $1.5 * IQR$

❑ Varianza y Desviación estándar (muestral: s, población: σ)

❑ Varianza:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$$

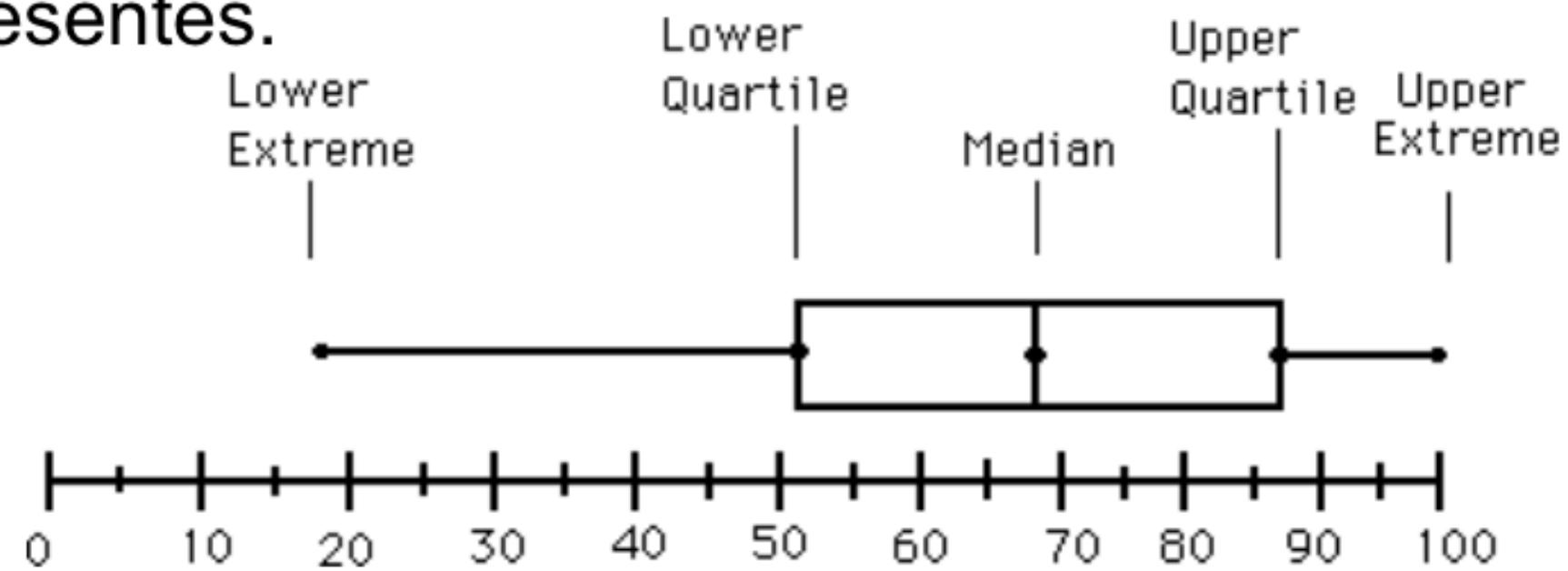
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

- ❑ Desviación estándar s (o σ) es la raíz cuadrada de s^2 (o σ^2)

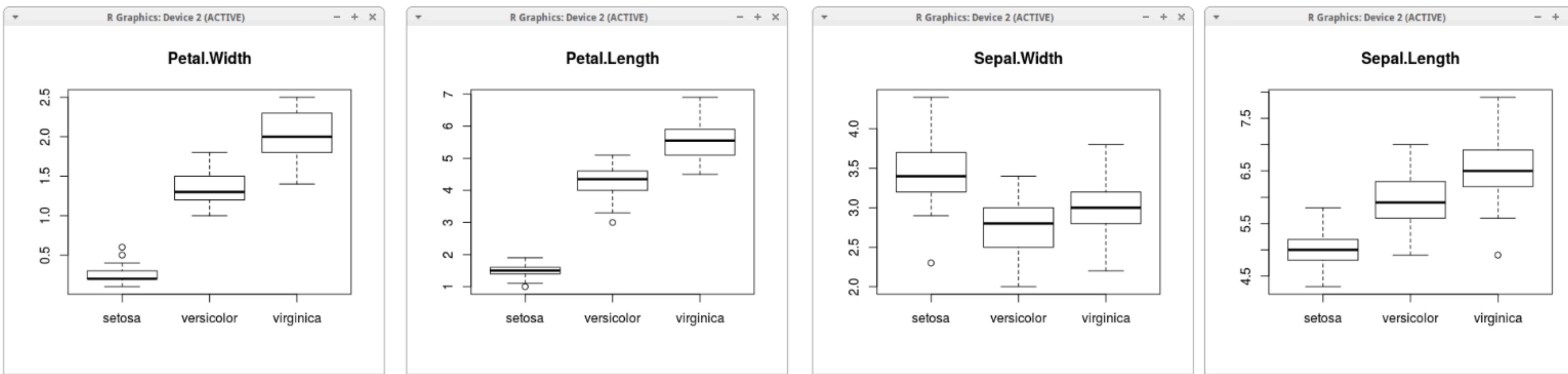
Boxplot

Los 5 números resumen de la distribución están presentes.

- Minimum, Q1, Median, Q3, Maximum

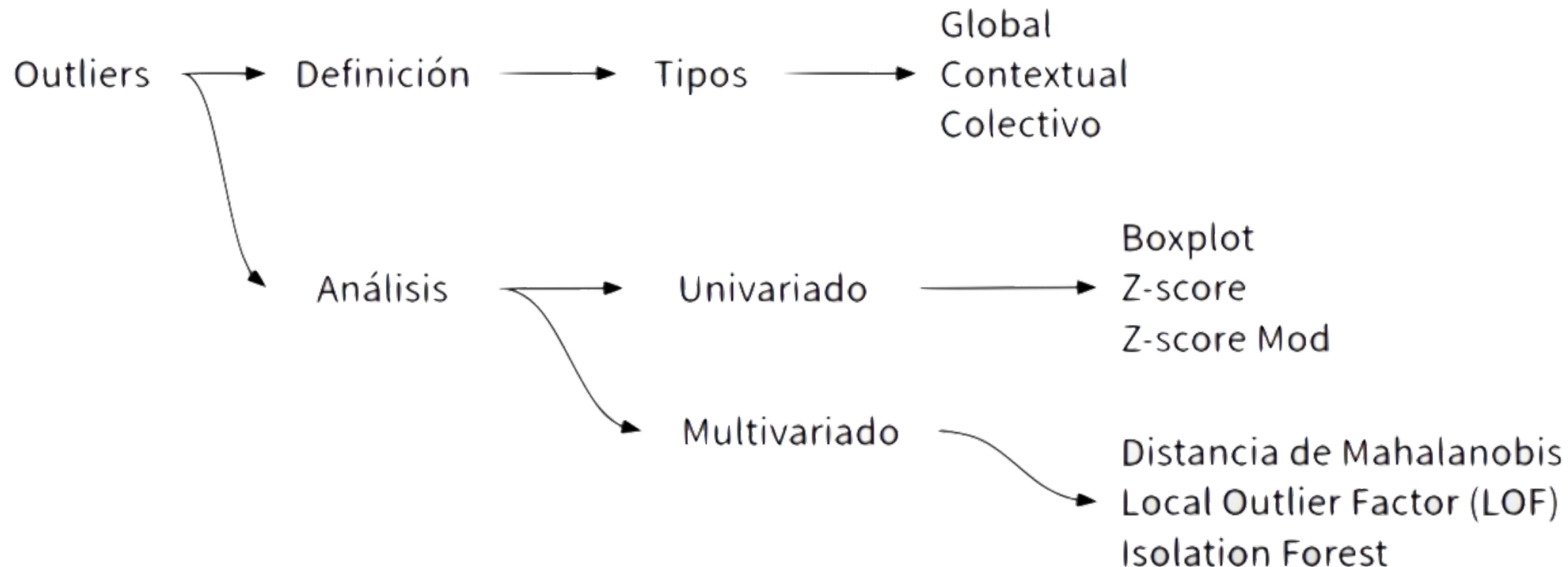


Análisis de separabilidad



Análisis de Outliers

Análisis de Outliers



Análisis de Outliers

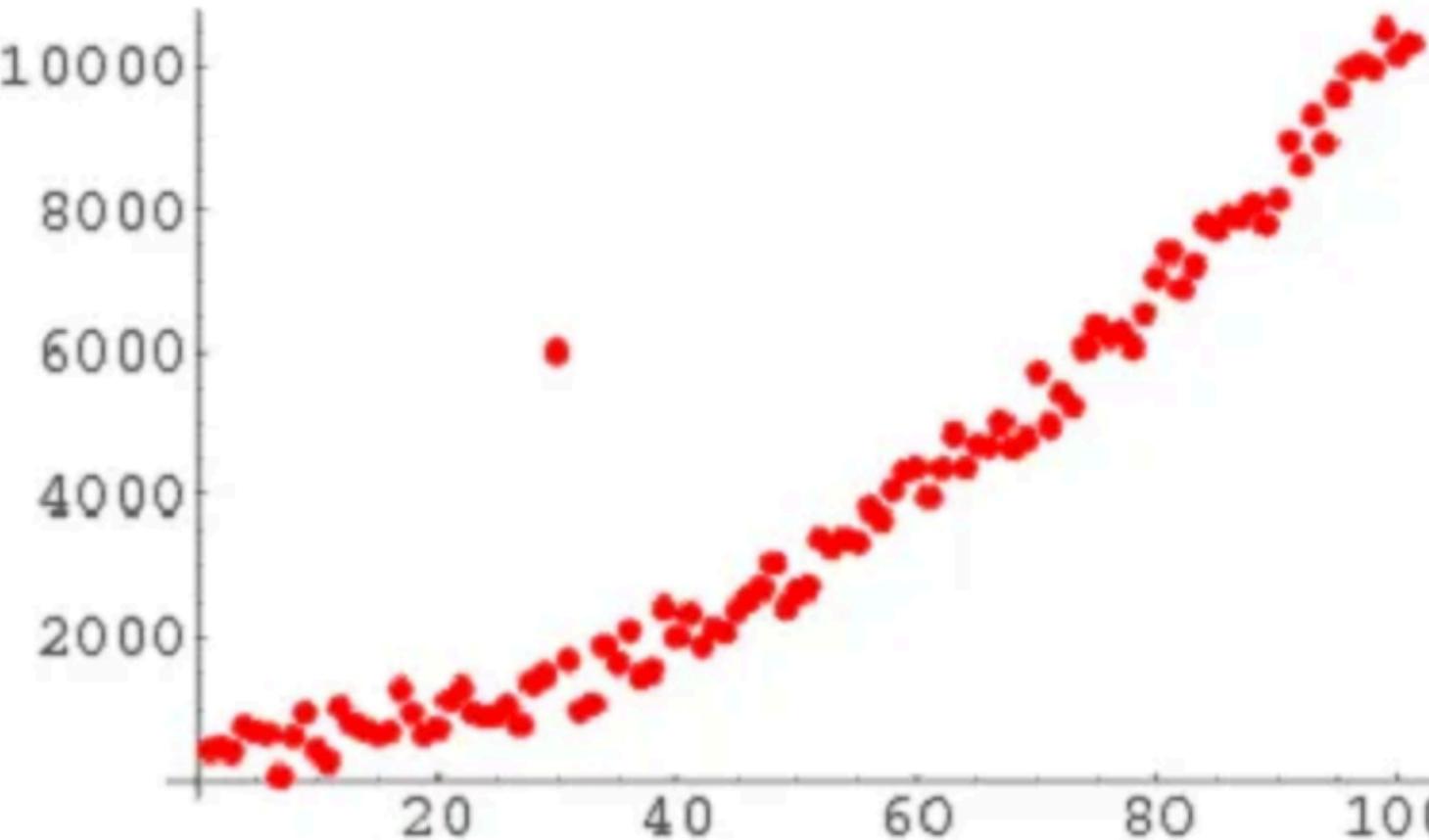
“Un *outlier* es una observación que se desvía tanto de las otras observaciones como para despertar sospechas que fue generado por un mecanismo diferente”

D. Hawkins. Identification of Outliers

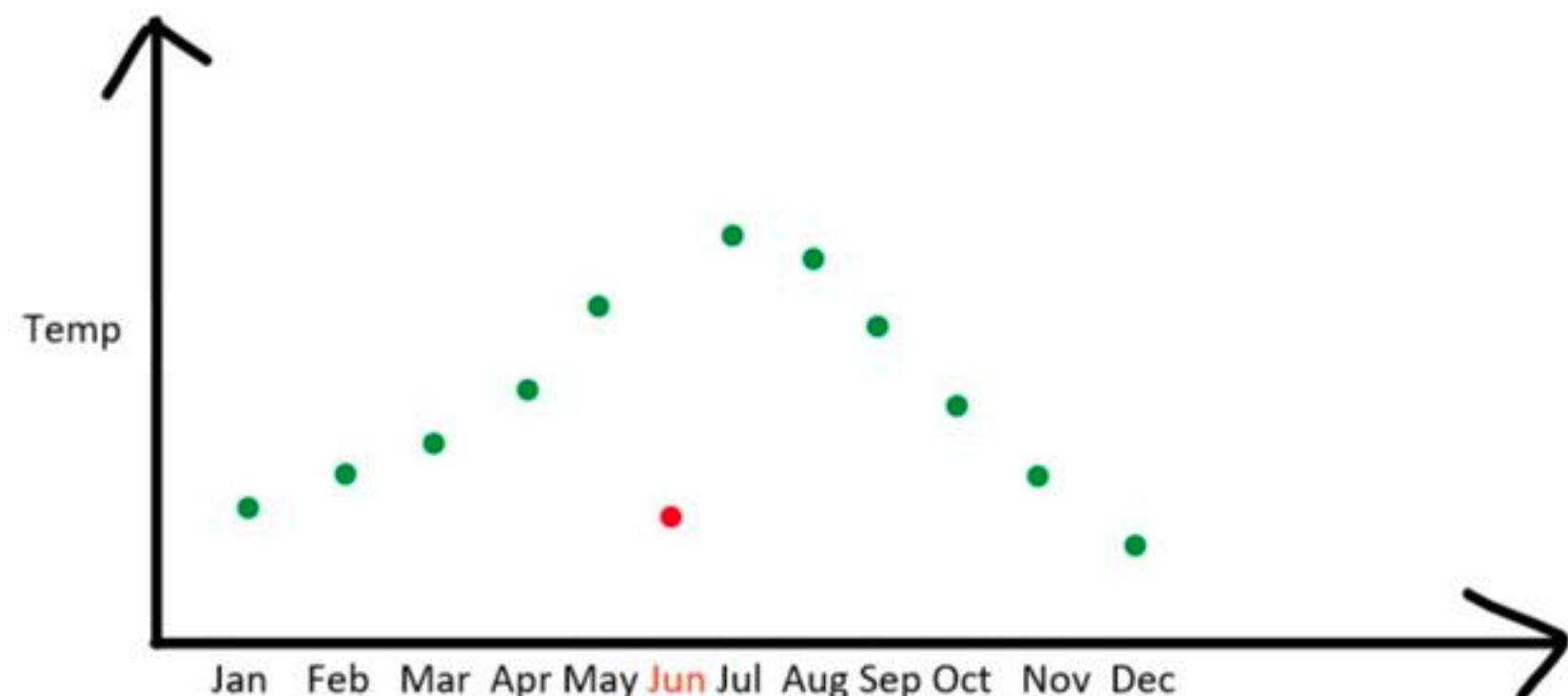
- Análisis de outliers.
 - Tipos de outliers
 - Revisión de métodos univariados: IQR, Boxplot, z-score....
 - Métodos multivariados: Local Outlier Factor (LOF)

Análisis de Outliers

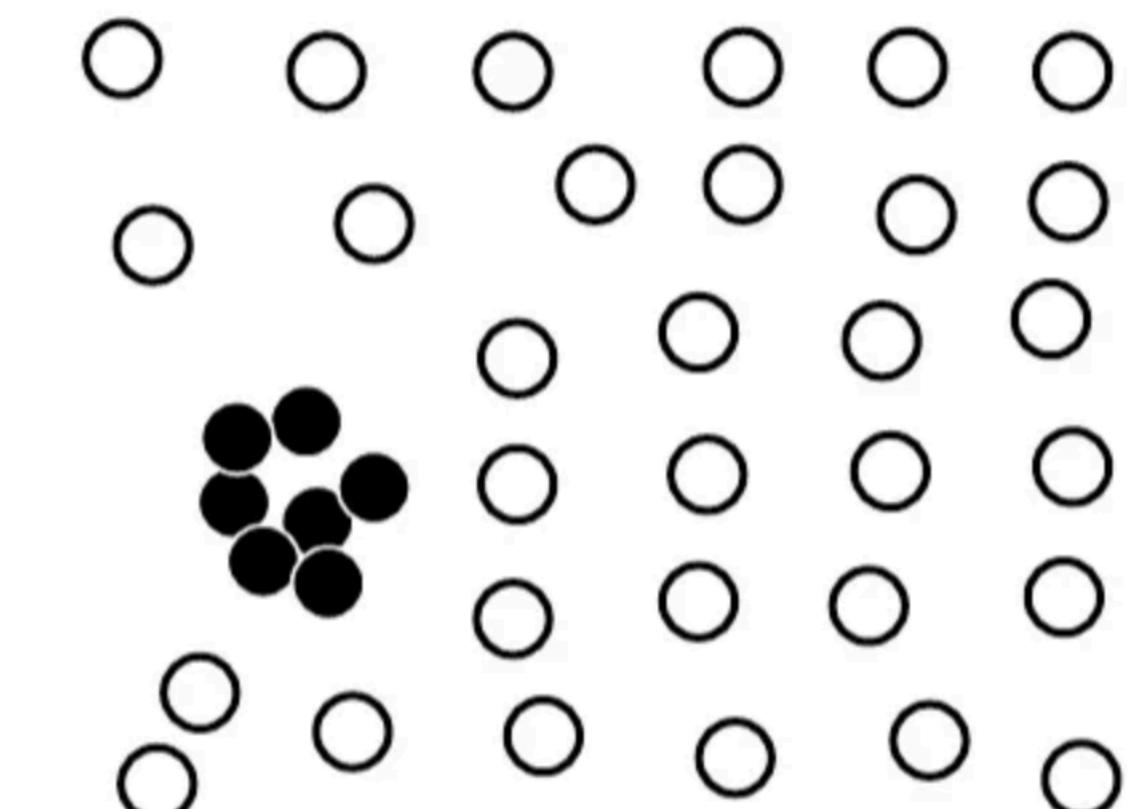
Global/point Outlier



Contextual/conditional Outlier



Collective Outlier



Análisis de Outliers

Univariado

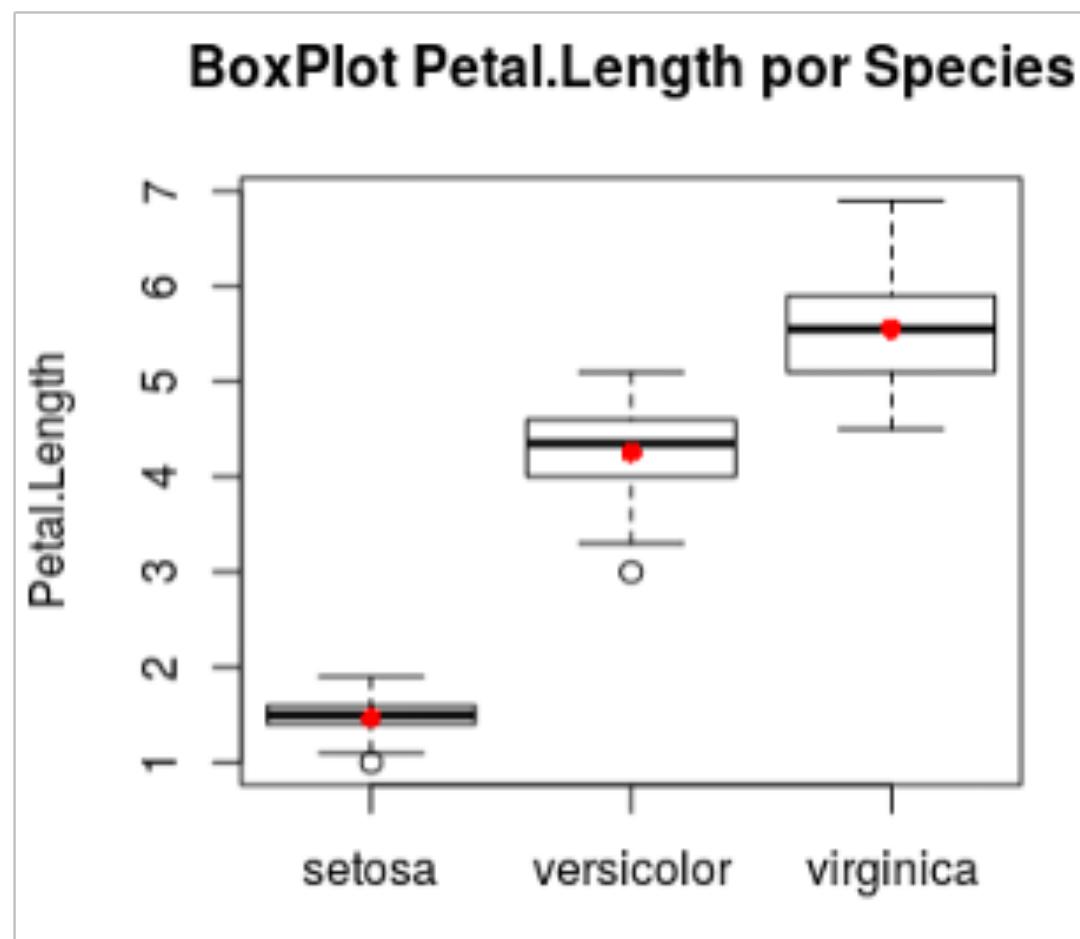
- Son valores **atípicos** que podemos encontrar en una simple variable.
- El problema de los enfoques univariados es que son buenos para detección de extremos pero no en otros casos.

Multivariado

- Los valores atípicos multivariados se pueden encontrar en un espacio n-dimensional.
- Para detectar valores atípicos en espacios n-dimensionales es necesario ajustar un modelo.

Análisis de Outliers

Métodos Univariados



- IQR: Analizar los valores que están por fuera del IRQ
- Z-score y Z-score Modificado
- Identificar valores extremos a partir de 1, 2 o 3 desvíos de la media.

Métodos Multivariados

- Análisis globales: Clustering.
 - Utilizando medidas de distancia como Mahalanobis.
- Local Outlier Factor (LOF)
 - Es un método de detección de outliers basado en distancias.
 - Calcula un score de *outlier* a partir de una distancia que se normaliza por densidad.
- Métodos basados en árboles de búsqueda: IsolationForest

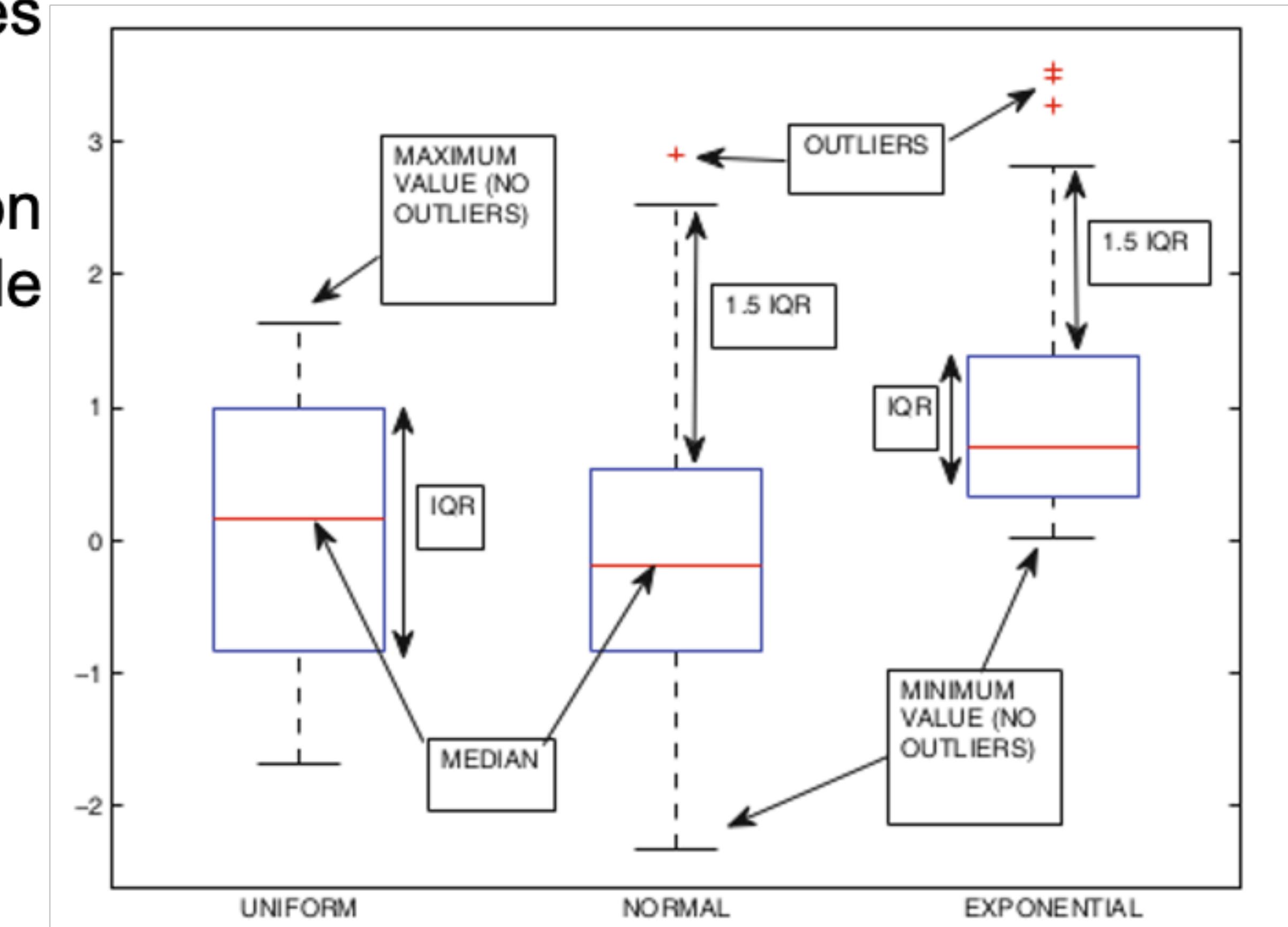
Métodos Univariados

Análisis de Box-Plot

Los Box-Plots permiten visualizar valores extremos univariados.

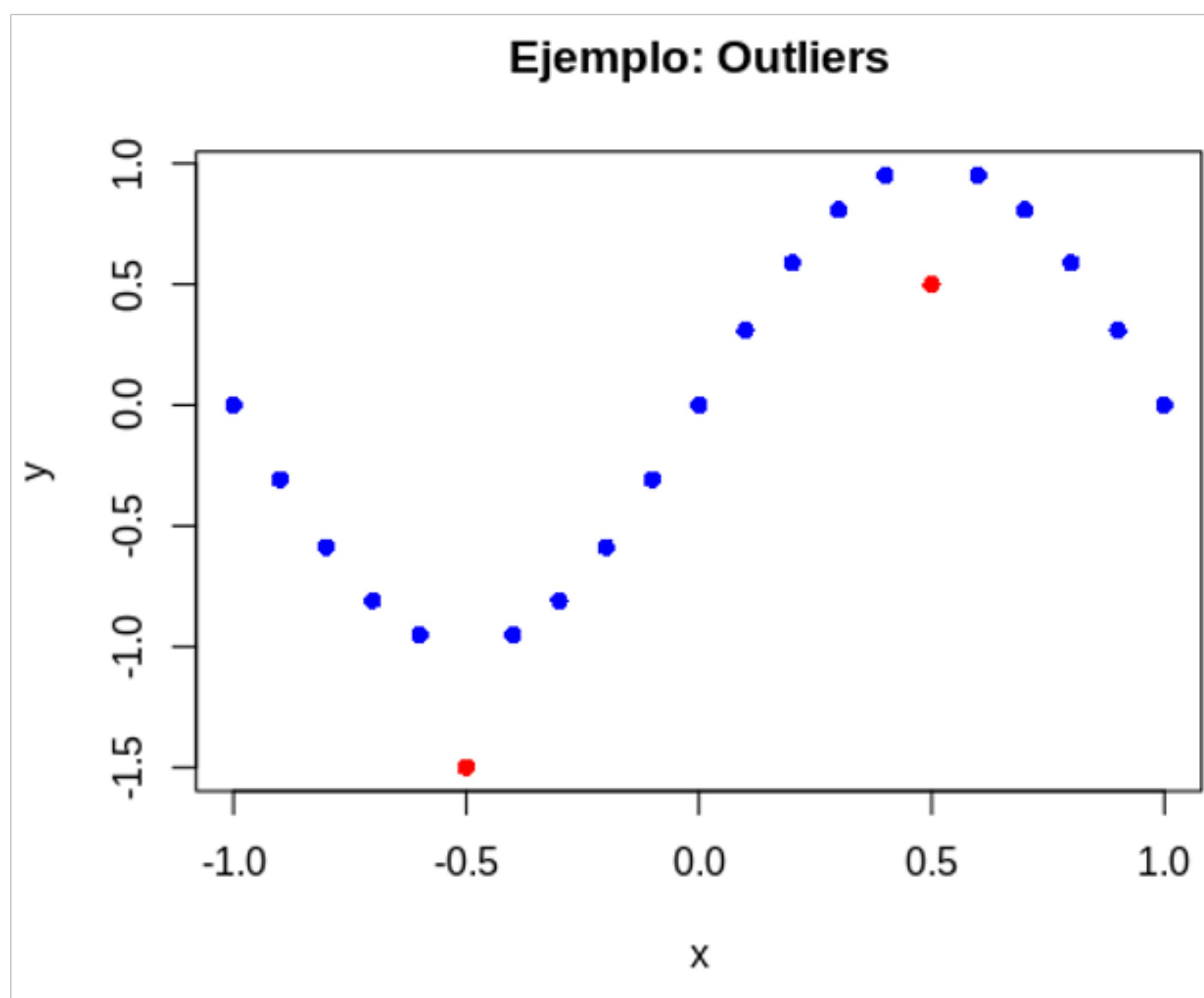
Las estadísticas de una distribución univariada se resumen en términos de cinco cantidades:

- Mínimo/máximo (bigotes)
- Primer y tercer cuantil (caja)
- Mediana (línea media de la caja)
- $IQR = Q3 - Q1$
- Generalmente la regla de decisión es $+/- 1.5*IQR$

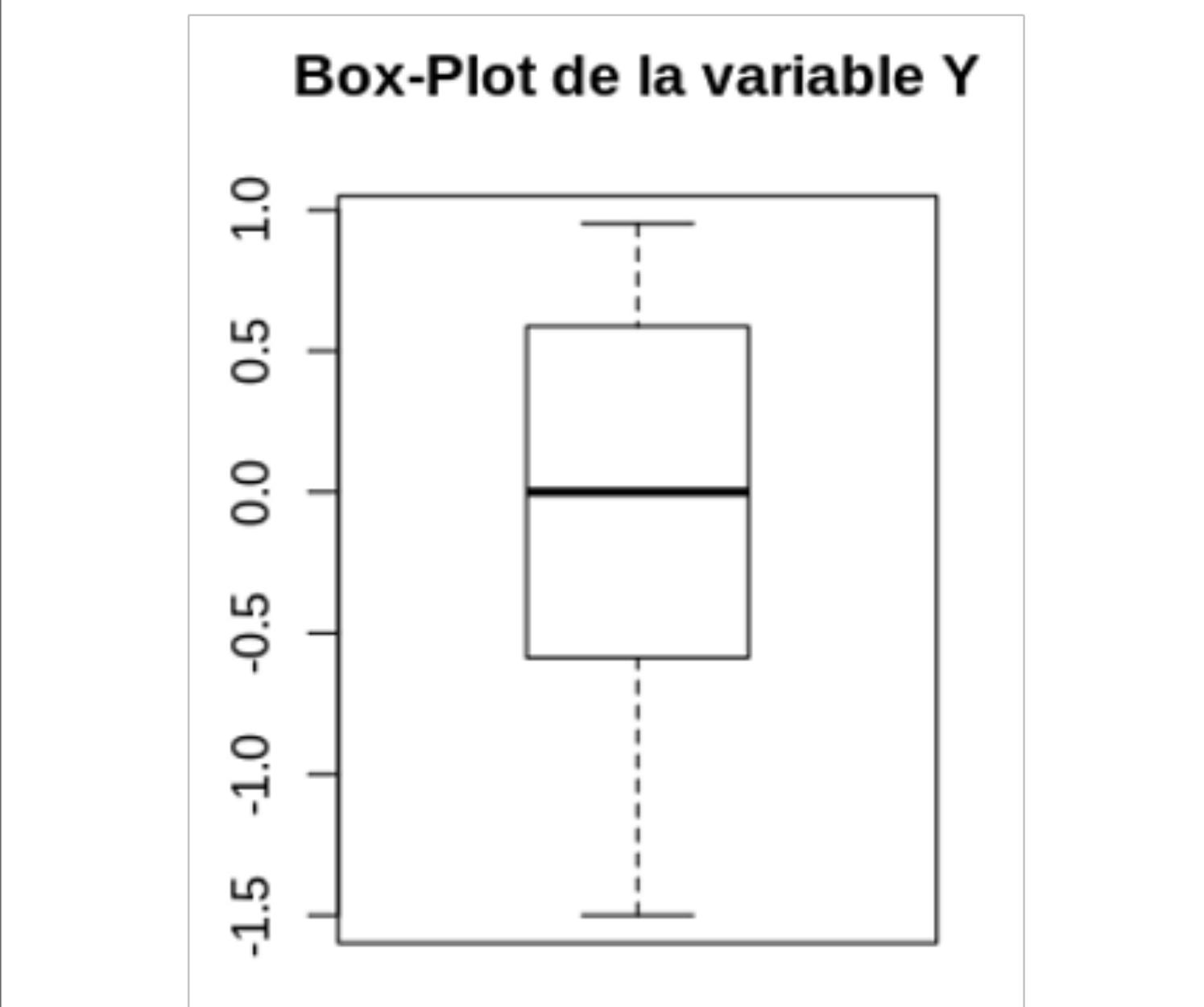


Limitaciones del Box-Plot

En el scatter se observan dos valores atípicos.



¿Qué pasa con el box-plot?

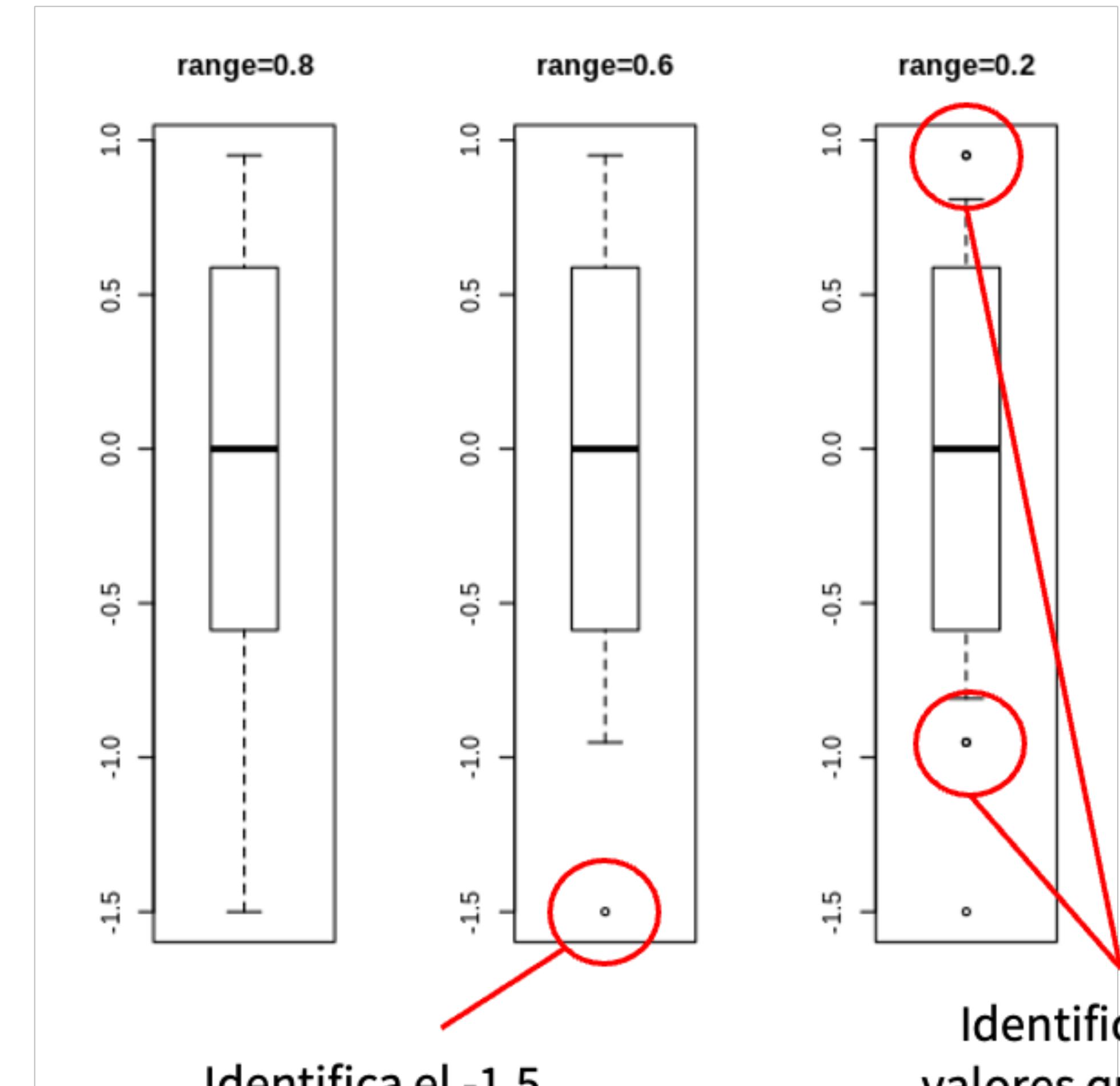


Limitaciones del Box-Plot

Podemos intervenir modificando el largo de los bigotes del gráfico.

En R podemos hacerlo con el parámetro *range*.

También es posible modificar los cuantiles para regular el tamaño de las cajas.



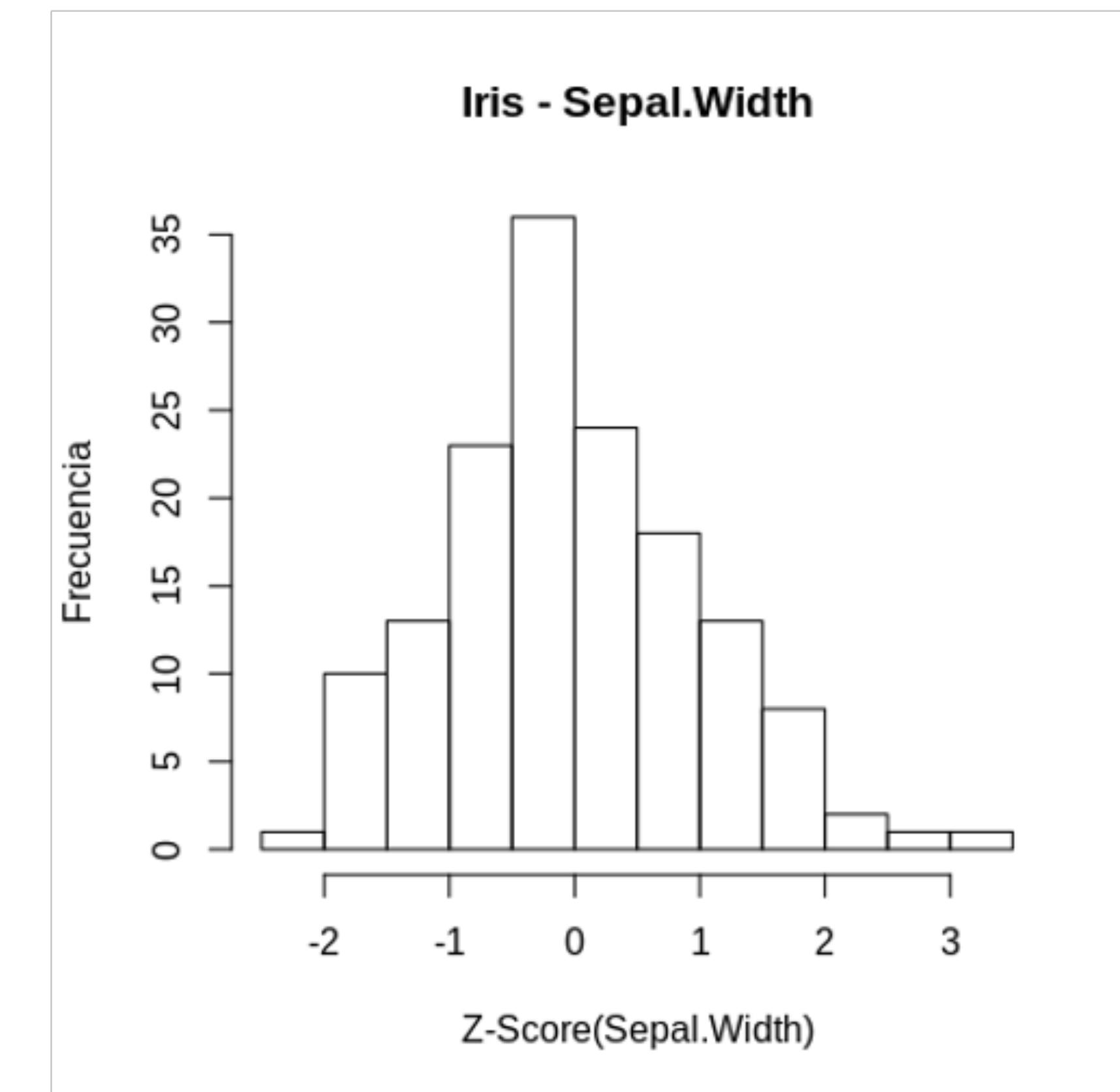
Identifica otros
valores que no son
outliers

Z-Score

Z-Score es una métrica que indica cuántas desviaciones estándar tiene una observación de la media muestral, asumiendo una distribución gaussiana.

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Cuando calculamos Z-Score para cada muestra debemos fijar un umbral.
- Un valor como “regla de oro” es $Z > 3$



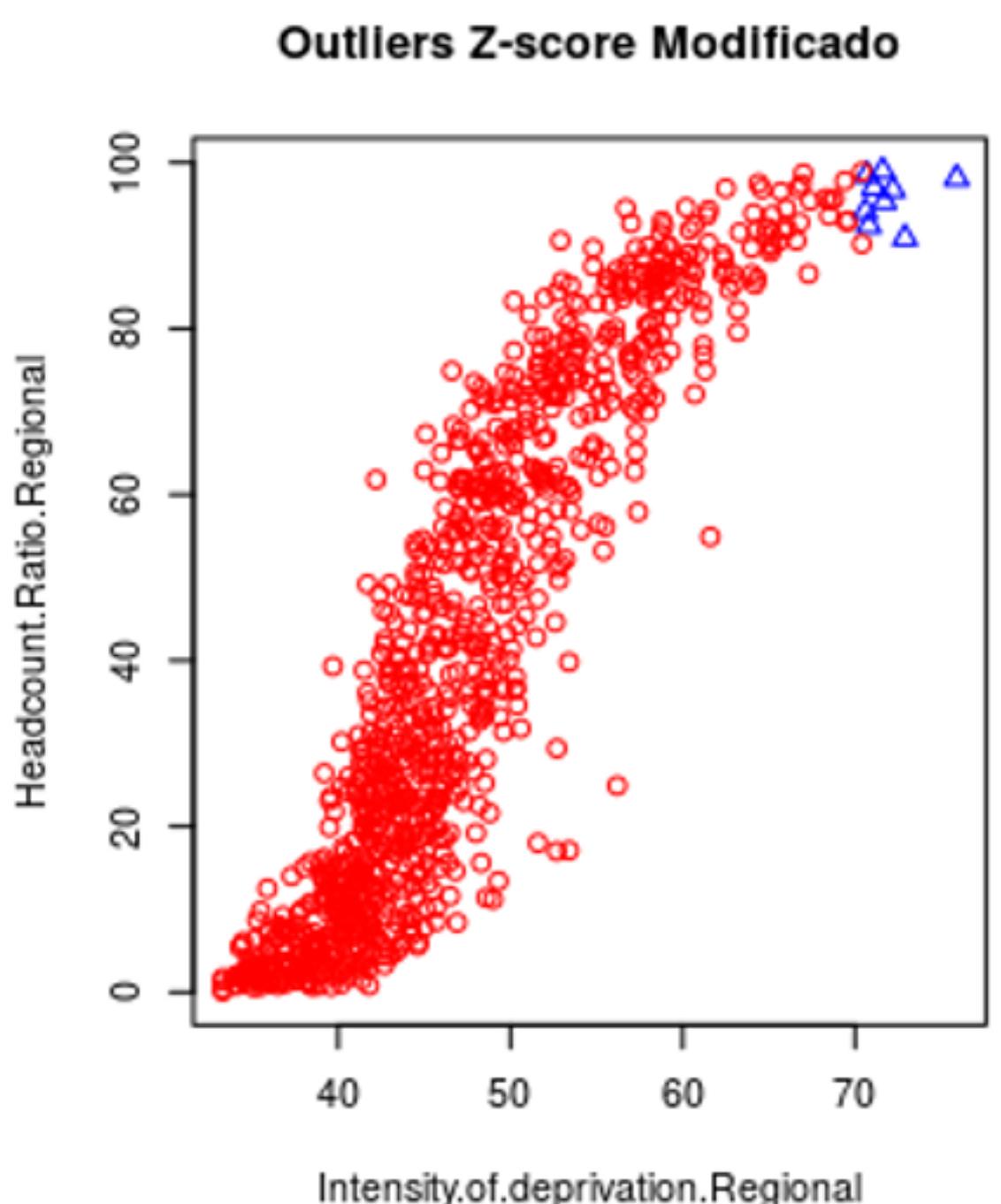
Z-Score Modificado

- La media de la muestra y la desviación estándar de la muestra, pueden verse afectados por los valores extremos presentes en los datos

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

$$MAD = median\{|x_i - \tilde{x}|\}$$

- Regla de oro. Valores mayores a 3.5 son considerados outliers



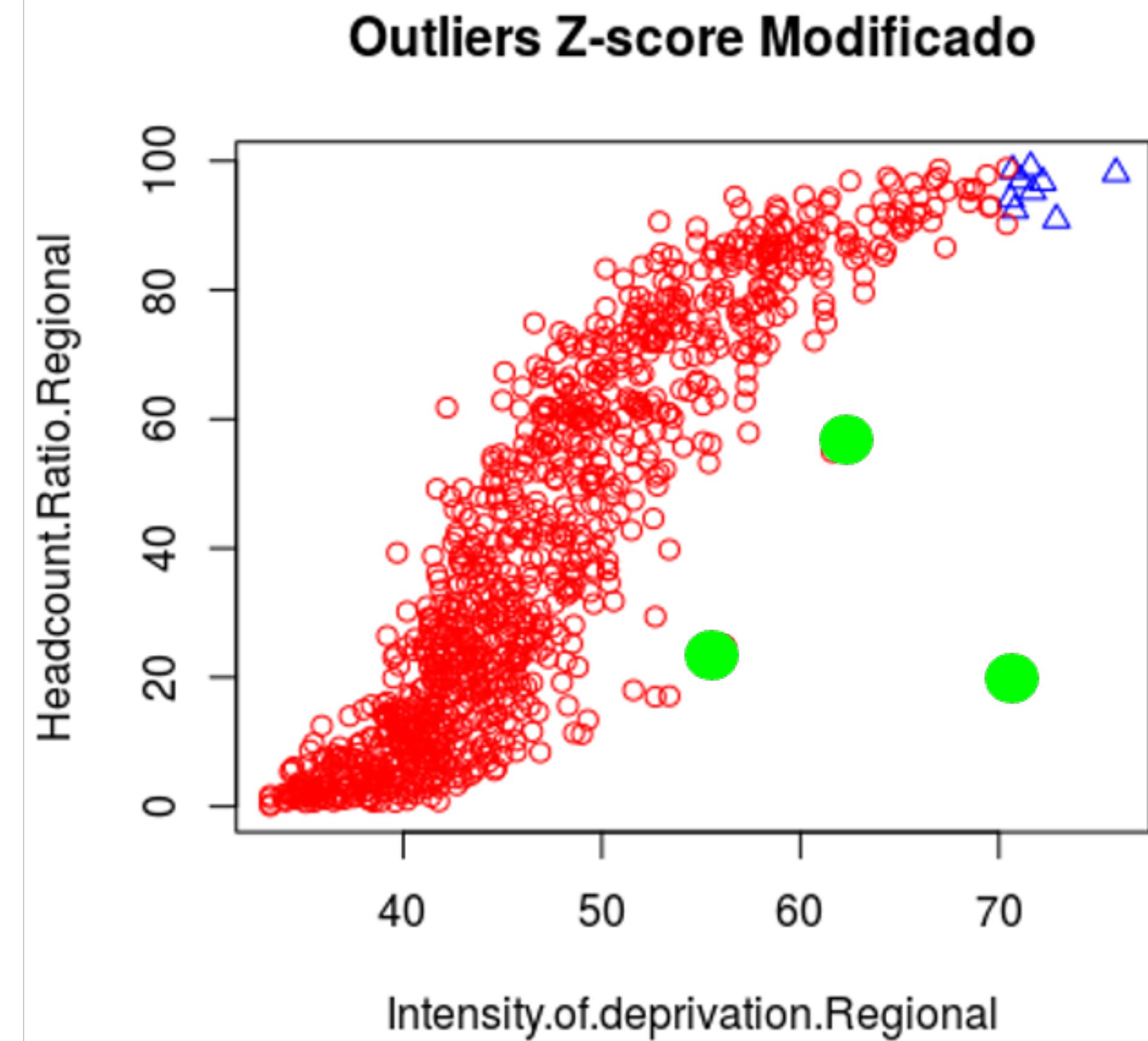
Métodos Multivariados

Problema

Una forma de tratar con tales valores es sacar los valores más altos y más bajos de una variable.

Esto puede funcionar bastante bien, pero no tiene en cuenta las **combinaciones de variables**.

¿Qué ocurre con los casos ● ?



Distancia de Mahalanobis

Es una medida de distancia entre el punto

$$\vec{x} = (x_1, x_2, x_3, \dots, x_N)^T$$

y un conjunto de observaciones con media

$$\vec{\mu} = (\mu_1, \mu_2, \mu_3, \dots, \mu_N)^T$$

y una matriz de covarianza S .

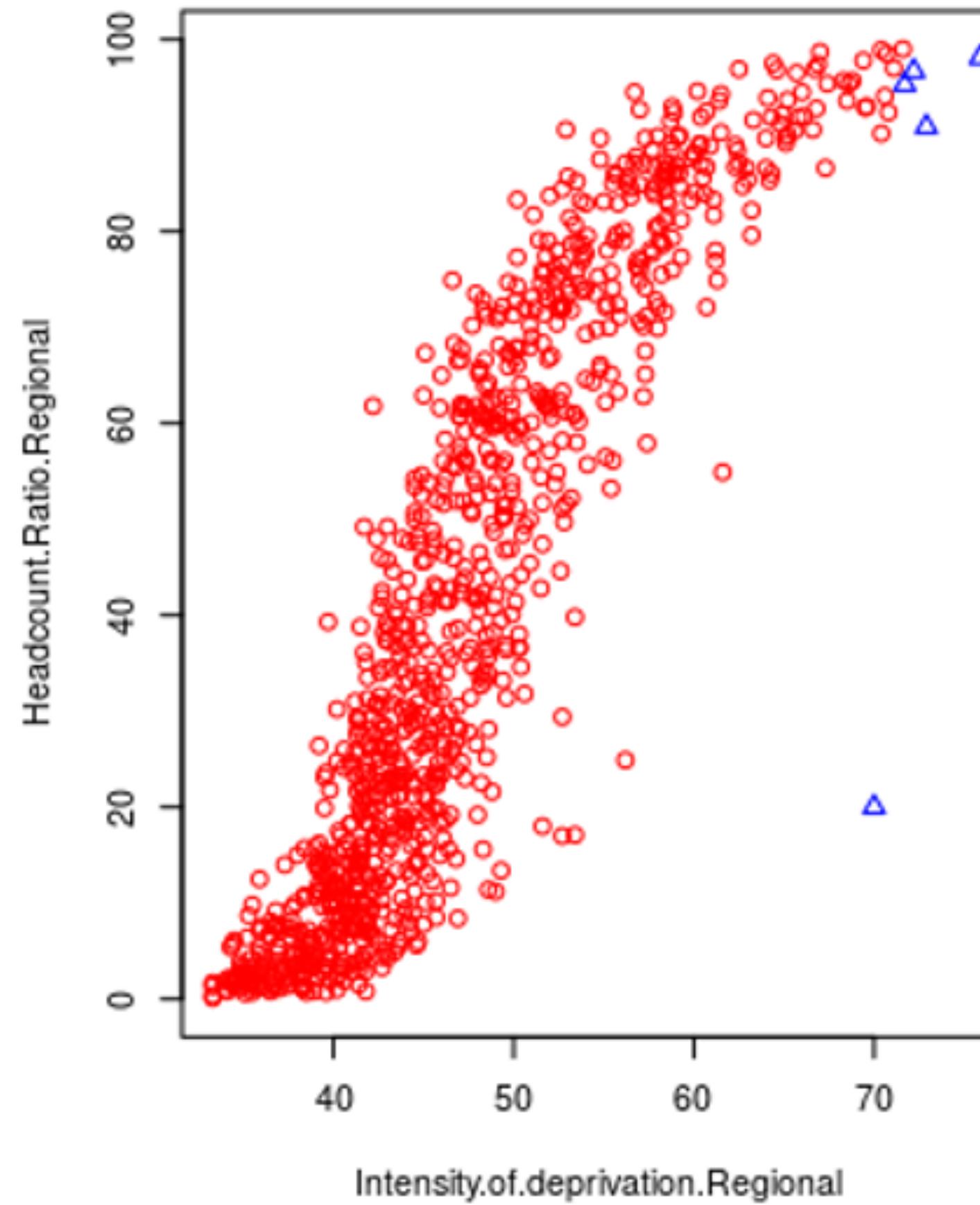
$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

Matriz de distancias con
respecto a la media

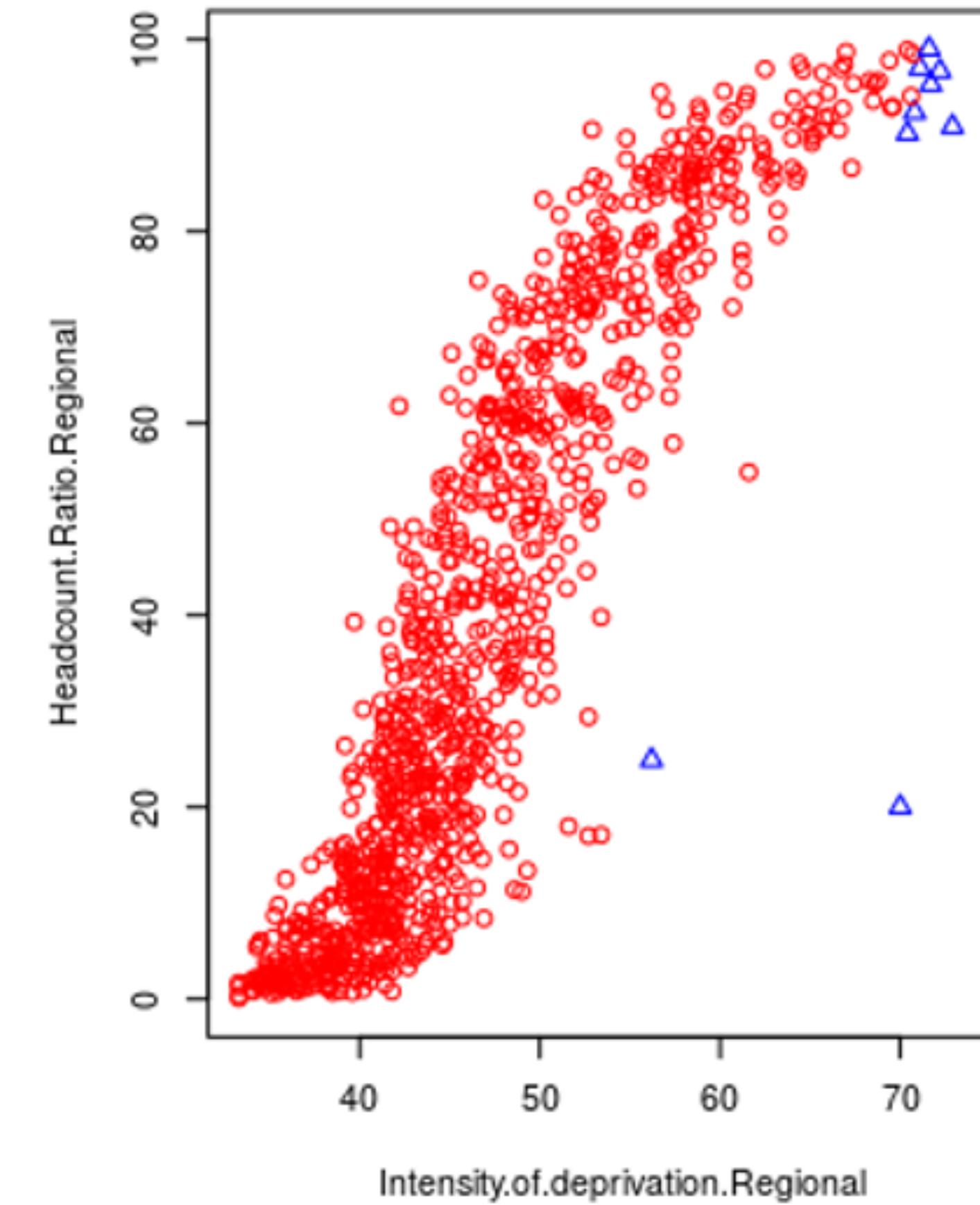
Inversa de la matriz de
covarianzas

Distancia de Mahalanobis

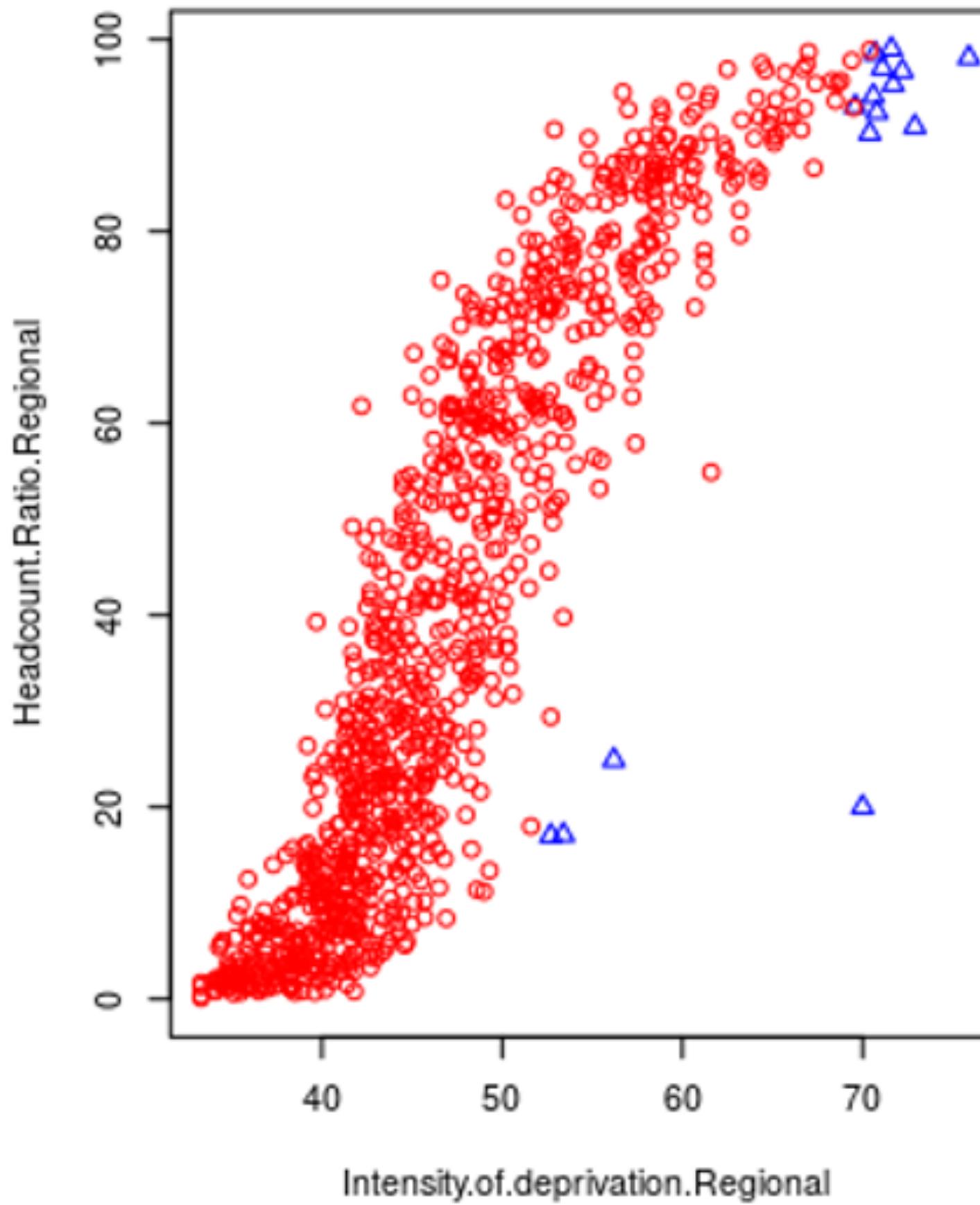
Mahalanobis (Top@5)



Mahalanobis (Top@10)



Mahalanobis (Top@15)

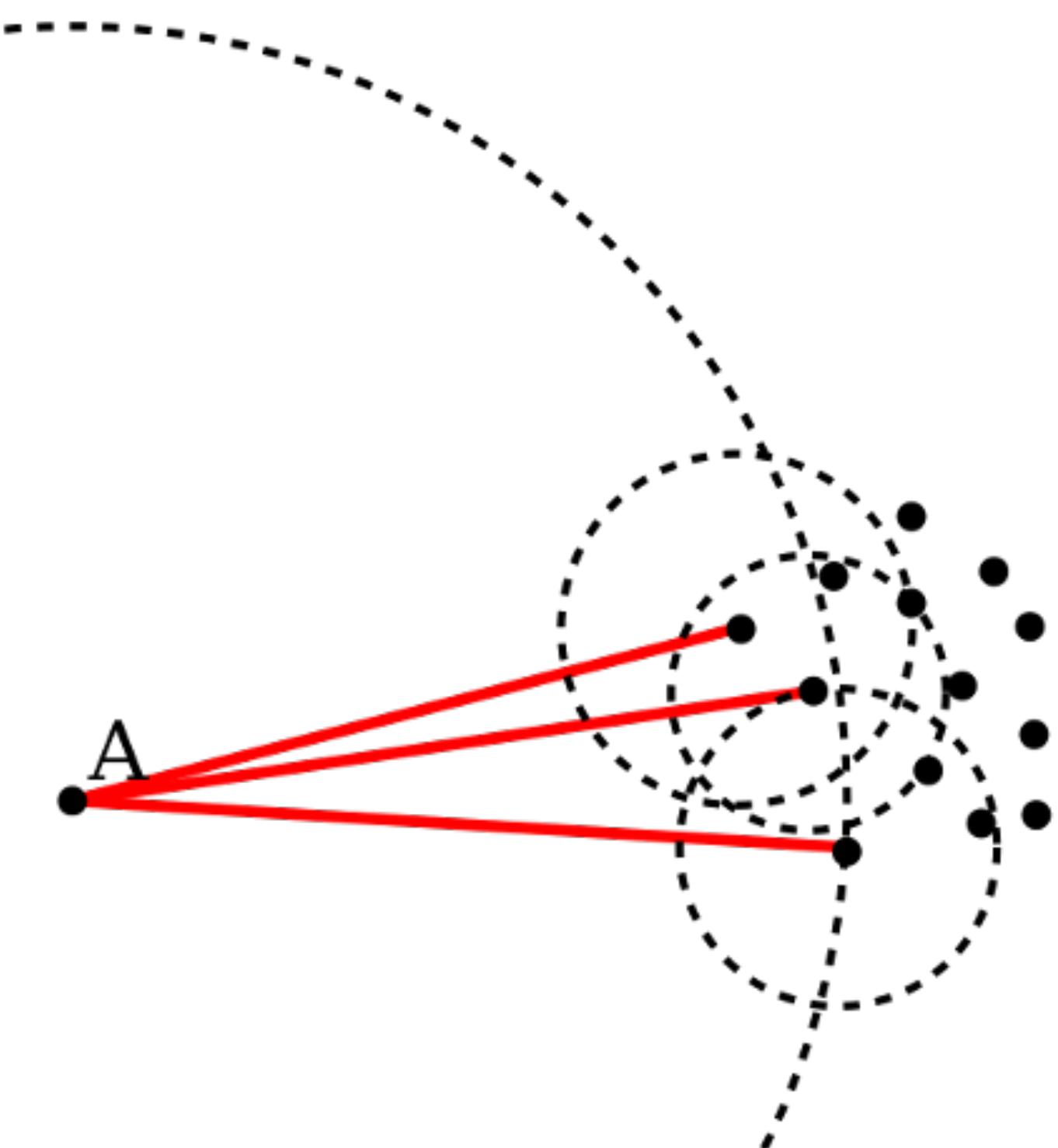


Local Outlier Factor (LOF)

El método del LOF valora puntos en un conjunto de datos multivariados cuyas filas se supone que se generan de forma independiente.

LOF es un **método basado en densidad** que utiliza la búsqueda de vecinos más cercanos.

El método calcula los **scores** para cada uno de los puntos a partir de la tasa promedio de densidad de los puntos vecinos con respecto a si mismo.



Local Outlier Factor (LOF)

K-Distancia de una observación \mathbf{o} , $dist_k(\mathbf{o})$: Es la distancia entre \mathbf{o} y su k -ésimo **vecino más cercano (kNN)**

Vecindad de \mathbf{o} , $N_k(\mathbf{o}) = \{\mathbf{o}' \mid \mathbf{o}' \text{ en } D, dist(\mathbf{o}, \mathbf{o}') \leq dist_k(\mathbf{o})\}$

La distancia de vecindad son todas las distancias dentro del radio de $dist_k(\mathbf{o})$

Distancia esperada: de \mathbf{o}' a \mathbf{o} : $reachdist(\mathbf{o} \leftarrow \mathbf{o}') = max(dist_k(\mathbf{o}), dist(\mathbf{o}, \mathbf{o}'))$

- Densidad Local Esperada de \mathbf{o} :
(Local Reachability Density)

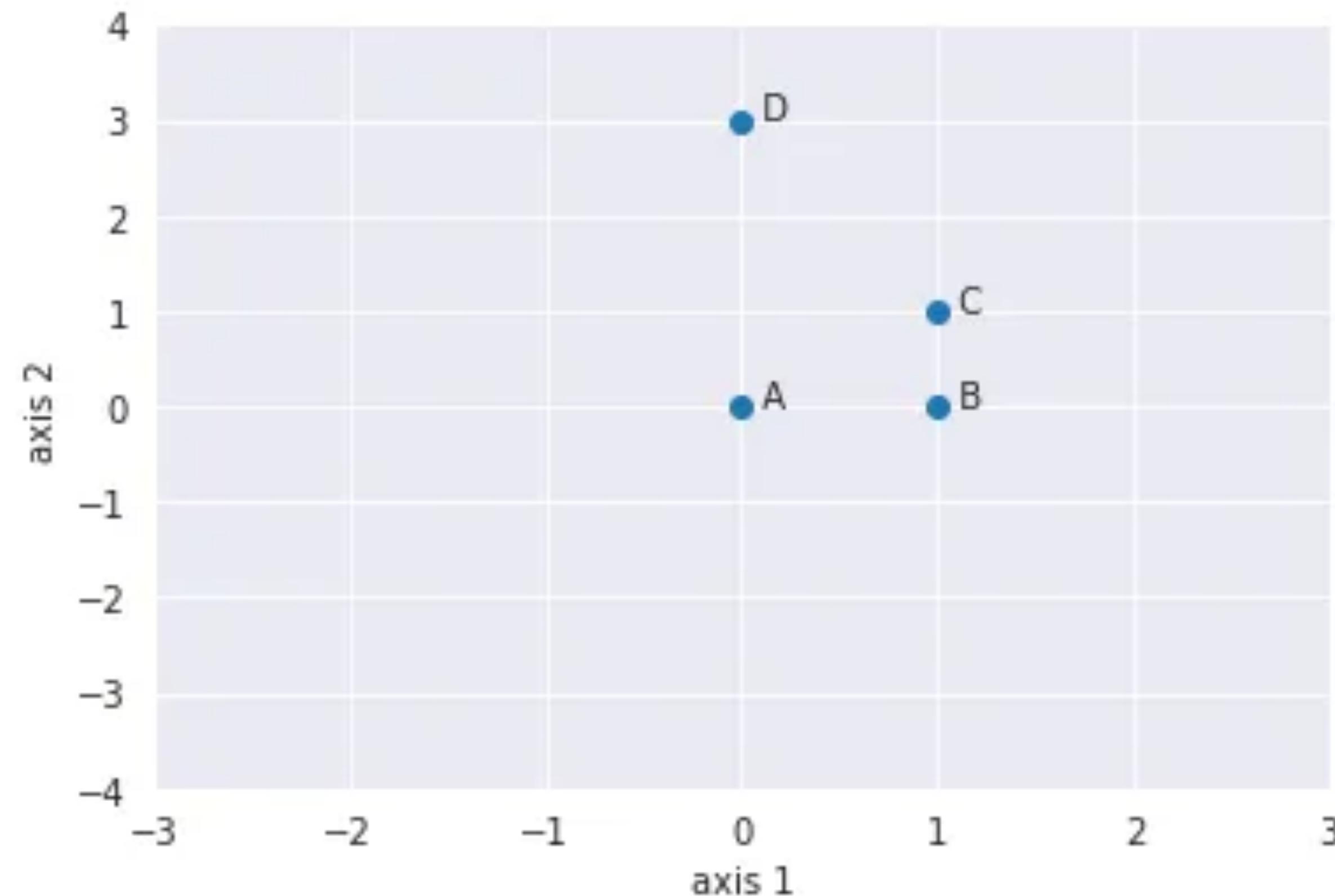
$$lrd_k(\mathbf{o}) = \frac{\|N_k(\mathbf{o})\|}{\sum_{\mathbf{o}' \in N_k(\mathbf{o})} reachdist_k(\mathbf{o}' \leftarrow \mathbf{o})}$$

$$LOF_k(\mathbf{o}) = \frac{\sum_{\mathbf{o}' \in N_k(\mathbf{o})} \frac{lrd_k(\mathbf{o}')}{lrd_k(\mathbf{o})}}{\|N_k(\mathbf{o})\|} =$$

$$\sum_{\mathbf{o}' \in N_k(\mathbf{o})} lrd_k(\mathbf{o}') \cdot \sum_{\mathbf{o}' \in N_k(\mathbf{o})} reachdist_k(\mathbf{o}' \leftarrow \mathbf{o})$$

Local Outlier Factor (LOF)-Ejemplo

- K -Distancia
- K-ésimo Vecino, $N_k(o)$
- LRD
- LOF

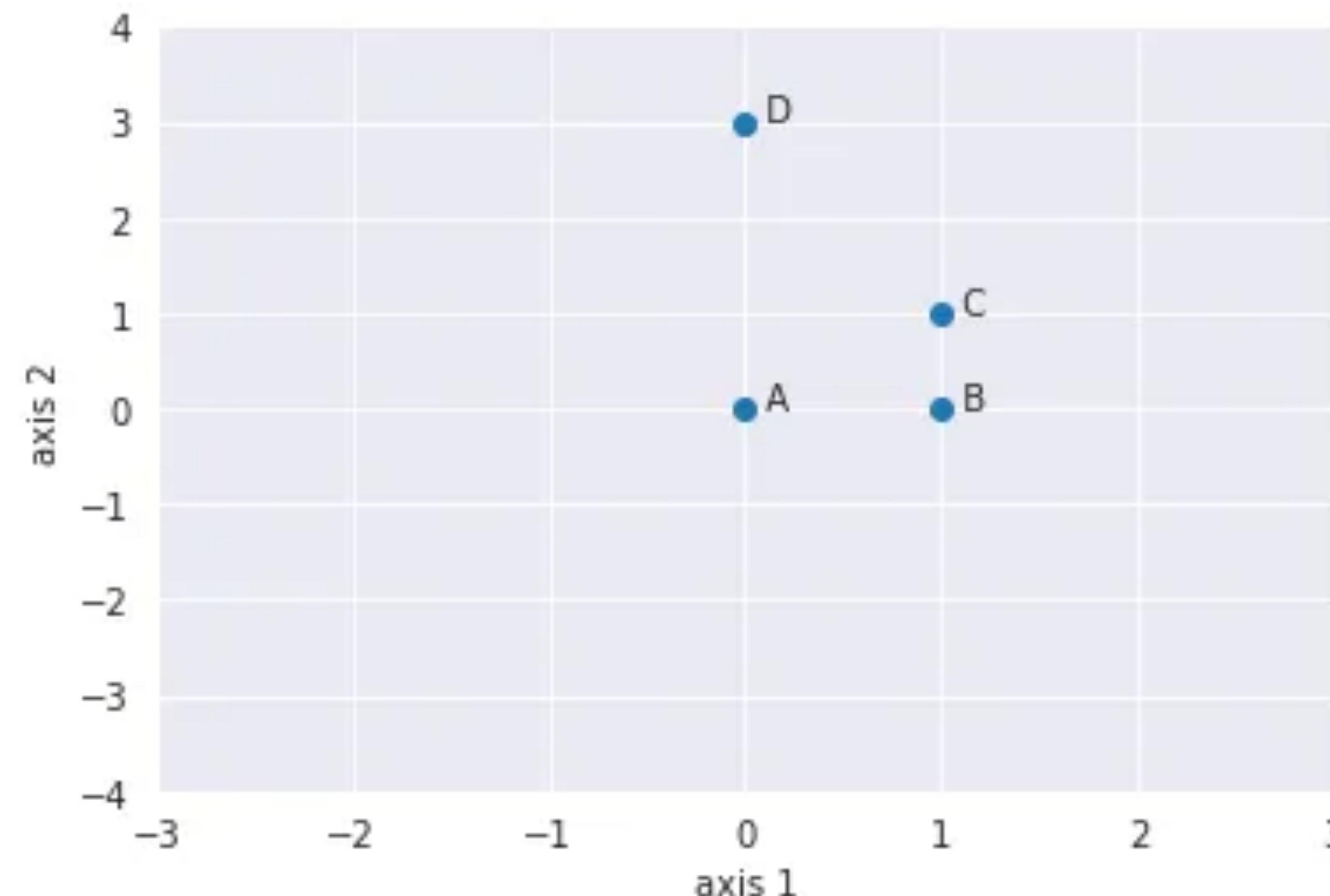


A(0,0), B(1,0), C(1,1), and D(0,3)

Local Outlier Factor (LOF)-Ejemplo

- K -Distancia, K=2. Usando distancia de Manhattan.

Manhattan_Distance(A,B) =	1
Manhattan_Distance(A,C) =	2
Manhattan_Distance(A,D) =	3
Manhattan_Distance(B,C) =	1
Manhattan_Distance(B,D) =	4
Manhattan_Distance(C,D) =	3



A(0,0), B(1,0), C(1,1), and D(0,3)

Local Outlier Factor (LOF)-Ejemplo

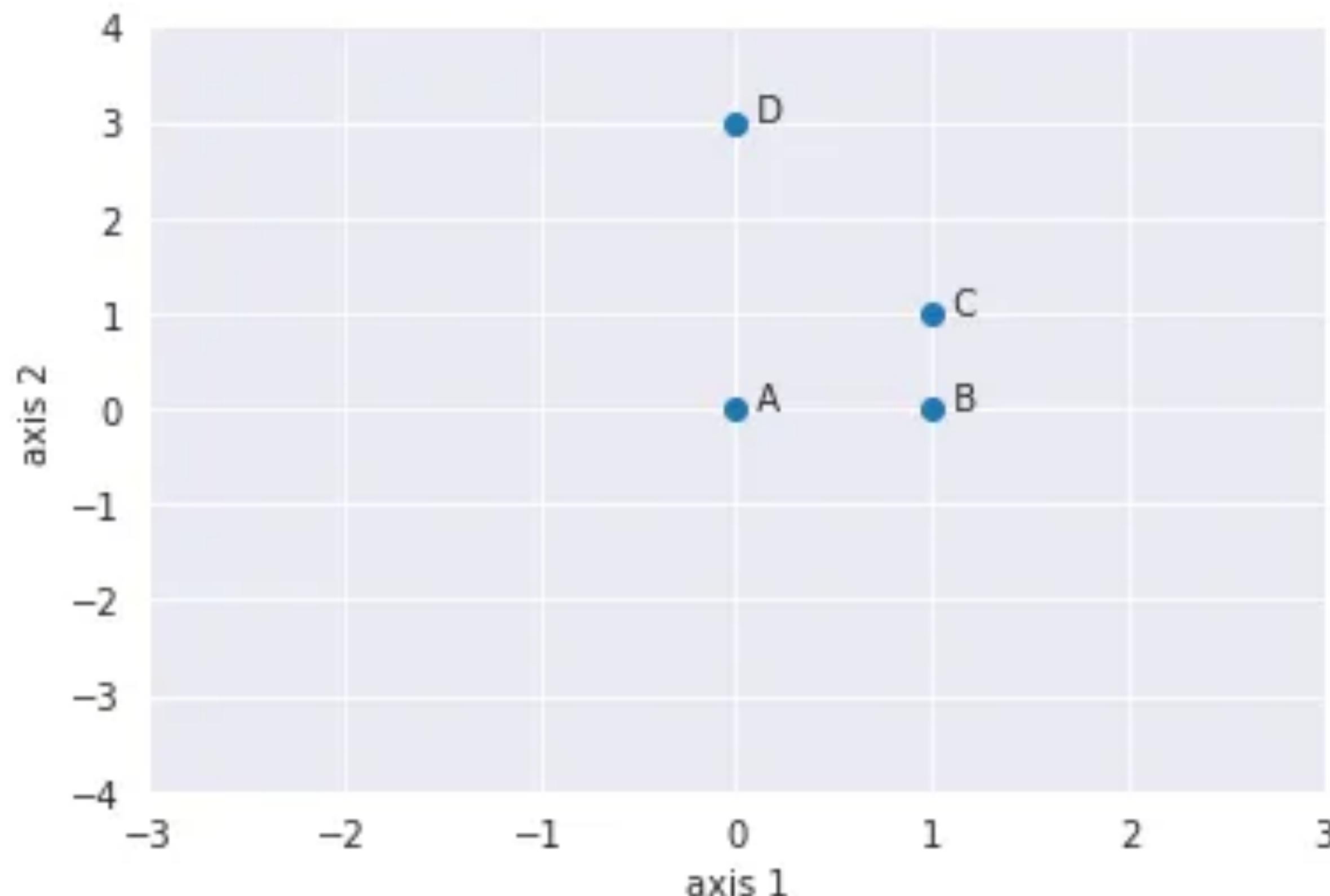
- **K-ésimo Vecino, $N_k(o)$**

K-neighborhood (A) = {B,C} , $\|N_2(A)\| = 2$

K-neighborhood (B) = {A,C}, $\|N_2(B)\| = 2$

K-neighborhood (C)= {B,A}, $\|N_2(C)\| = 2$

K-neighborhood (D) = {A,C}, $\|N_2(D)\| = 2$



A(0,0), B(1,0), C(1,1), and D(0,3)

Local Outlier Factor (LOF)-Ejemplo

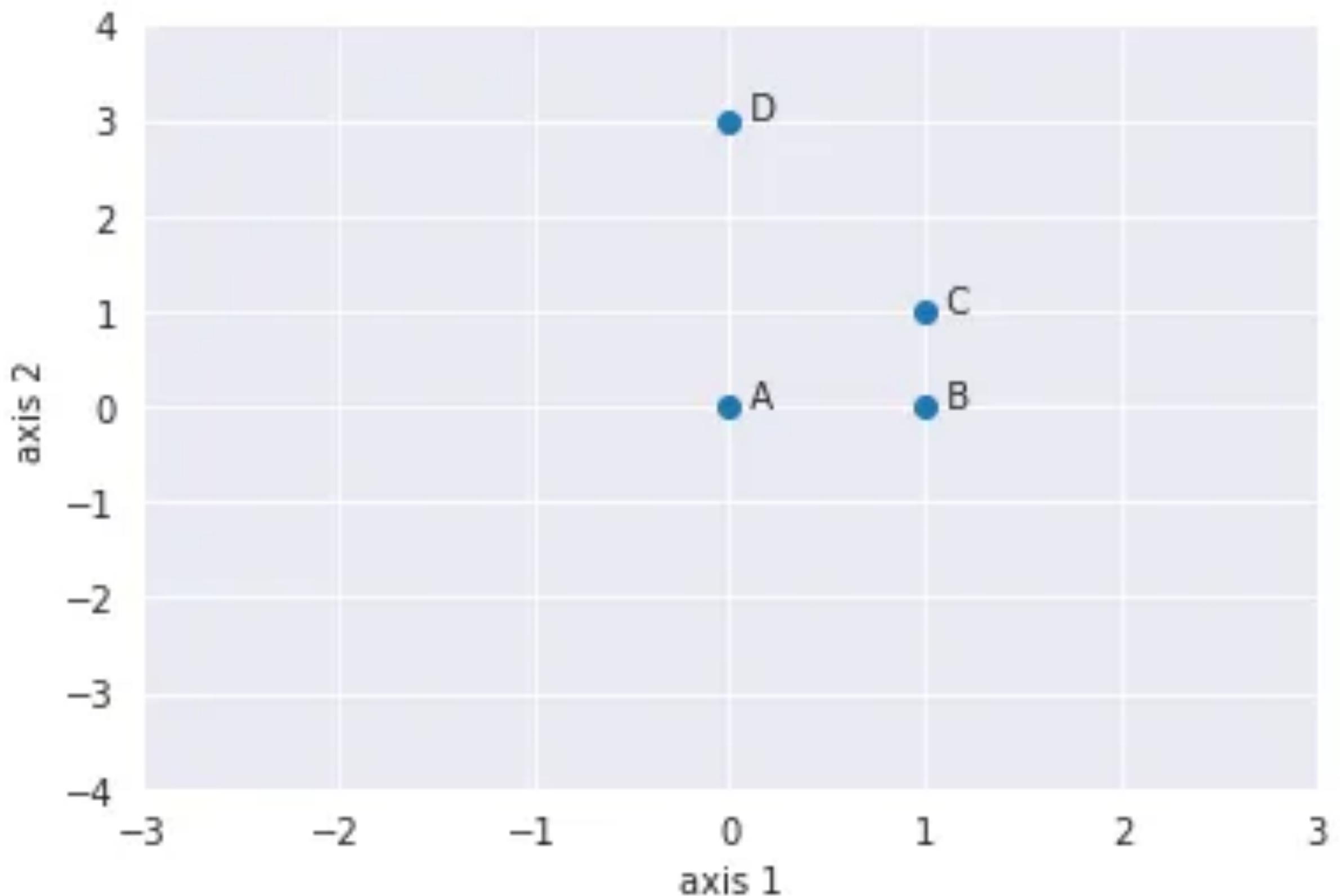
- LRD

$$LRD_2(A) = \frac{1}{\frac{RD(A,B)+RD(A,C)}{\|N_2(A)\|}} = \frac{\frac{1}{1+2}}{2} = 0.667$$

$$LRD_2(B) = \frac{1}{\frac{RD(B,A)+RD(B,C)}{\|N_2(B)\|}} = \frac{\frac{1}{2+2}}{2} = 0.50$$

$$LRD_2(C) = \frac{1}{\frac{RD(C,B)+RD(C,A)}{\|N_2(C)\|}} = \frac{\frac{1}{1+2}}{2} = 0.667$$

$$LRD_2(D) = \frac{1}{\frac{RD(D,A)+RD(D,C)}{\|N_2(D)\|}} = \frac{\frac{1}{3+3}}{2} = 0.337$$



A(0,0), B(1,0), C(1,1), and D(0,3)

Local Outlier Factor (LOF)-Ejemplo

- LOF

$$LOF_2(A) = \frac{LRD_2(B) + LRD_2(C)}{\|N_2(A)\|} \times \frac{1}{LRD_2(A)} = \frac{0.5 + 0.667}{2} \times \frac{1}{0.667} = 0.87$$

$$LOF_2(B) = \frac{LRD_2(A) + LRD_2(C)}{\|N_2(B)\|} \times \frac{1}{LRD_2(B)} = \frac{0.667 + 0.667}{2} \times \frac{1}{0.5} = 1.334$$

$$LOF_2(C) = \frac{LRD_2(B) + LRD_2(A)}{\|N_2(C)\|} \times \frac{1}{LRD_2(C)} = \frac{0.5 + 0.667}{2} \times \frac{1}{0.667} = 0.87$$

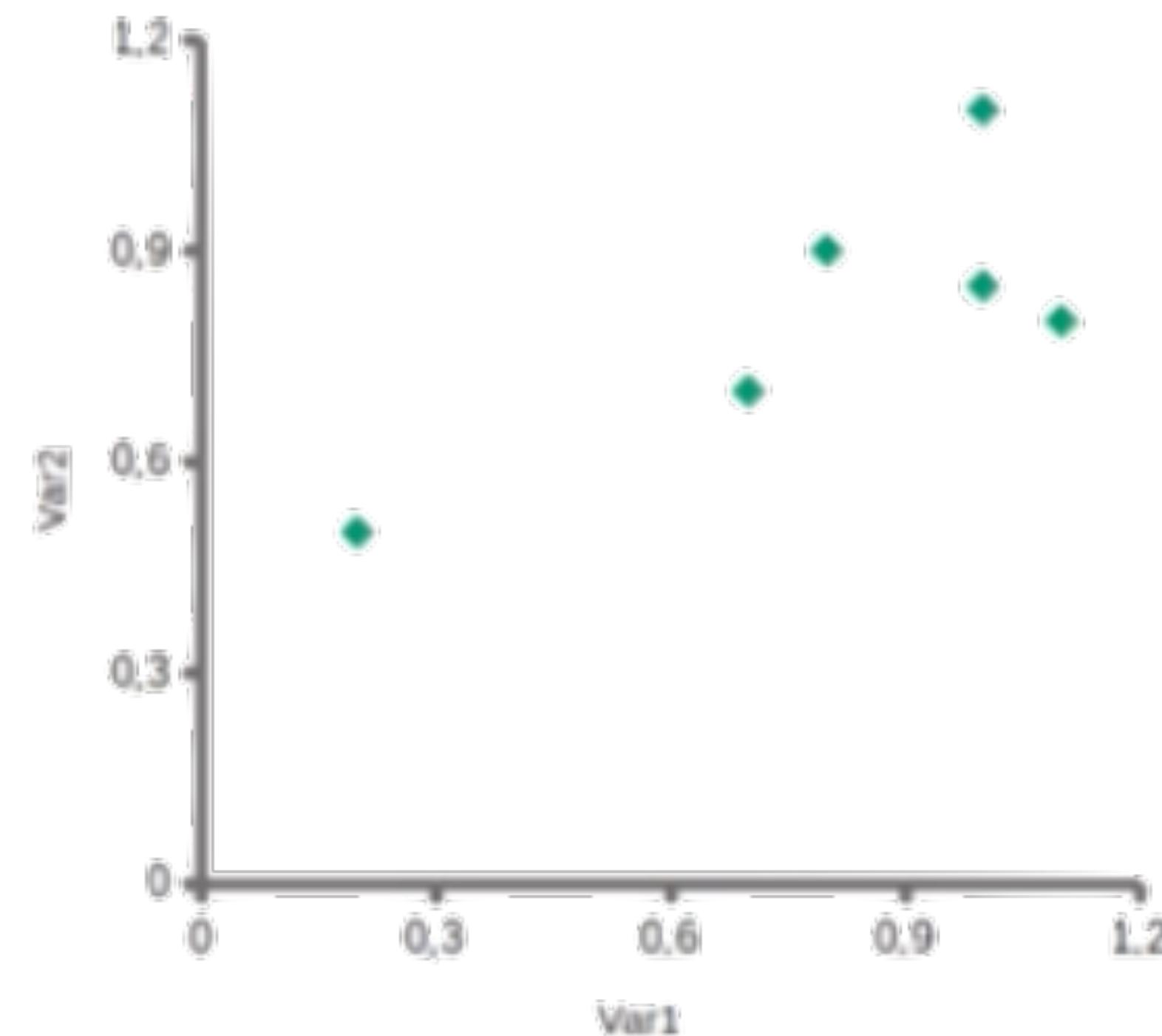
$$LOF_2(D) = \frac{LRD_2(A) + LRD_2(C)}{\|N_2(D)\|} \times \frac{1}{LRD_2(D)} = \frac{0.667 + 0.667}{2} \times \frac{1}{0.337} = 2$$



A(0,0), B(1,0), C(1,1), and D(0,3)

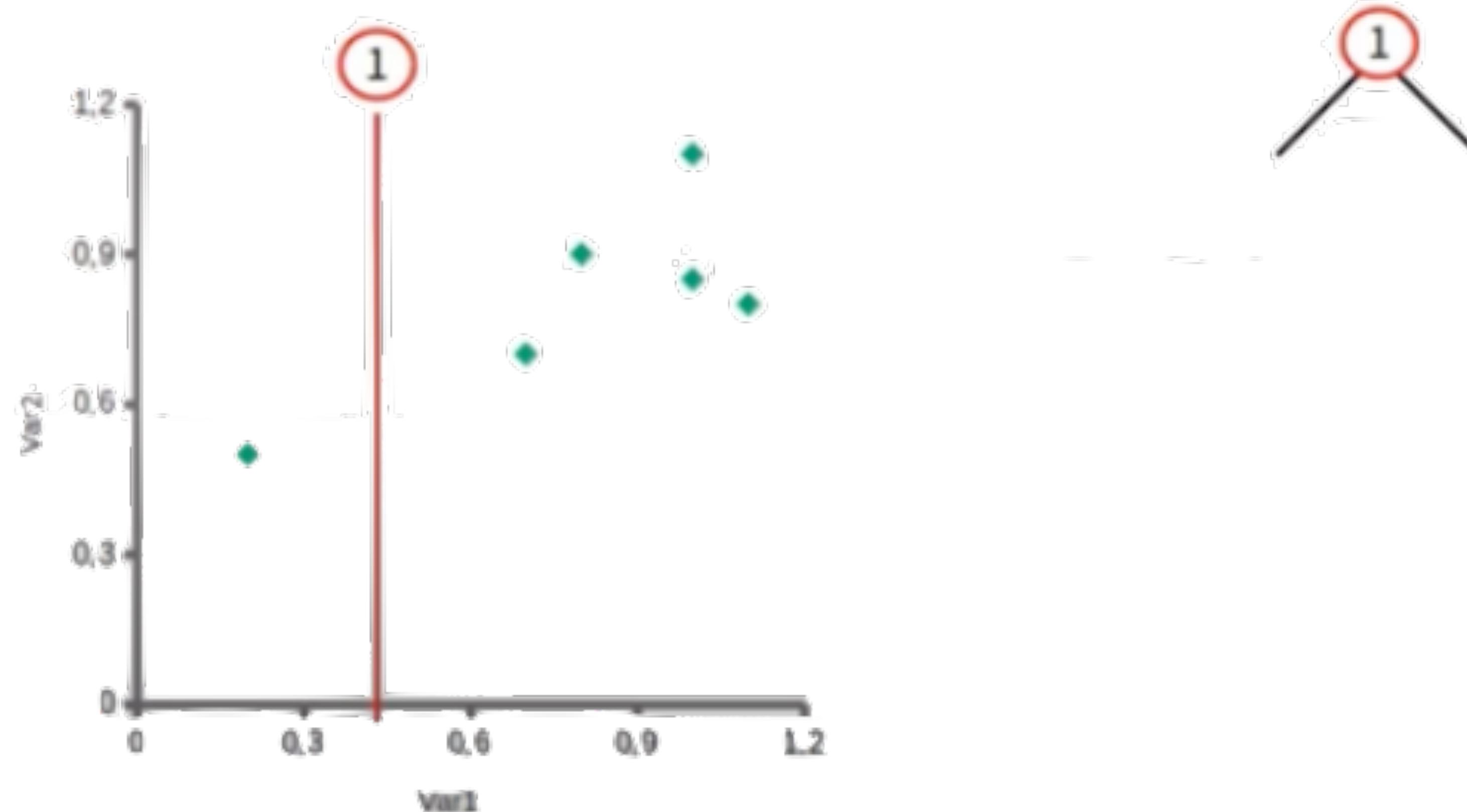
Isolation Forest

- **Idea principal:** las instancias de datos anómalas se pueden aislar de los datos normales mediante la partición recursiva del conjunto de datos.



Isolation Forest

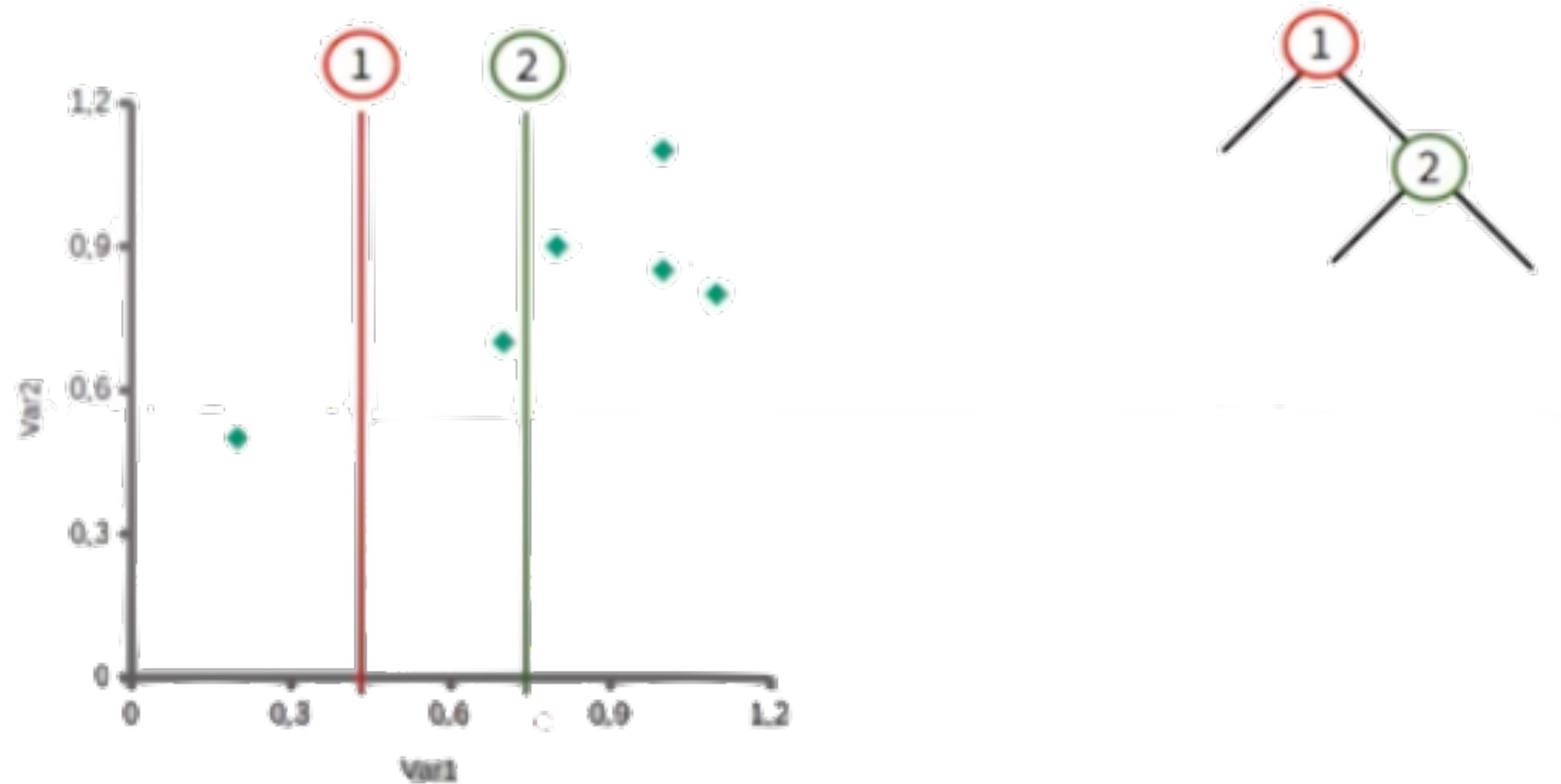
- **Idea principal:** las instancias de datos anómalas se pueden aislar de los datos normales mediante la partición recursiva del conjunto de datos.



Árbol Binario de Búsqueda (binary search tree:
https://es.wikipedia.org/wiki/%C3%81rbol_binario_de_b%C3%BAqueda

Isolation Forest

- Idea principal: las instancias de datos anómalas se pueden aislar de los datos normales mediante la partición recursiva del conjunto de datos.

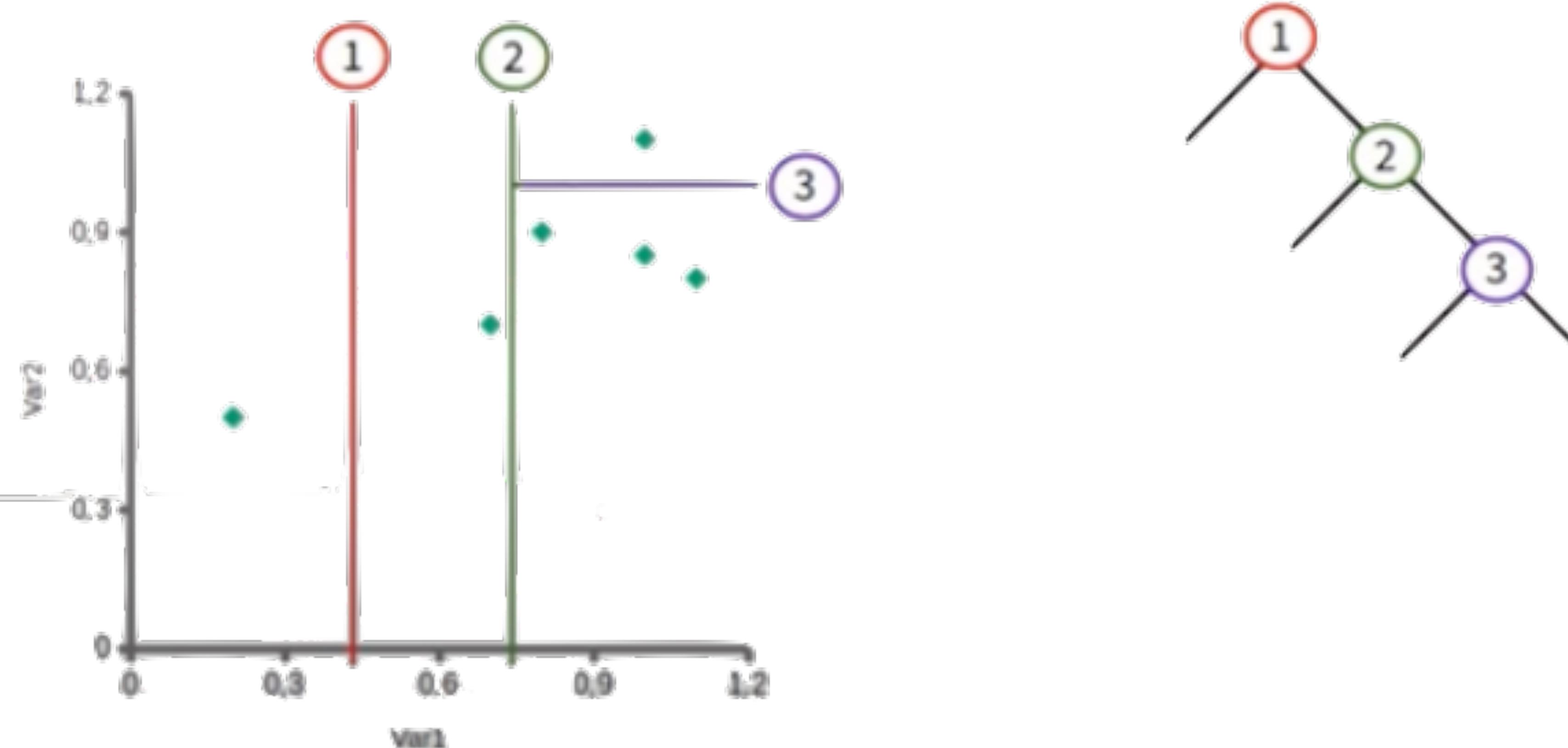


Árbol Binario de Búsqueda (binary search tree:

https://es.wikipedia.org/wiki/%C3%81rbol_binario_de_b%C3%A1squeda

Isolation Forest

- Idea principal: las instancias de datos anómalas se pueden aislar de los datos normales mediante la partición recursiva del conjunto de datos.

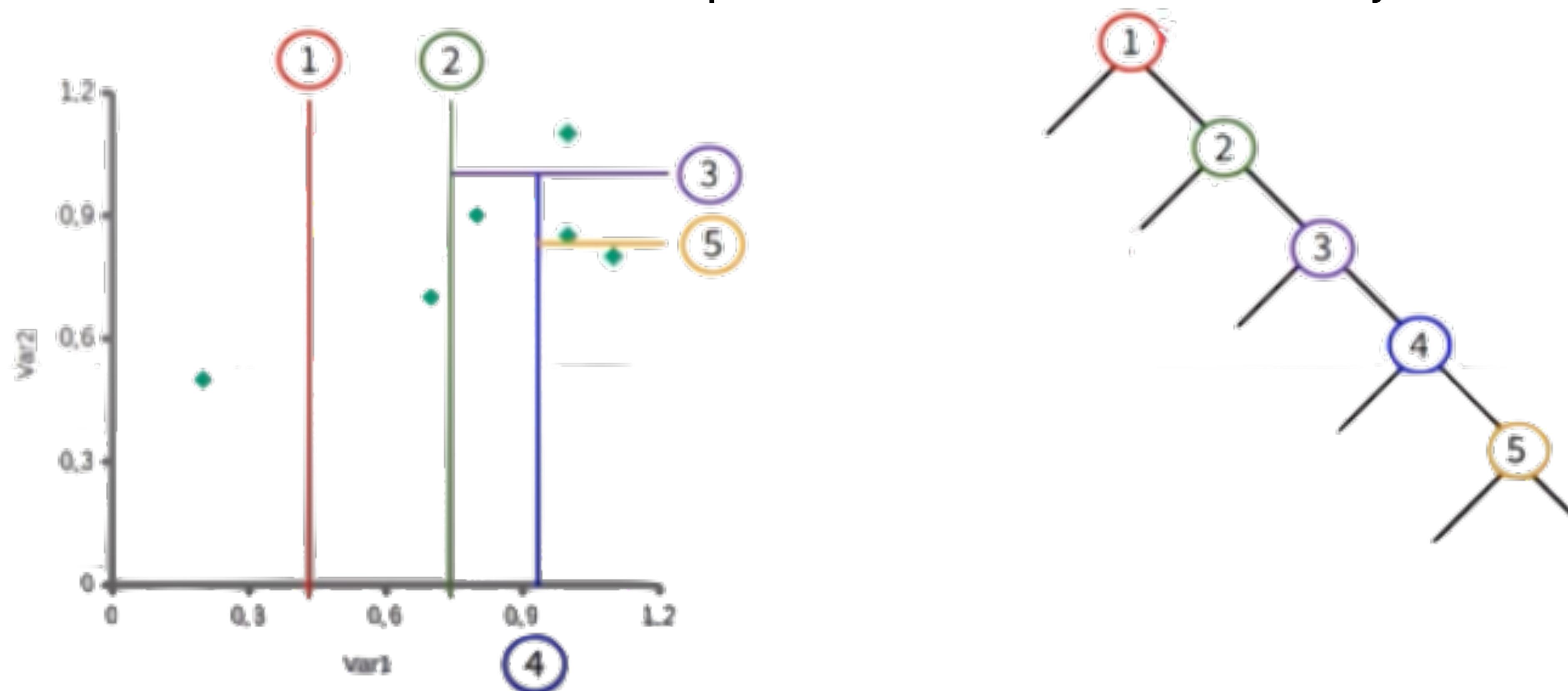


Árbol Binario de Búsqueda (binary search tree:

https://es.wikipedia.org/wiki/%C3%81rbol_binario_de_b%C3%BAqueda

Isolation Forest

- Idea principal: las instancias de datos anómalas se pueden aislar de los datos normales mediante la partición recursiva del conjunto de datos.

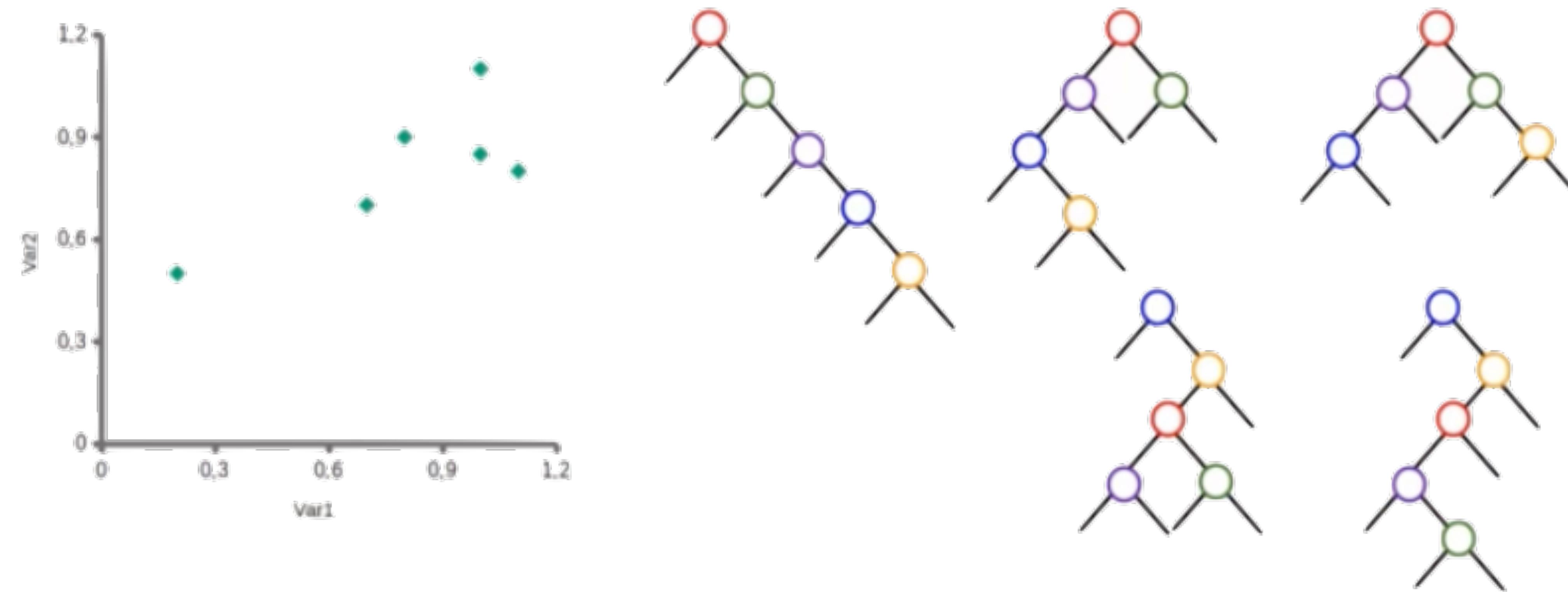


Árbol Binario de Búsqueda (binary search tree:

https://es.wikipedia.org/wiki/%C3%81rbol_binario_de_b%C3%BAqueda

Isolation Forest

- Idea principal: las instancias de datos anómalas se pueden aislar de los datos normales mediante la partición recursiva del conjunto de datos.



Score

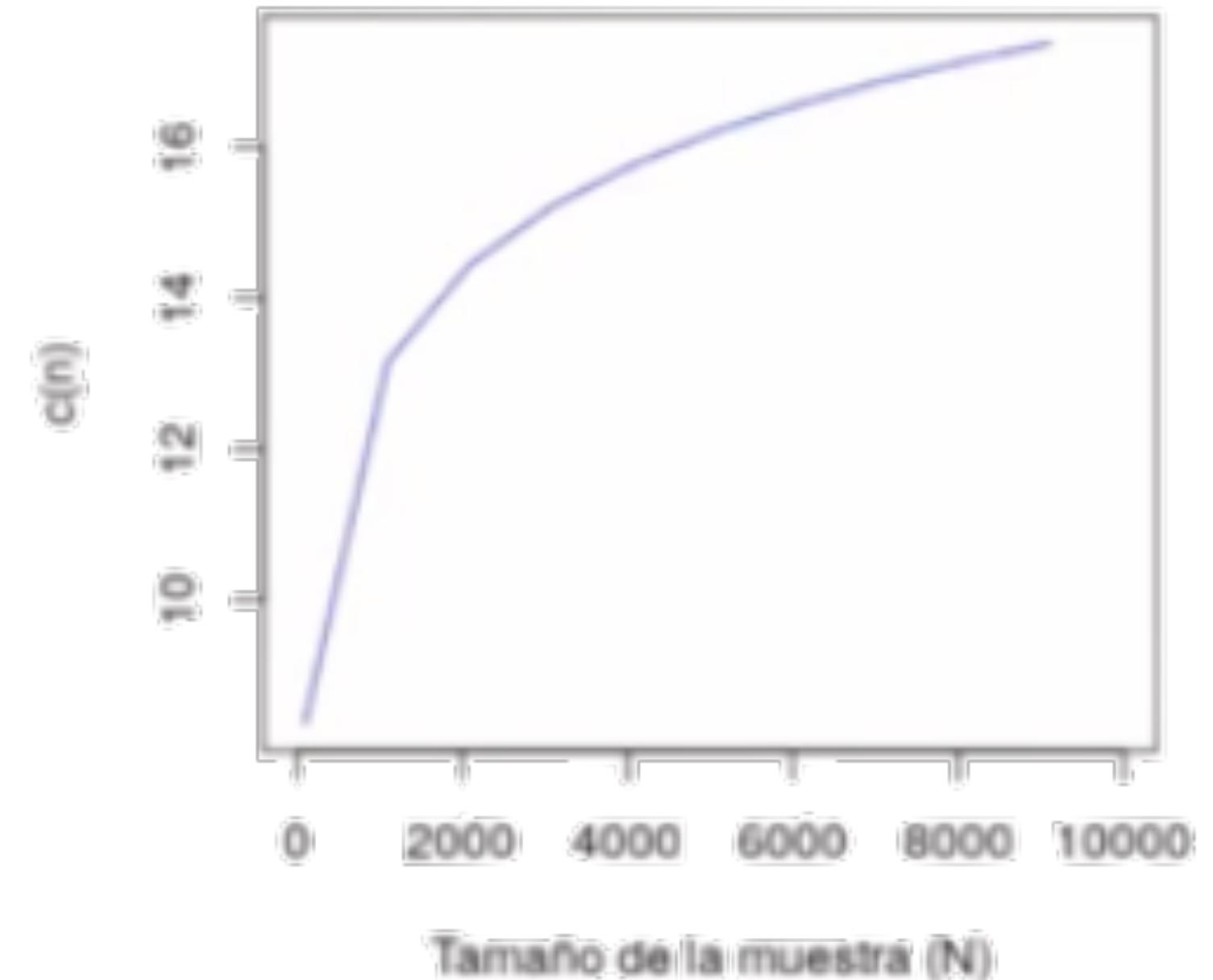
$$score = 2^{-\frac{E(h(x))}{c(n)}}$$

Donde:

- $h(x)$ es la longitud del camino a x
- $E(h(x))$ es la media de las alturas de x en todos los iForest
- $c(n)$ es una constante de normalización puede ser calculado como:

$$c(n) = 2H(n-1) = (2(n-1)/n),$$

$H(i)$ es el número armónico $H(i) = \ln(i) + 0.5772156649$ (constante de Euler)



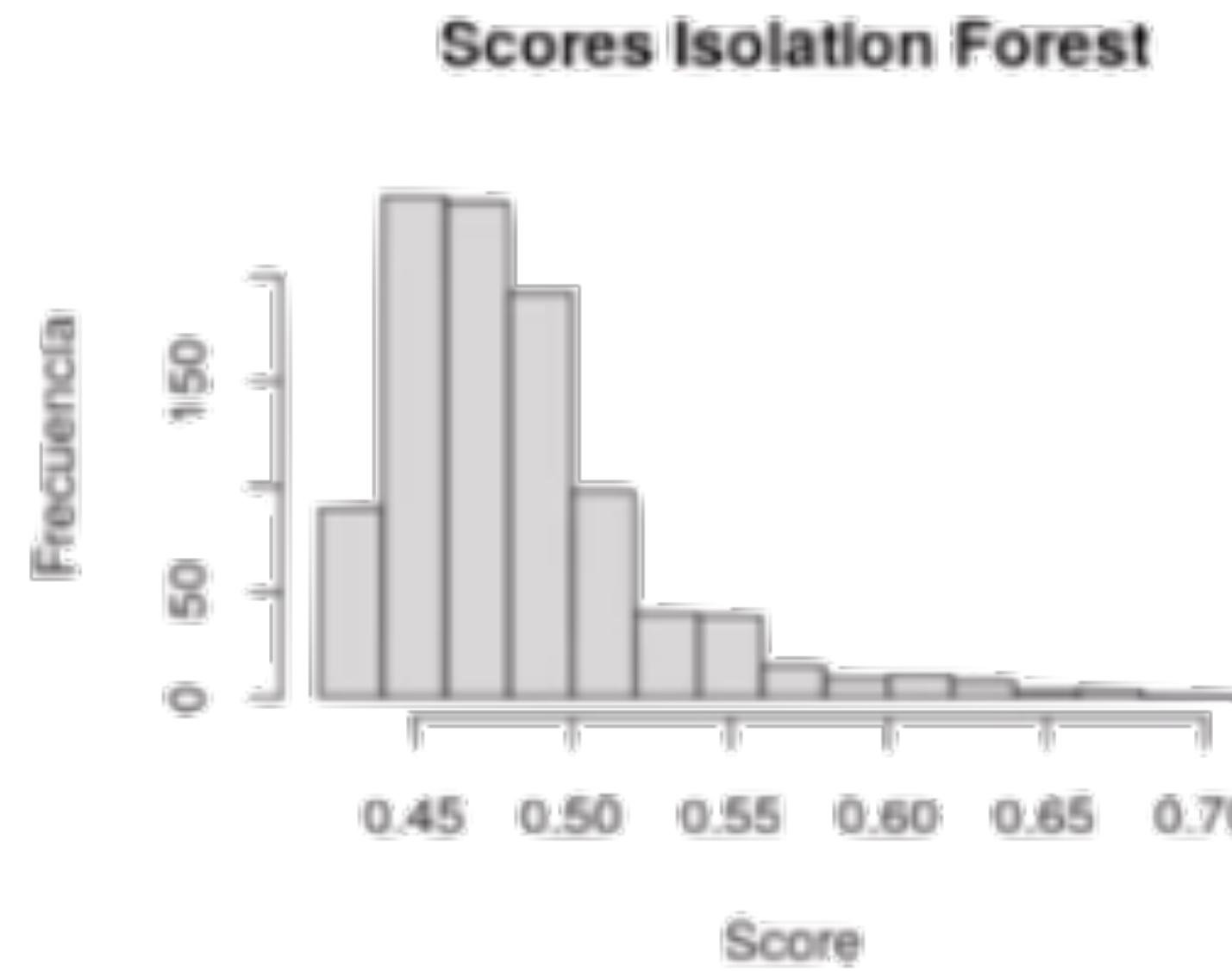
Árbol Binario de Búsqueda (binary search tree:

https://es.wikipedia.org/wiki/%C3%81rbol_binario_de_b%C3%BAqueda

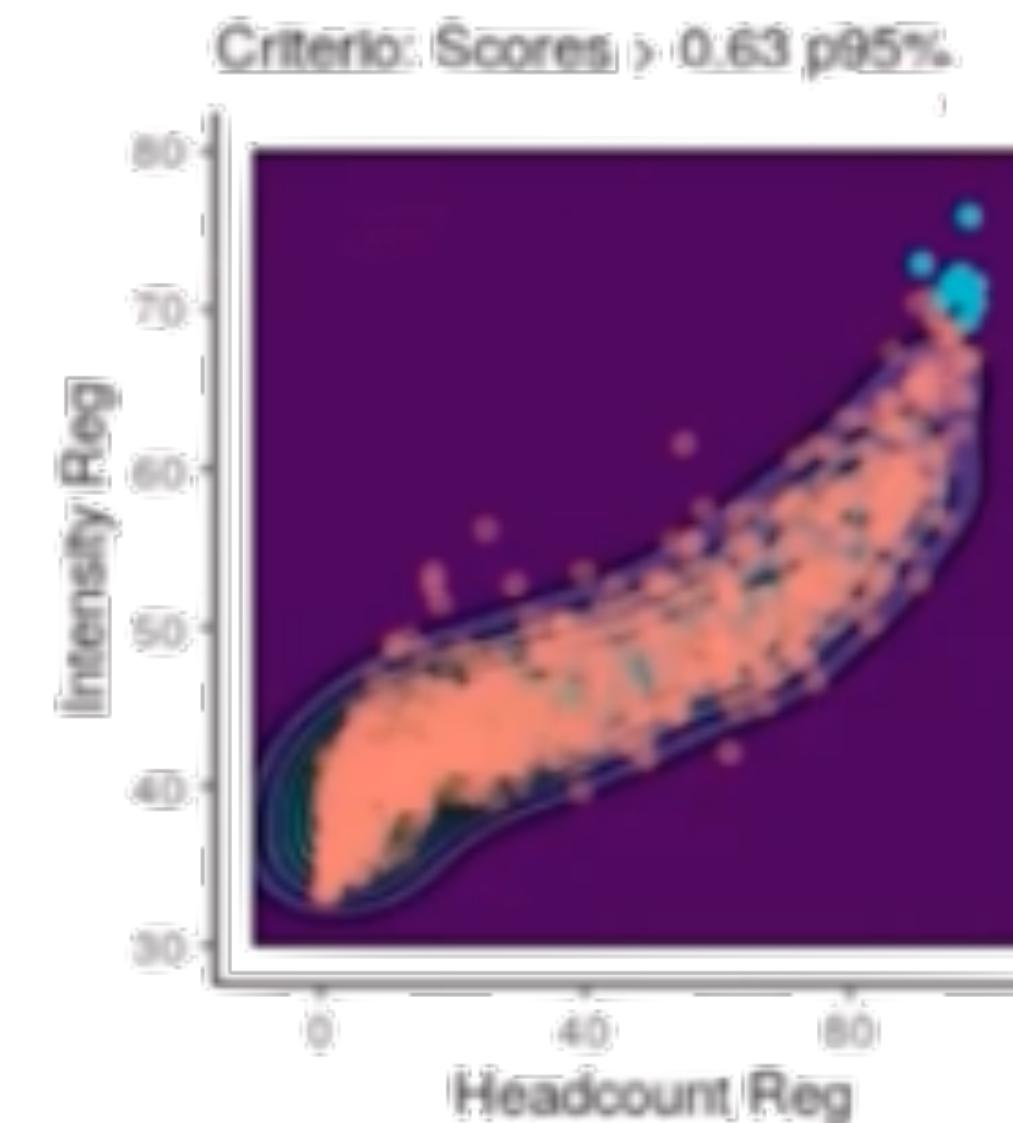
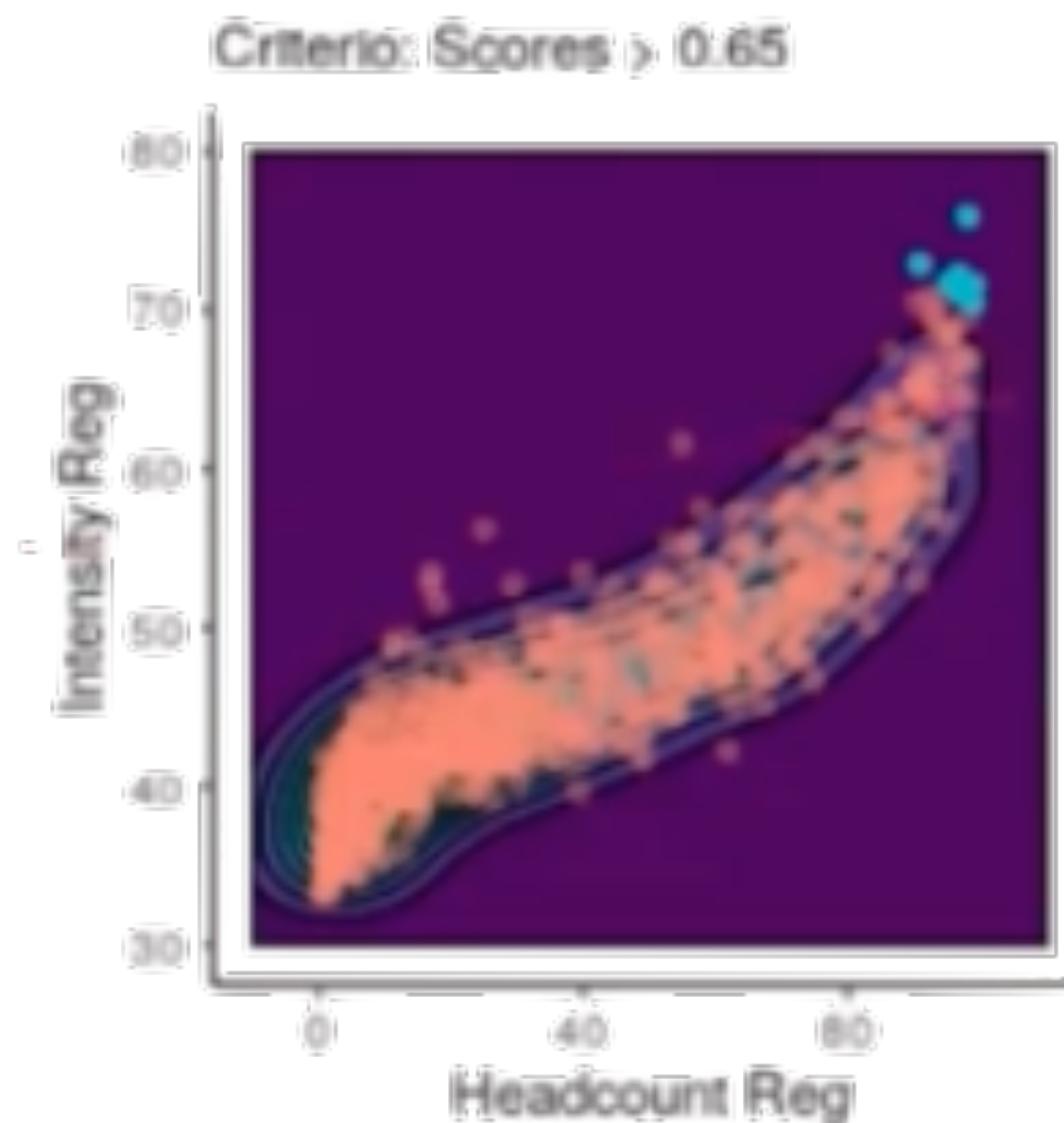
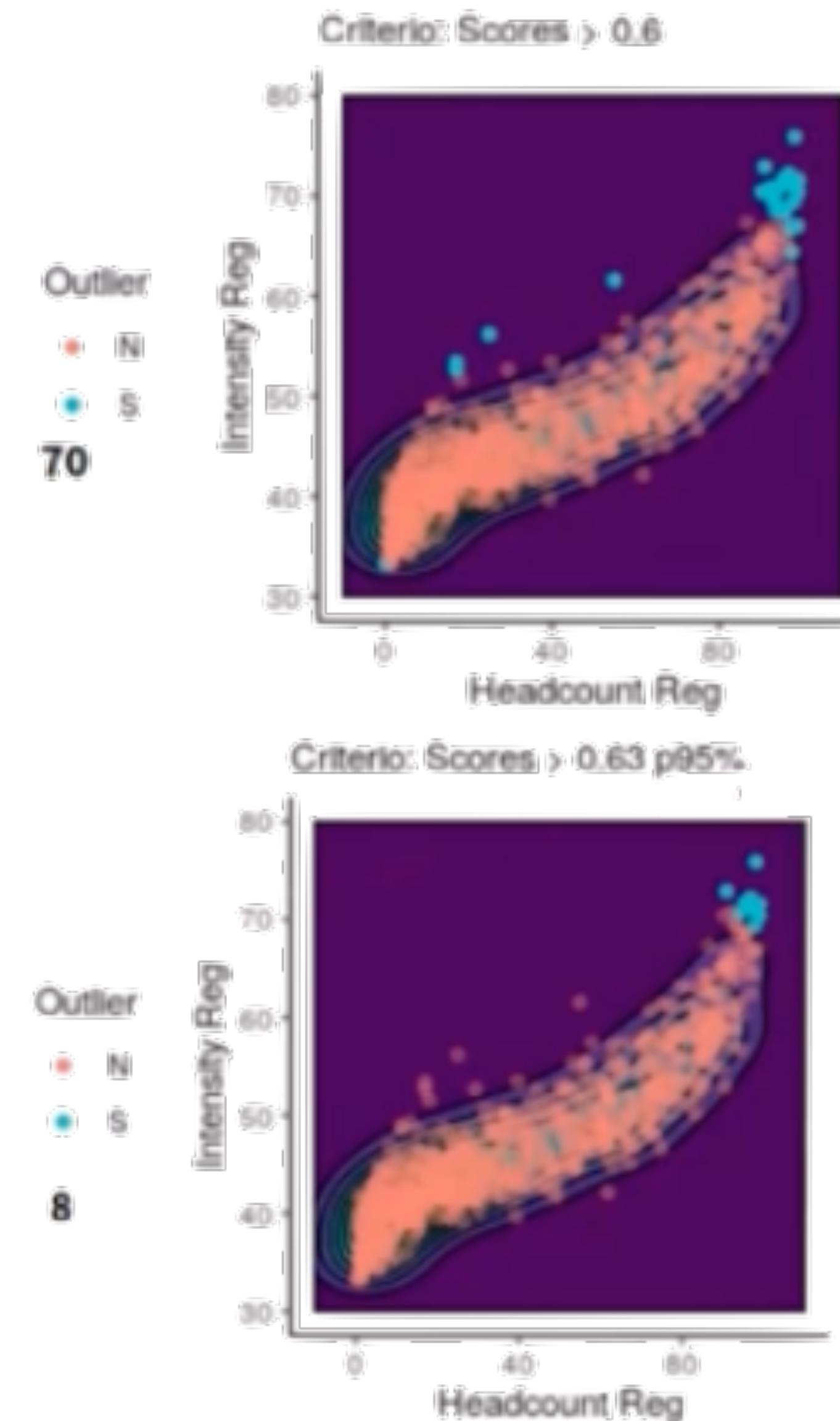
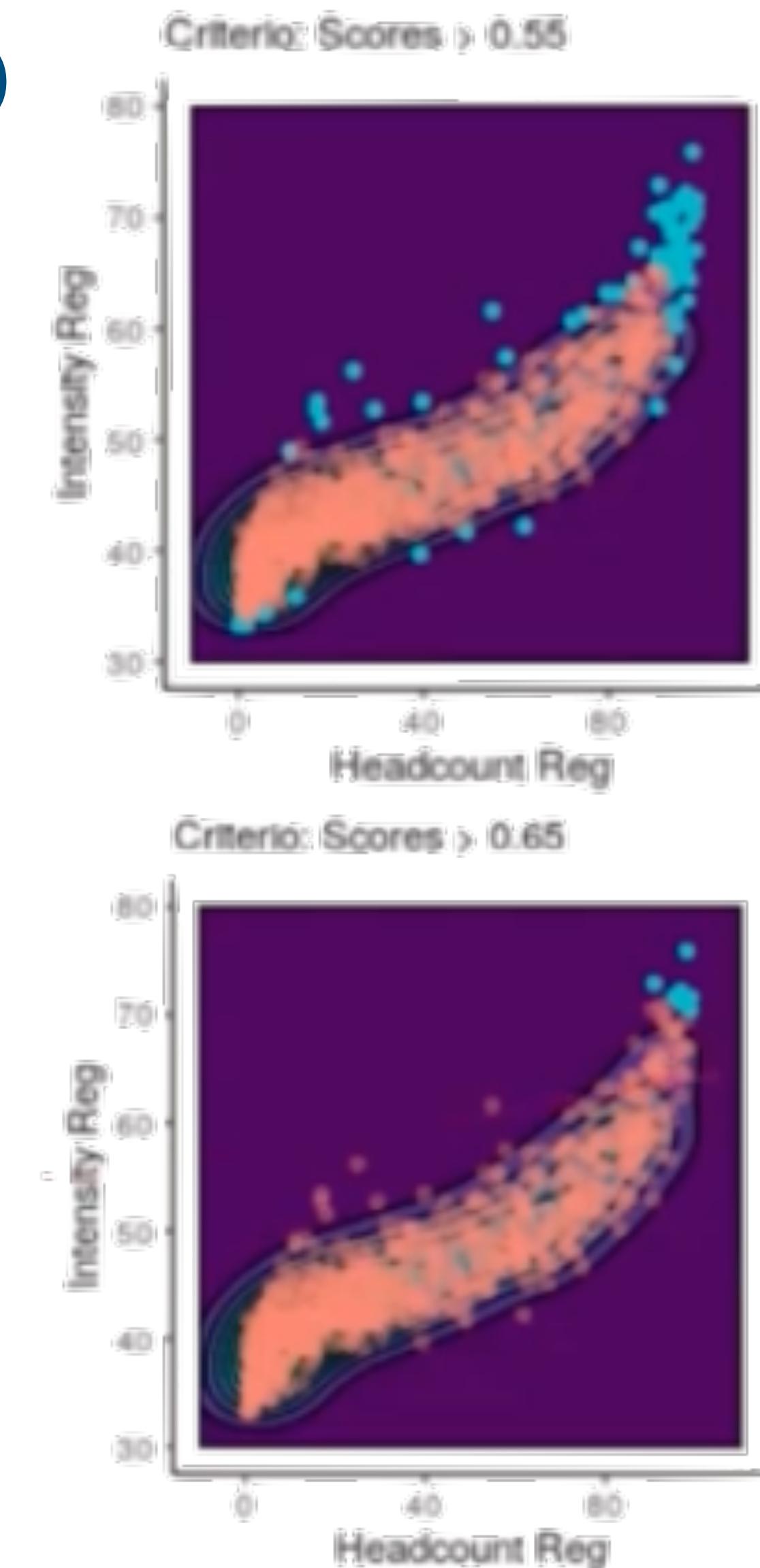
Propiedades de Isolation Forest

- Construye un árbol de manera no supervisada.
- Utiliza la altura del árbol en lugar de medidas de distancia o densidad.
- Puede escalar para manejar datos de gran dimensión.

Ejemplo



Isolation Forest:
Árboles: 500
Muestras: 256 casos
ndim (#features): 1



DM UBA 2021

Datos Faltantes

Datos Faltantes

- Tratamiento de datos faltantes.
 - Origen de los faltantes.
 - Tratamientos:
 - Eliminación de filas o columnas.
 - Imputaciones puntuales (imputación por media, moda, etc.).
 - Métodos cold/hot deck
 - Imputaciones múltiples (MICE).

Tratamiento Datos Faltantes

- El problema de los *missing data* está presente en el análisis de datos desde los **orígenes del almacenamiento**.
- Este problema de missing data se vuelve incluso un dilema propio del **proceso de KDD** donde los volúmenes de datos se incrementan y con ello la probabilidad de inconsistencias y faltantes.

Razones de la inconsistencia

- Factores propios del procedimiento.**

Formularios mal diseñados, errores de programación, etc.

- Negativa a responder.**

Ej: Cuestiones relacionadas con la edad, ¿cuánto gana? afiliación política, religión, etc.

- Respuestas inaplicables.**

Ej: ¿Cuánto gasto en juguetes para sus hijos el último año?...no tengo hijos!

Problemas de trabajar con faltantes

- Los datos faltantes (DF) dificultan el **análisis de datos**.
- Un manejo inapropiado de DF en el análisis puede introducir sesgos y puede resultar en **conclusiones engañosas**.
- También pueden **limitar la generalización** del conocimiento alcanzado.

Tipos de datos faltantes

Existen diferentes **mecanismos de faltantes**, los tipos estándar de DF son:

- Outliers tratados como datos faltantes:** Cuando se conocen los límites de las diferentes variables del dataset, los datos que caen fuera del rango definido se deben etiquetar como faltantes.
- Missing At Random MAR***
 - La probabilidad de que la variable Y tenga un dato faltante depende de X, pero no de Y.
 - Es decir, el patrón de los datos faltantes se puede predecir a partir de otras variables de la base de datos.

Tipos de datos faltantes

Missing completely at random MCAR

- La probabilidad de que la variable Y tenga un dato faltante es independiente de X.
- Los datos existentes en Y son una muestra al azar de los valores de Y

Non-Ignorable missing data. (Missing not at random MNAR)

- el valor de la variable que falta está relacionado con la razón por la que falta

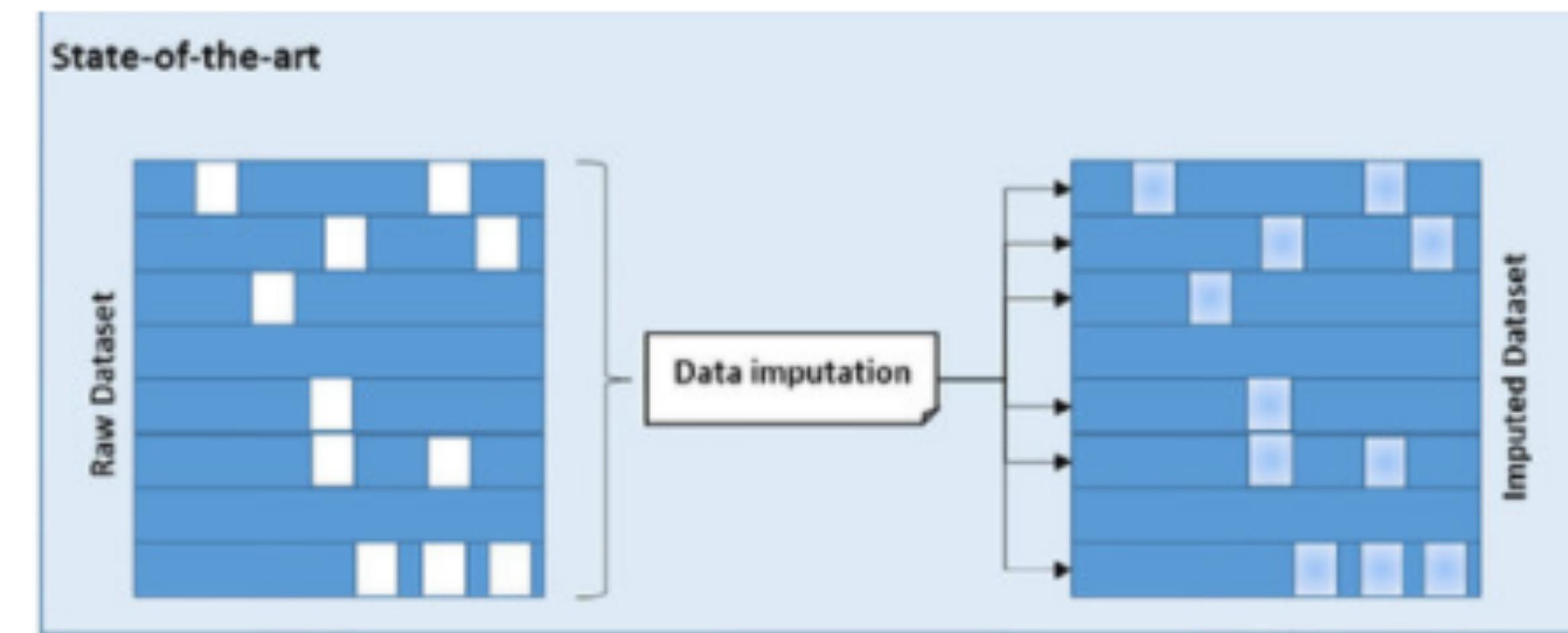
Métodos para tratar datos faltantes

- **Utilizar solo registros completos:** Esto va a depender de cómo es el origen del faltante.
 - Se recomienda que se utilice solo en los casos en que el mecanismo de faltante sea MCAR.
- **Borrar casos seleccionados o variables:** La eliminación de registros ante la presencia de faltantes puede utilizarse cuando hay un **patrón no aleatorio de datos faltantes**.
 - Si la eliminación de un subconjunto disminuye significativamente la utilidad de los datos, la eliminación del caso puede no ser efectiva.
- **Imputación de datos:** Son métodos de relleno de faltantes.
- Aproximaciones basadas en modelado: Múltiples imputaciones

Imputación de Datos

Imputación de datos

- Son métodos de relleno de datos ante la presencia de faltantes.
- Imputar es el proceso de estimar datos faltantes de una observación a partir de valores válidos de otras variables.
- Se debe tener precaución al emplear métodos de imputación, ya que pueden generar sesgos importantes entre los datos reales y los imputados.



Imputación de datos

- **Sustitución de casos.** Se reemplaza con valores no observados. Debería ser realizado por un experto en esos datos.
- **Sustitución de Medias.** Se reemplaza utilizando el promedio calculado de los valores presentes. Debe verificarse que los datos ajusten a una distribución normal, si los datos están sesgados es mejor utilizar la mediana.

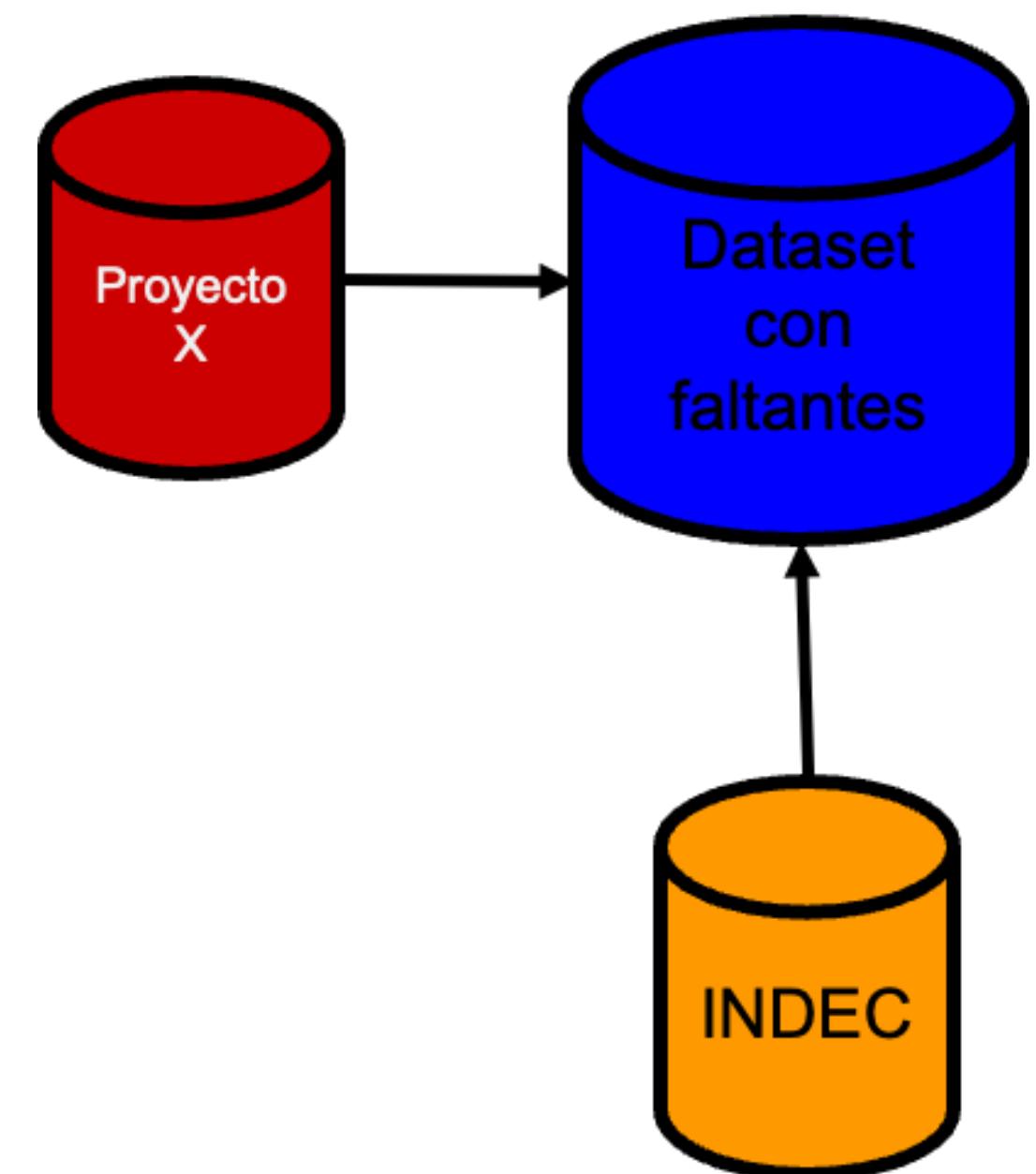
Hay tres desventajas en el uso de la media:

- La varianza estimada de la nueva variable no es válida porque está atenuada por los valores repetidos
- Se distorsiona la distribución
- Las correlaciones que se observen estarán deprimidas debido a la repetición de un solo valor constante.

Imputación de datos: Cold Deck

Cold Deck

- Selecciona valores o usar relaciones obtenidas de fuentes distintas de la base de datos actual.
- Se **sustituye un valor constante** derivado de fuentes externas o de investigaciones previas.
- Tiene las mismas desventajas de aplicar la media.



Imputación de datos: Hot Deck

Hot Deck: Se reemplazan los faltantes con valores obtenidos de registros que son los más similares.

Ventajas de hot deck:

- Conceptualmente simple
- Conserva los niveles de medición adecuados para las variables
- Finaliza el proceso de imputación con un conjunto completo de datos.

Desventajas:

- la dificultad para definir qué es similar.

Table I

Illustration of hot deck imputation: incomplete data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	???
4	2	5	10	2

Table II

Illustration of hot deck imputation: imputed data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	13
4	2	5	10	2

Imputación de datos: Regresión

- Se utiliza el análisis de regresión para predecir valores faltantes a partir variables relacionadas en el conjunto de datos.
- Se pueden utilizar regresiones simples o múltiples.
- Se identifican las variables independientes y dependiente.

```
# Imputación por Regresión
income = read.csv("missing_income.csv", sep = ';', dec = ',')
f_income = lm(income~age + years_of_college, data=income[1:17,])
summary(f_income)

income$imputado = 33912.1 + 300.9*income$age + 1554.2*income$years_of_college
```

	caso	income	age	years_of_college
1	1	45251.25	26	4
2	2	62498.27	45	6
3	3	49350.32	28	5
4	4	46424.92	28	4
5	5	56077.27	46	4
6	6	51776.24	38	4
7	7	51410.97	35	4
8	8	64102.33	50	6
9	9	45953.96	45	3
10	10	50818.87	52	5
11	11	49078.98	30	0
12	12	61657.42	50	6
13	13	54479.90	46	6
14	14	64035.71	48	6
15	15	51651.50	50	6
16	16	46326.93	31	3
17	17	53742.71	50	4
18	18	NA	55	6
19	19	NA	35	4
20	20	NA	39	5

Imputaciones Múltiples MICE

Imputación de datos múltiple: MICE

Es una técnica de imputación y su acrónimo significa: Multivariate imputation by chained equations (MICE).

MICE opera bajo el supuesto de que el origen de los faltantes es Missing At Random (MAR).

La probabilidad de que falte un valor depende solo de los valores observados y no de los valores no observados.

Imputación de datos múltiple: MICE

El proceso de *Chained Equation* se puede dividir en cuatro pasos generales:

Paso 1: Imputación Simple

A partir de un reemplazo por medias o modas u otra imputación se completan todos los faltantes.

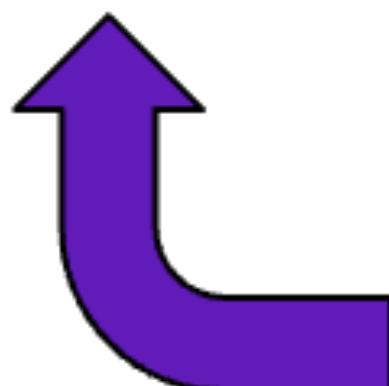
Paso 2: Place Holders

Se toma una de las variables (**X**) con faltantes y se la vuelve a poner NA en sus faltantes originales.

Paso 3: Modelado

Se ajusta un modelo que permita predecir a partir de todas (o algunas de las variables completas) los NA de **X**.

Los **Pasos 2 a 4** son repetidos para cada variable con datos faltantes.



Una vez que se procesan todas las variables con faltantes se cumple un **ciclo o iteración**.

Paso 4: Imputaciones

Los faltantes de **X** son reemplazados por las predicciones. Luego, **X** será usada como predictor para completar otra variable.

Estos ciclos pueden repetirse varias veces, según la bibliografía 10 veces es suficiente.



MICE: ejemplo

Conjunto inicial con datos faltantes

Age	Income	Gender
33	N.A.	F
18	12,000	N.A.
N.A.	13,542	M

Paso 1: Imputación puntual

Age	Income	Gender
33	12,771	F
18	12,000	F
25.5	13,542	M

Paso 2: Pasamos a faltante uno de los casos imputados en 1

Age	Income	Gender
33	12,771	F
18	12,000	F
N.A.	13,542	M

Paso 3: Ajustamos un modelo con los casos completos

Age	Income	Gender
33	12,771	F
18	12,000	F
N.A.	13,542	M

Paso 4: Imputamos con el valor que surja de aplicar el modelo

Age	Income	Gender
33	12,771	F
18	12,000	F
35.3	13,542	M

MICE: ejemplo

Paso 2: Pasamos a faltante uno de los casos imputados en 1

Age	Income	Gender
33	N.A	F
18	12,000	F
35.3	13,542	M

Paso 3: Ajustamos un modelo con los casos completos

Age	Income	Gender
33	N.A	F
18	12,000	F
35.3	13,542	M

Paso 4: Imputamos con el valor que surja de aplicar el modelo

Age	Income	Gender
33	13,103	F
18	12,000	F
35.3	13,542	M

Age	Income	Gender
33	13,103	F
18	12,000	M
35.3	13,542	M

Por último se modela Gender y se completa la primer iteración.

Con los valores imputados en este 1er ciclo y utilizando la máscara de faltantes se vuelve a iterar.

MICE

Ventajas

No produce sesgo (Esto dependerá del modelo de imputación)

Puede ser utilizado para cualquier tipo de análisis.

Es fácil de usar.

Desventajas

Hay que pensar en el modelo de imputación además del modelo de análisis.

Puede ser costoso computacionalmente

Genera un dataset completo por cada iteración



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Gracias