

Análisis de los datos disponibles

□ Tipos de Variables

- Cuantitativas y cualitativas

□ Descripciones estadísticas

- Medidas de tendencia central
- Medidas de dispersión

□ Gráficos

- Diagrama de barras
- Diagrama de torta
- Histograma
- Diagrama de caja
- Diagrama de dispersión

Tipos de variables

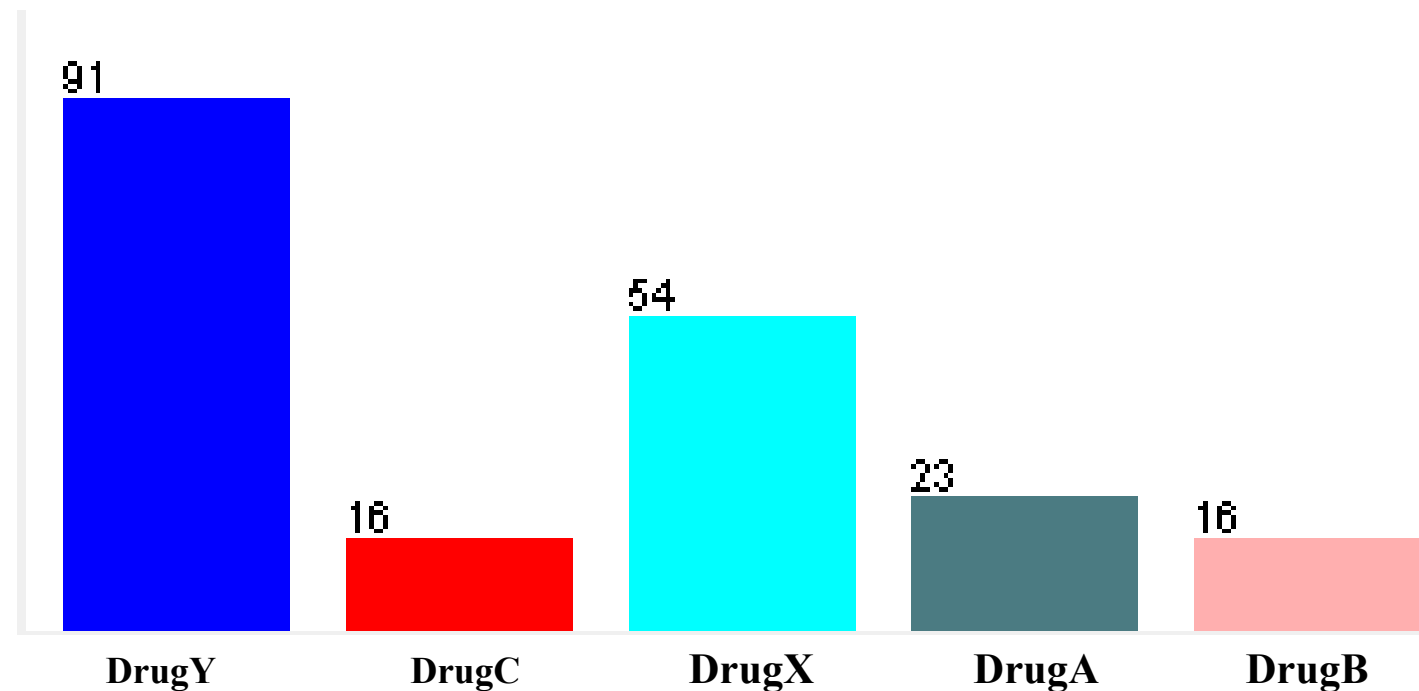
□ **Cuantitativas o numéricas**

- ▣ DISCRETAS (cant. de empleados, cant. de alumnos, etc)
- ▣ CONTINUAS (sueldo, metros cuadrados, beneficios, etc)

□ **Cualitativas o categóricas**

- ▣ NOMINALES: nombran al objeto al que se refieren sin poder establecer un orden (estado civil, raza, idioma, etc.)
- ▣ ORDINALES: se puede establecer un orden entre sus valores (alto, medio, bajo, etc)

- Se busca predecir si el tipo de fármaco que se debe administrar a un paciente afectado de rinitis alérgica es el habitual o no.



- Se dispone de información de pacientes afectados de rinitis alérgica:
 - ▣ Age: Edad
 - ▣ Sex: Sexo
 - ▣ BP (Blood Pressure): Tensión sanguínea.
 - ▣ Cholesterol: nivel de colesterol.
 - ▣ Na: Nivel de sodio en la sangre.
 - ▣ K: Nivel de potasio en la sangre.
 - ▣ Cada paciente ha sido medicado con un único fármaco de entre cinco posibles: DrugA, DrugB, DrugC, DrugX, DrugY.

Ejemplo

DRUG5.CSV

- Drug5.csv contiene 200 muestras de pacientes atendidos previamente

Nro.	Age	Sex	BP	Colesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0,792535	0,031258	drugY
2	47	M	LOW	HIGH	0,739309	0,056468	drugC
3	47	M	LOW	HIGH	0,697269	0,068944	drugC
4	28	F	NORMAL	HIGH	0,563682	0,072289	drugX
5	61	F	LOW	HIGH	0,559294	0,030998	drugY
...
...
...
197	16	M	LOW	HIGH	0,743021	0,061886	drugC
198	52	M	NORMAL	HIGH	0,549945	0,055581	drugX
199	23	M	NORMAL	NORMAL	0,78452	0,055959	drugX
200	40	F	LOW	NORMAL	0,683503	0,060226	drugX

Ejemplo

- Drug5.csv contiene 200 muestras de pacientes atendidos previamente

Nro.	Age	Sex	BP	Colesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0,792535	0,031258	drugY
2	47	M	LOW	HIGH	0,739309	0,056468	drugC
3	47	M	LOW	HIGH	0,697269	0,068944	drugC
4	28	F	NORMAL	HIGH	0,563682	0,072289	drugX
5	61	F	LOW	HIGH	0,559294	0,030998	drugY
...

- ¿Cuántos atributos tiene la tabla?
- ¿De qué tipo es cada uno de ellos?

Análisis de los datos disponibles

□ Tipos de Variables

- ▣ Cuantitativas y cualitativas

□ Descripciones estadísticas

- ▣ Medidas de tendencia central
 - Media, moda, mediana, rango medio
- ▣ Medidas de dispersión
 - Varianza, Desviación estándar, Rango
 - Cuartiles, Rango intercuartil

□ Gráficos

- ▣ Diagrama de barras
- ▣ Diagrama de torta
- ▣ Histograma
- ▣ Diagrama de caja
- ▣ Diagrama de dispersión

Descripciones estadísticas básicas

- Identifican propiedades de los datos y destacan qué valores deben tratarse como ruido o valores atípicos

MEDIDAS DE TENDENCIA CENTRAL

- Media
- Mediana
- Moda
- Rango medio

MEDIDAS DE DISPERSION

- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Intercuartil

MEDIA

- La **MEDIA** es el promedio de los valores del atributo. Dicho atributo debe ser numérico.

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N}$$

N es la cantidad de valores a promediar

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\bar{X} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58$$

MEDIA

- La **MEDIA** es el promedio de los valores del atributo. Dicho atributo debe ser numérico.

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N}$$

N es la cantidad de valores a promediar

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

↑
 $\bar{X} = 58$

MEDIA TRUNCADA
¿cómo se calcula?
¿para qué sirve?

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad impar** de valores

30 36 47 50 52 52 56 57 60 63 70 70 110



$$\tilde{X} = x_{(N+1)/2} = 56$$

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad impar** de valores

30 36 47 50 52 52 56 57 60 63 70 70 110



$$\tilde{X} = 56$$

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad par** de valores

30 36 47 50 52 52 56 60 63 70 70 110



$$\tilde{X} = \frac{x_{N/2} + x_{(N+1)/2}}{2} = \frac{52 + 56}{2} = 54$$

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad par** de valores

30 36 47 50 52 52 56 60 63 70 70 110



$$\tilde{X} = 54$$

MEDIANA

- También puede calcularse sobre **atributos ordinales**. En tal caso, el resultado será o bien el valor que divide al conjunto en dos partes iguales o bien se dirá que “la mediana está entre los valores ...”.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo ordinal con una **cantidad impar** de valores

chico	chico	chico	chico	medio	medio	grande	grande	grande
-------	-------	-------	-------	-------	-------	--------	--------	--------



$$\tilde{X} = \text{medio}$$

MEDIANA

- También puede calcularse sobre **atributos ordinales**. En tal caso, el resultado será o bien el valor que divide al conjunto en dos partes iguales o bien se dirá que “la mediana está entre los valores ...”.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo ordinal con una **cantidad par** de valores

chico	chico	chico	medio	medio	grande	grande	grande
-------	-------	-------	-------	-------	--------	--------	--------



$$\tilde{X} = \text{medio}$$

MEDIANA

- También puede calcularse sobre **atributos ordinales**. En tal caso, el resultado será o bien el valor que divide al conjunto en dos partes iguales o bien se dirá que “la mediana está entre los valores ...”.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo ordinal con una **cantidad par** de valores

chico	chico	chico	chico	medio	grande	grande	grande
-------	-------	-------	-------	-------	--------	--------	--------



\tilde{X} está entre “chico” y “medio”

MODA

- ❑ La moda es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.
- ❑ Es posible que la mayor frecuencia corresponda a varios valores diferentes, lo que da lugar a más de una MODA.
- ❑ Los conjuntos de datos con uno, dos o tres modas se denominan unimodal, bimodal y trimodal, respectivamente.
- ❑ En general, un conjunto de datos con dos o más modas es multimodal.
- ❑ Si cada valor de los datos ocurre sólo una vez, entonces no hay moda.

MODA

- La moda es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.

- Ejemplo: atributo numérico

30	36	47	50	52	52	56	60	63	70	70	110
----	----	----	----	----	----	----	----	----	----	----	-----

- ▣ Hay 2 modas y sus valores son 52 y 70

- Ejemplo: atributo nominal

español	inglés	chino	inglés	chino	chino
---------	--------	-------	--------	-------	-------

- ▣ La moda es “chino” por ser el valor que aparece más veces

RANGO MEDIO

- El rango medio es fácil de calcular y también puede utilizarse para evaluar la tendencia central de un conjunto de datos numéricos.
- Es la media de los valores máximo y mínimo del conjunto.

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\text{rango medio} = \frac{\text{maximo} + \text{minimo}}{2} = \frac{110 + 30}{2} = \frac{140}{2} = 70$$

Medidas descriptivas

Atributo AGE - DRUG5.CSV

MINIMO	15
MEDIA	44.3
MEDIANA	45
MAXIMO	74
RANGO MEDIO	44.5
MODA	47

Atributo AGE - DRUG5_ATIPICOS.CSV

MINIMO	15
MEDIA	45
MEDIANA	45
MAXIMO	174
RANGO MEDIO	94.5
MODA	47

Analisis_Drug5.ipynb

Descripciones estadísticas básicas

- Identifican propiedades de los datos y destacan qué valores deben tratarse como ruido o valores atípicos

MEDIDAS DE TENDENCIA CENTRAL

- Media
- Mediana
- Moda
- Rango medio

MEDIDAS DE DISPERSION

- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Intercuartil

VARIANZA Y DESVIACION ESTANDARD

- La varianza mide la dispersión de los datos con respecto a la media.
- Valores bajos indican que las observaciones de los datos tienden a estar muy cerca de la media, mientras que valores altos indican que los datos están muy dispersos.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2$$

- La desviación estándar σ es la raíz cuadrada de la varianza

VARIANZA Y DESVIACION ESTANDARD

□ Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

VARIANZA POBLACIONAL

$$\sigma^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2 = \frac{1}{12} (30^2 + 36^2 + \dots + 110^2) - 58^2 \approx 379.17$$

$$\sigma \approx \sqrt{379.17} \approx 19.47$$

VARIANZA Y DESVIACION MUESTRAL

□ Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

VARIANZA MUESTRAL

$$S^2 = \left(\frac{1}{N-1} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2 = \frac{1}{11} (30^2 + 36^2 + \dots + 110^2) - 58^2 \approx 413.64$$

$$S \approx \sqrt{413.64} \approx 20.34$$

RANGO

- El rango de un conjunto de valores numéricos es la diferencia entre los valores máximo y mínimo de dicho conjunto.

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\text{rango} = \text{maximo} - \text{minimo} = 110 - 30 = 80$$

Cuantiles, Cuartiles y Percentiles

- Los cuantiles son valores que dividen un conjunto numérico ordenado en partes iguales. Es decir que determinan intervalos que comprenden el mismo número de valores.
- Los cuantiles más usados son los siguientes:
 - ▣ CUARTILES: dividen la distribución en cuatro partes.
 - ▣ DECILES: dividen la distribución en diez partes.
 - ▣ Centiles o PERCENTILES: dividen la distribución en cien partes.
 - *El percentil es una medida de posición usada en estadística que indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo.*

CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110



$$Q_1 = 49.25$$



$$Q_2 = 54$$



$$Q_3 = 64.75$$

CUARTILES

- Los cuartiles suelen representarse como Q1, Q2 y Q3.
- El 2do. cuartil o Q2 coincide con la MEDIANA.
- Para hallar las posiciones de Q1 y Q3 usaremos $(N+1)/4$ y $3(N+1)/4$ respectivamente, siendo N la cantidad de valores disponibles.
 - ▣ Si no hay parte decimal, se toma directamente el elemento.
 - ▣ Si la posición corresponde a un número con parte decimal entre el elemento i y el $i+1$, se determina un factor realizando una **interpolación lineal**.

El cuartil será:

$$Q = x_i + (x_{i+1} - x_i) * factor$$

CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110

- La ubicación de Q_1 es $(N+1)/4$, es decir, $(12+1)/4=13/4=3.25$
- Como no es un número entero calculamos su valor entre el 3ro y el 4to elemento.

$$Q_1 = x_3 + (x_4 - x_3) * \underbrace{factor}_{\uparrow}$$

CUARTILES – cálculo del factor

i	F_i
1	0.00
2	0.09
3	0.18
4	0.27
5	0.36
6	0.45
7	0.55
8	0.64
9	0.73
10	0.82
11	0.91
12	1.00

$$N = 12$$

$$F_i = \frac{i - 1}{N - 1}$$

CUARTILES – cálculo del factor

Ubicación de Q1

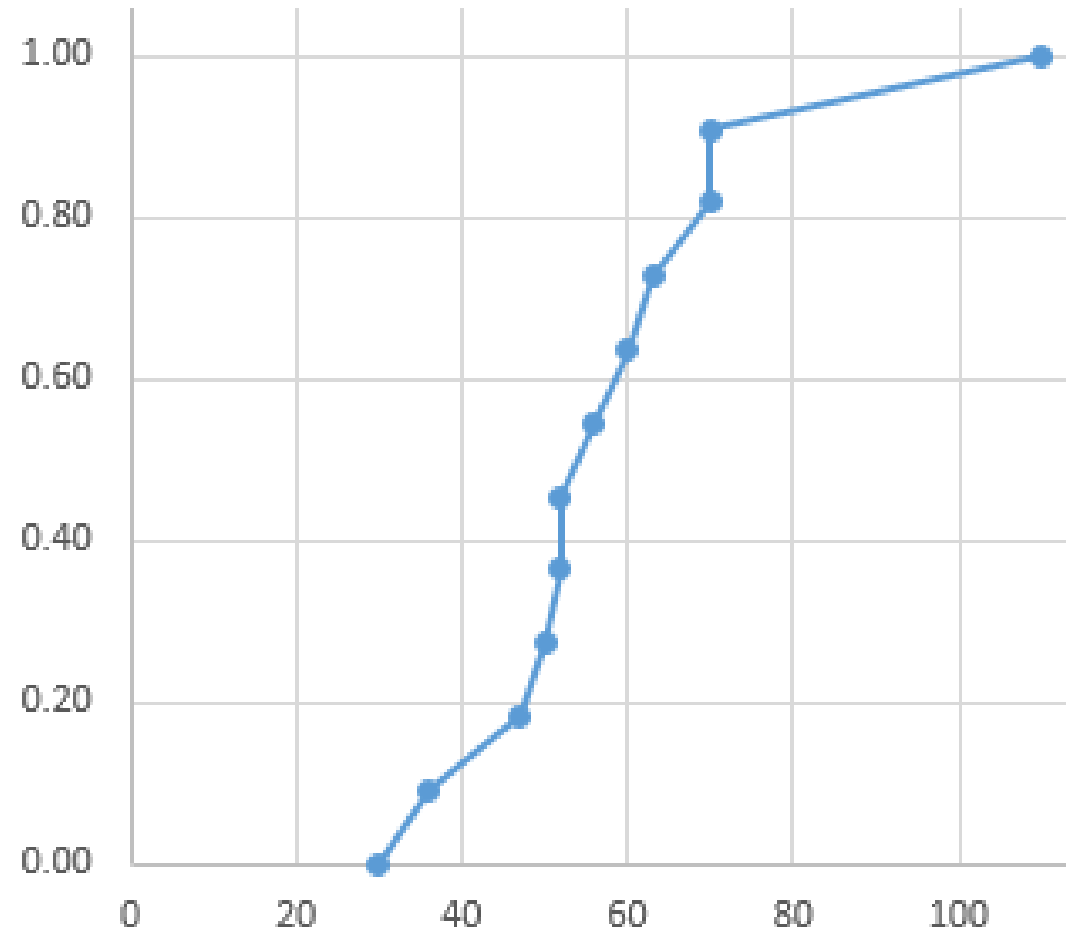
$$(N + 1)/4 = 13/4 = 3.25$$

Q1 →

X	F_i
30	0.00
36	0.09
47	0.18
50	0.27
52	0.36
52	0.45
56	0.55
60	0.64
63	0.73
70	0.82
70	0.91
110	1.00

$$N = 12$$

$$F_i = \frac{i - 1}{N - 1}$$



CUARTILES

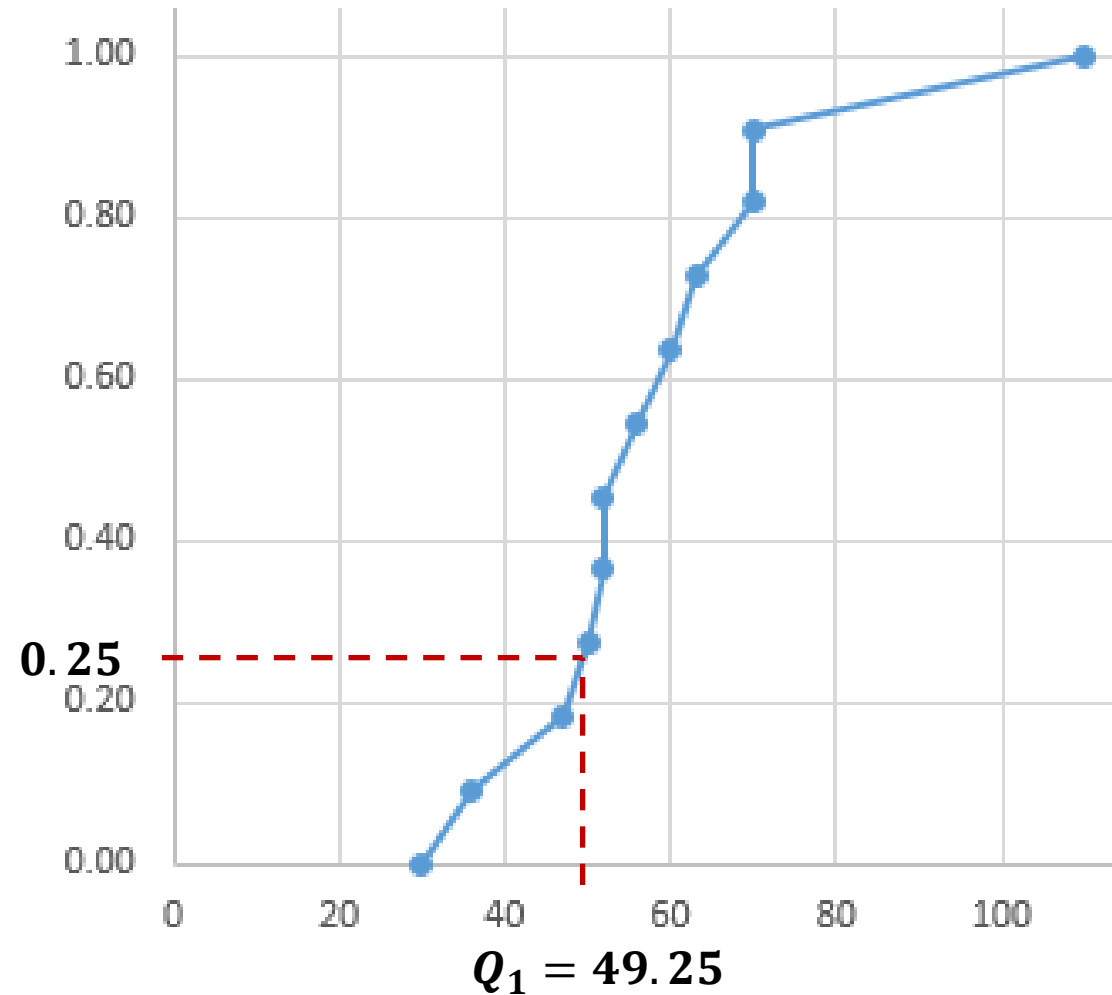
$$F_i = \frac{i - 1}{N - 1}$$

Q1 →

X	F_i
30	0.00
36	0.09
47	0.18
50	0.27
52	0.36
52	0.45
56	0.55
60	0.64
63	0.73
70	0.82
70	0.91
110	1.00

Ubicación de Q1

$$(N + 1)/4 = 3.25$$



CUARTILES

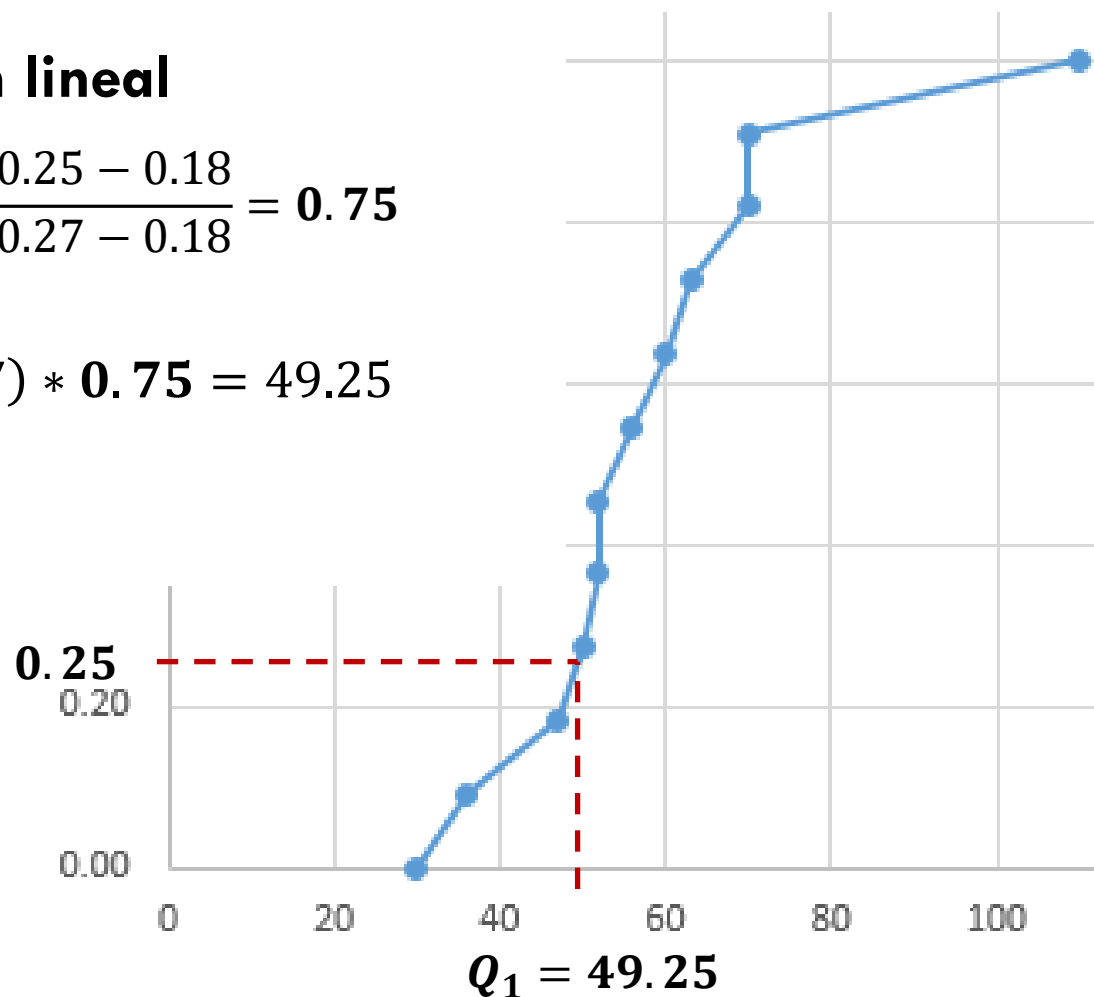
$$F_i = \frac{i - 1}{N - 1}$$

	<i>X</i>	<i>F_i</i>
	30	0.00
	36	0.09
Q1 →	47	0.18
	50	0.27
	52	0.36
	52	0.45
	56	0.55
	60	0.64
	63	0.73
	70	0.82
	70	0.91
	110	1.00

Interpolación lineal

$$factor = \frac{0.25 - F_3}{F_4 - F_3} = \frac{0.25 - 0.18}{0.27 - 0.18} = \mathbf{0.75}$$

$$Q_1 = 47 + (50 - 47) * \mathbf{0.75} = 49.25$$



CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110

- La ubicación de Q3 es $3(N+1)/4 = 3*(12+1)/4 = 3*13/4 = 9.75$
- Como no es un número entero calculamos su valor entre el 9no y el 10mo elemento.

$$\begin{aligned} Q_3 &= x_9 + (x_{10} - x_9) * factor \\ &= 63 + (70 - 63) * 0.25 = 64.75 \end{aligned}$$

CUARTILES

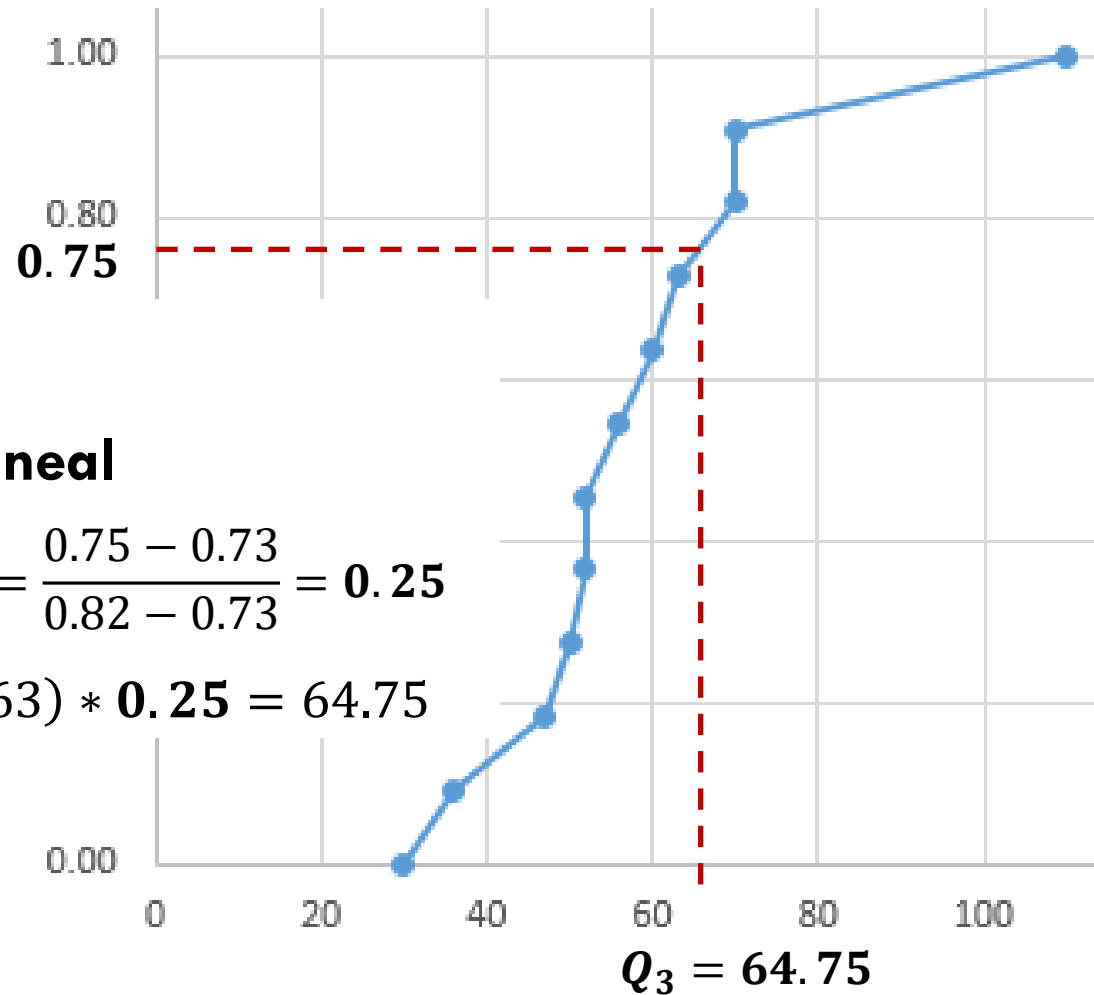
$$F_i = \frac{i - 1}{N - 1}$$

X	F_i
30	0.00
36	0.09
47	0.18
50	0.27
52	0.36
52	0.45
56	0.55
60	0.64
63	0.73
70	0.82
70	0.91
110	1.00

Interpolación lineal

$$factor = \frac{0.75 - F_9}{F_{10} - F_9} = \frac{0.75 - 0.73}{0.82 - 0.73} = 0.25$$

$$Q_3 = 63 + (70 - 63) * 0.25 = 64.75$$



CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110



$$Q_1 = 49.25$$



$$Q_2 = 54$$



$$Q_3 = 64.75$$


RANGO INTERCUARTIL


- La distancia entre Q_1 y Q_3 es una medida sencilla de dispersión que da el rango cubierto por la mitad de los datos.
- Esta distancia se denomina **rango intercuartil (RIC)** y se define como


$$\text{RIC} = Q_3 - Q_1$$

- Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110


 $Q_1 = 49.25$


 $Q_2 = 54$


 $Q_3 = 64.75$

$$\text{RIC} = Q_3 - Q_1 = 64.75 - 49.25 = 15.50$$

Análisis de los datos disponibles

□ Tipos de Variables

- ▣ Cuantitativas y cualitativas

□ Descripciones estadísticas

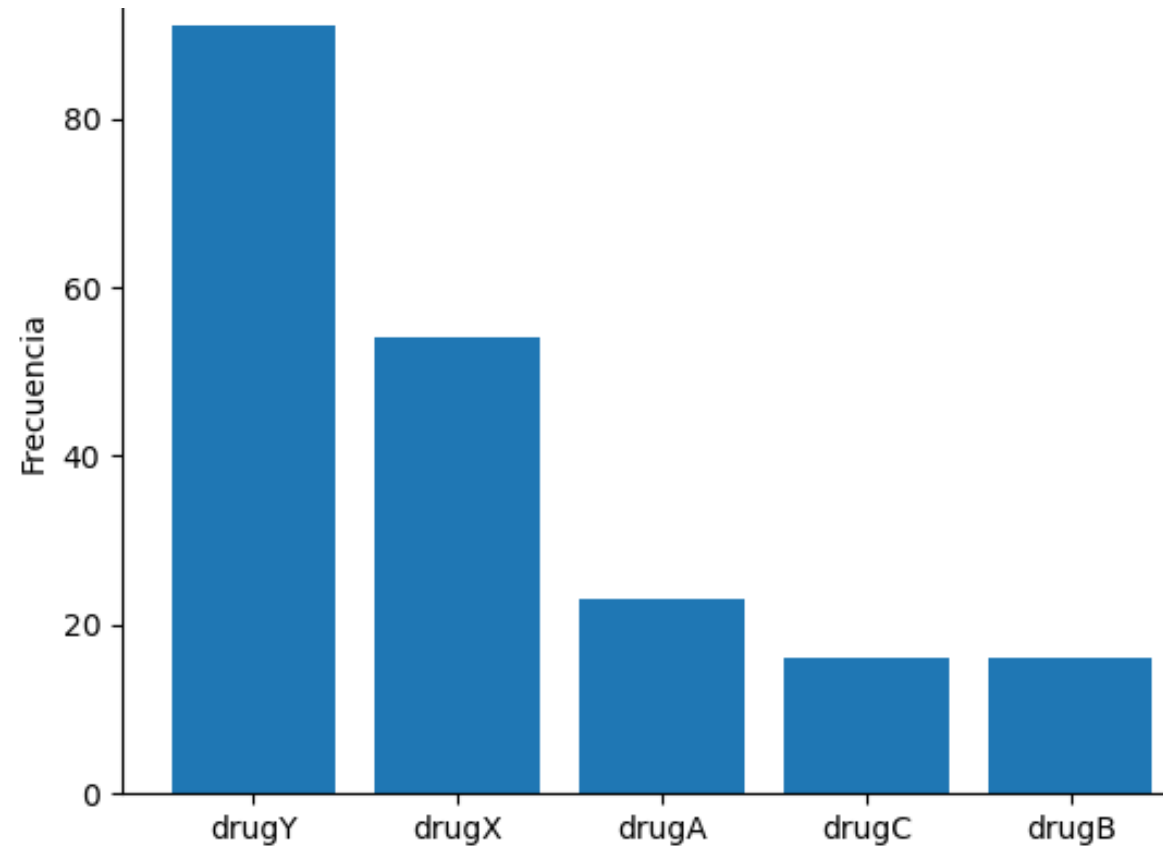
- ▣ Medidas de tendencia central
 - Media, moda, mediana, rango medio
- ▣ Medidas de dispersión
 - Varianza, Desviación estándar, Rango
 - Cuartiles, Rango intercuartil

□ Gráficos

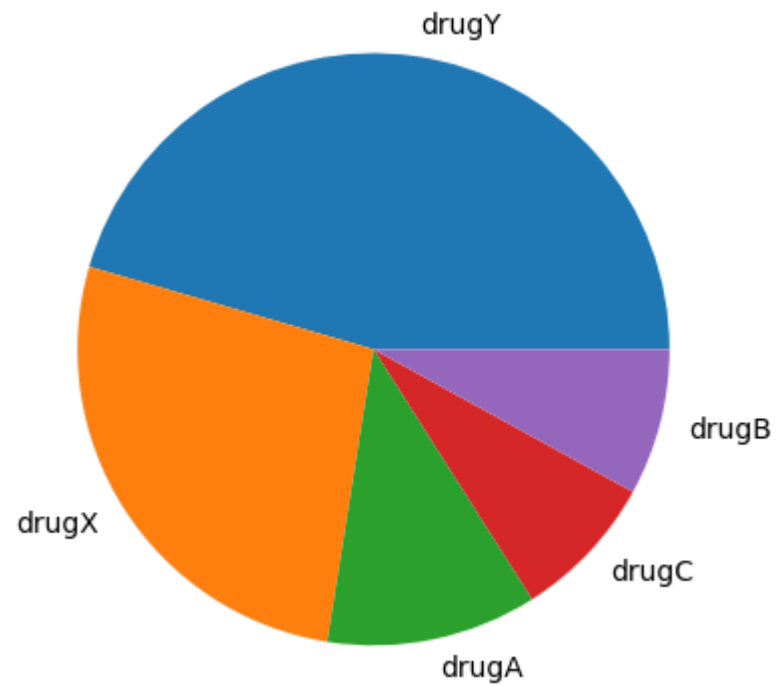
- ▣ Diagrama de barras
- ▣ Diagrama de torta
- ▣ Histograma
- ▣ Diagrama de caja
- ▣ Diagrama de dispersión

Analisis_Drug5.ipynb

Atributo Drug - Diagrama de barras

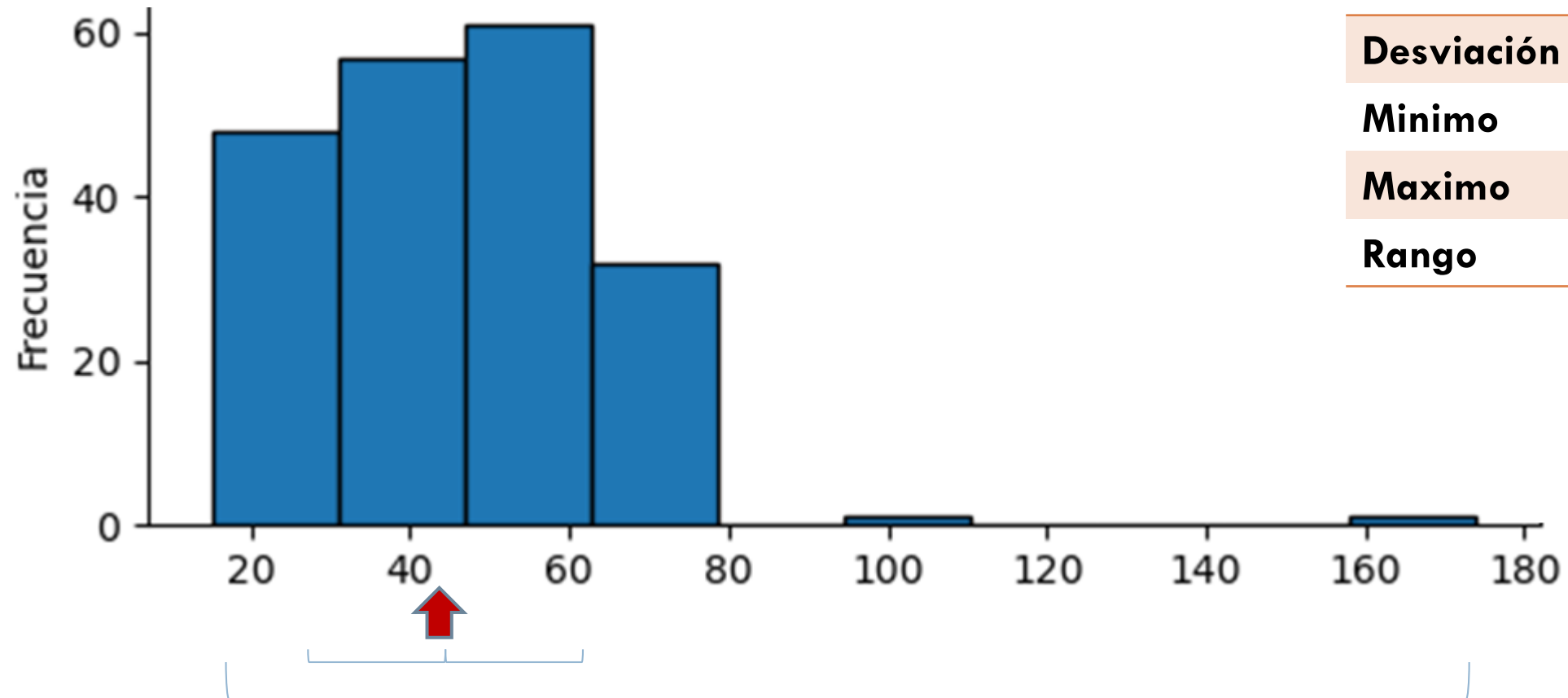


Atributo Drug - Gráfico de Torta



Atributo AGE – Histograma

(Atributo AGE del archivo Drug5_atipicos.CSV)



Media	44.965
--------------	---------------

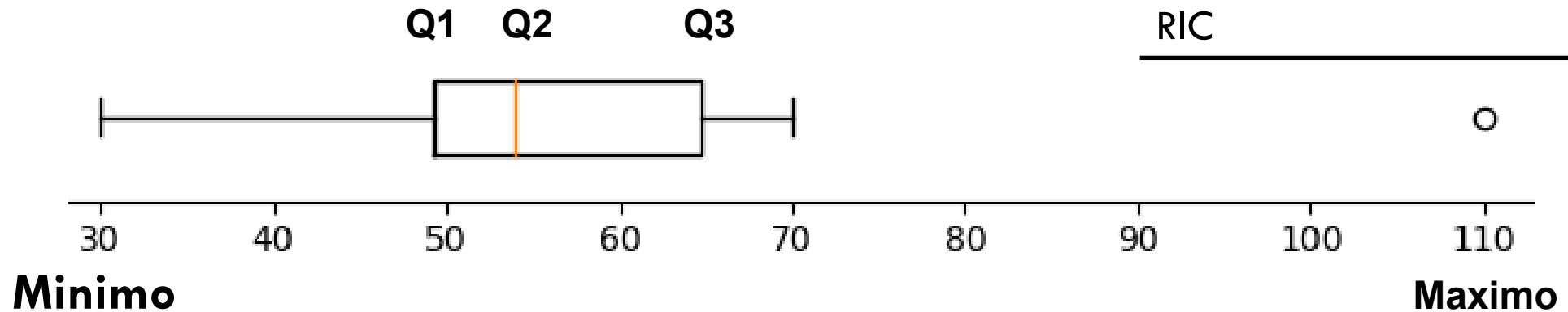
Desviación	19.145
-------------------	---------------

Minimo	15
---------------	-----------

Maximo	174
---------------	------------

Rango	159
--------------	------------

Diagrama de caja - Ejemplo



Mínimo	30
Q1	49.25
Q2 (mediana)	54
Q3	64.75
Maximo	100
RIC	

- Se consideran **valores atípicos leves** a los que se encuentran a $1.5 \times \text{RIC}$ más allá de los límites de la caja y **atípicos extremos** a los que están más allá de $3 \times \text{RIC}$.

Determine si hay valores atípicos y si son leves o extremos

Cuartiles y RIC del atributo AGE

(Atributo AGE del archivo Drug5_atipicos.CSV)

- Luego de ordenar los valores del atributo AGE deben identificarse los valores que los dividen en cuatro partes iguales.

$$Q_1 = 31$$



$$Q_2 = 45$$



$$Q_3 = 58$$



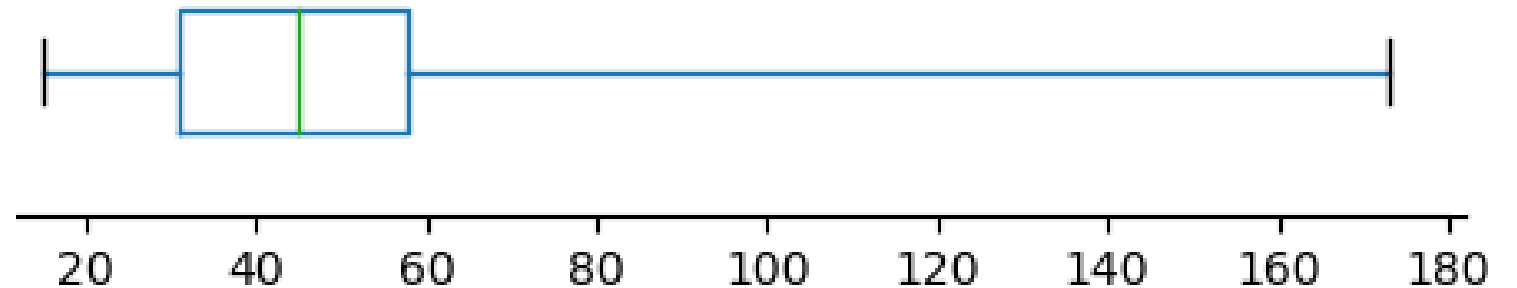
15	...	31	31	...	43	45	45	45	...	58	58	...	174
1	...	50	51	...	99	100	101	102	...	150	151	...	200

$$RIC = Q_3 - Q_1 = 58 - 31 = 27$$

Diagrama de caja (en construcción)

□ Atributo AGE (archivo Drug5_atipicos.csv)

Minimo	15
Q1	31
Q2	45
Q3	58
Maximo	174



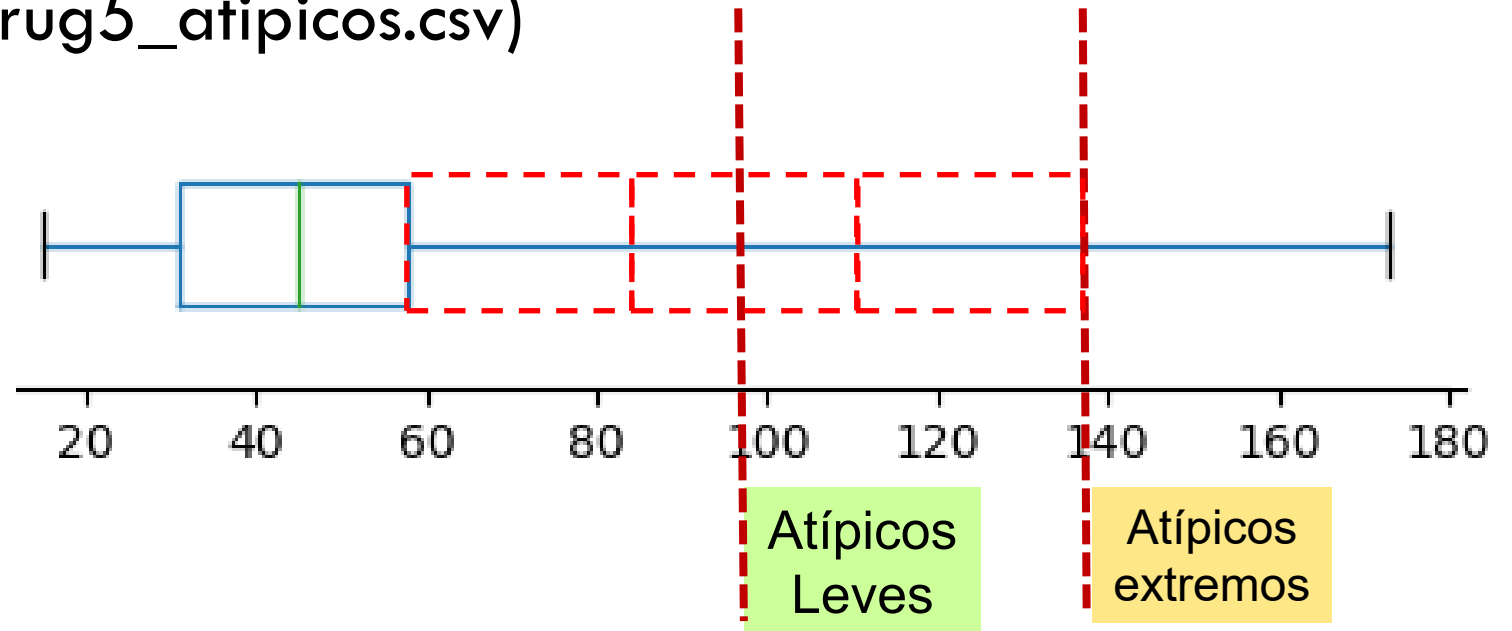
RIC	$Q3 - Q1 = 58 - 31 = 27$
Lim.Inf	$Q1 - 1.5 * RIC = 31 - 1.5 * 27 = -9.5$
Lim.Sup	$Q3 + 1.5 * RIC = 58 + 1.5 * 27 = 98.5$

Hay valores fuera de rango?

Diagrama de caja (en construcción)

□ Atributo AGE (archivo Drug5_atipicos.csv)

Minimo	15
Q1	31
Q2	45
Q3	58
Maximo	174



RIC	$Q3 - Q1 = 58 - 31 = 27$
Lim.Inf	$Q1 - 1.5 * RIC = 31 - 1.5 * 27 = -9.5$
Lim.Sup	$Q3 + 1.5 * RIC = 58 + 1.5 * 27 = 98.5$

Valor atípico o fuera de rango

- Los valores de la muestra que pertenezcan a alguno de estos intervalos

$$[Q1 - 3*RIC ; Q1 - 1.5*RIC) \text{ o } (Q3 + 1.5*RIC ; Q3 + 3*RIC]$$

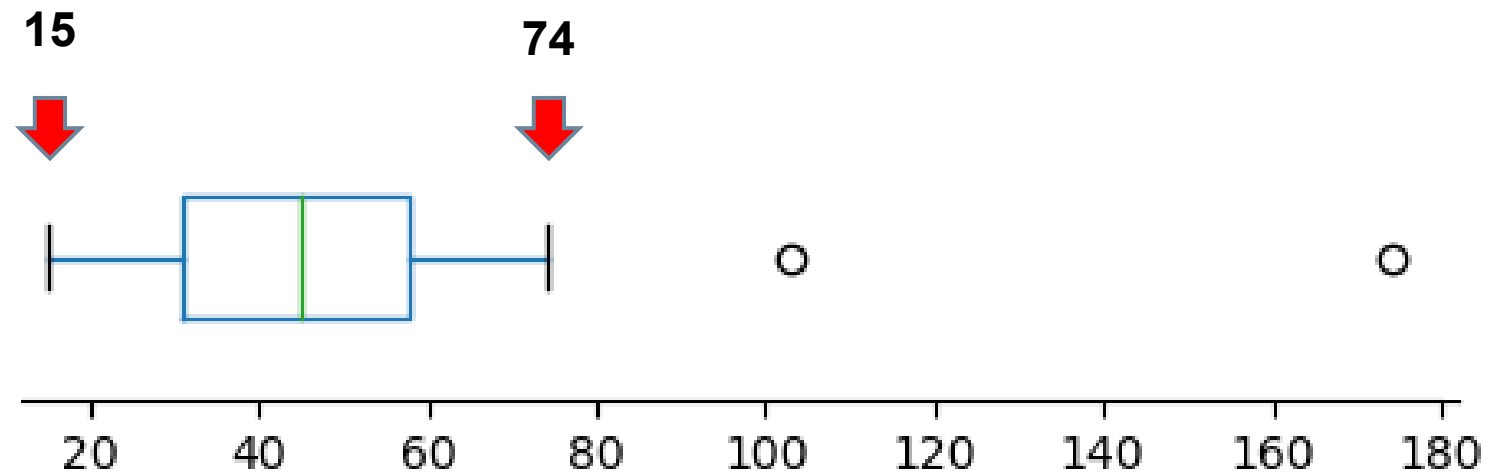
serán considerados **valores fuera de rango leves**.

- Los valores de la muestra inferiores a $Q1 - 3*RIC$ o superiores a $Q3 + 3*RIC$ serán considerados **valores fuera de rango extremos**.

Diagrama de caja

□ Atributo AGE

Minimo	15
Q1	31
Q2	45
Q3	58
Maximo	174



RIC	$Q3 - Q1 = 27$
Lim.Inf	$Q1 - 1.5 * RIC = -9.5$
Lim.Sup	$Q3 + 1.5 * RIC = 98.5$

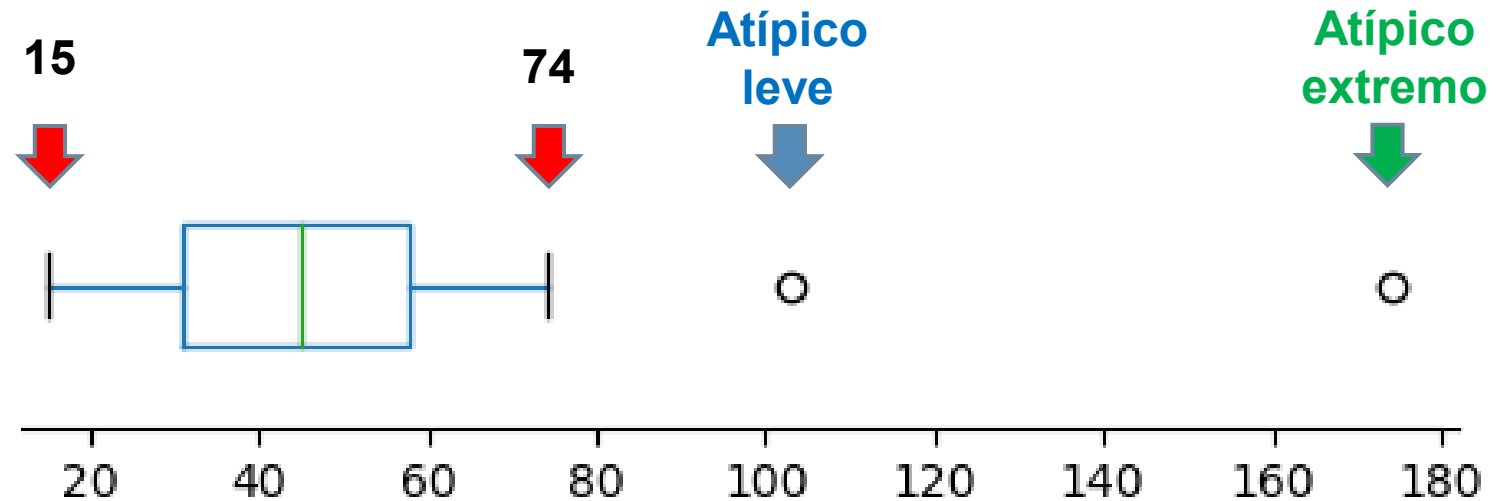
Los bigotes indican el rango de los valores de la muestra comprendidos en el intervalo

$$[Q1 - 1.5 * RIC ; Q3 + 1.5 * RIC] = [-9.5, 98.5]$$

Diagrama de caja

□ Atributo AGE

Minimo	15
Bigote Inferior	15
Q1	31
Q2	45
Q3	58
Bigote Superior	74
Maximo	174



- Los valores de AGE que pertenezcan a $[-50; -9.5)$ o $(98.5; 139]$ se considerarán **atípicos leves**.
- Los valores del atributo AGE inferiores a -50 o superiores a 139 se considerarán **atípicos extremos**.

Histograma y diagrama de caja

(Atributo AGE archivo Drug5_atipicos.CSV)

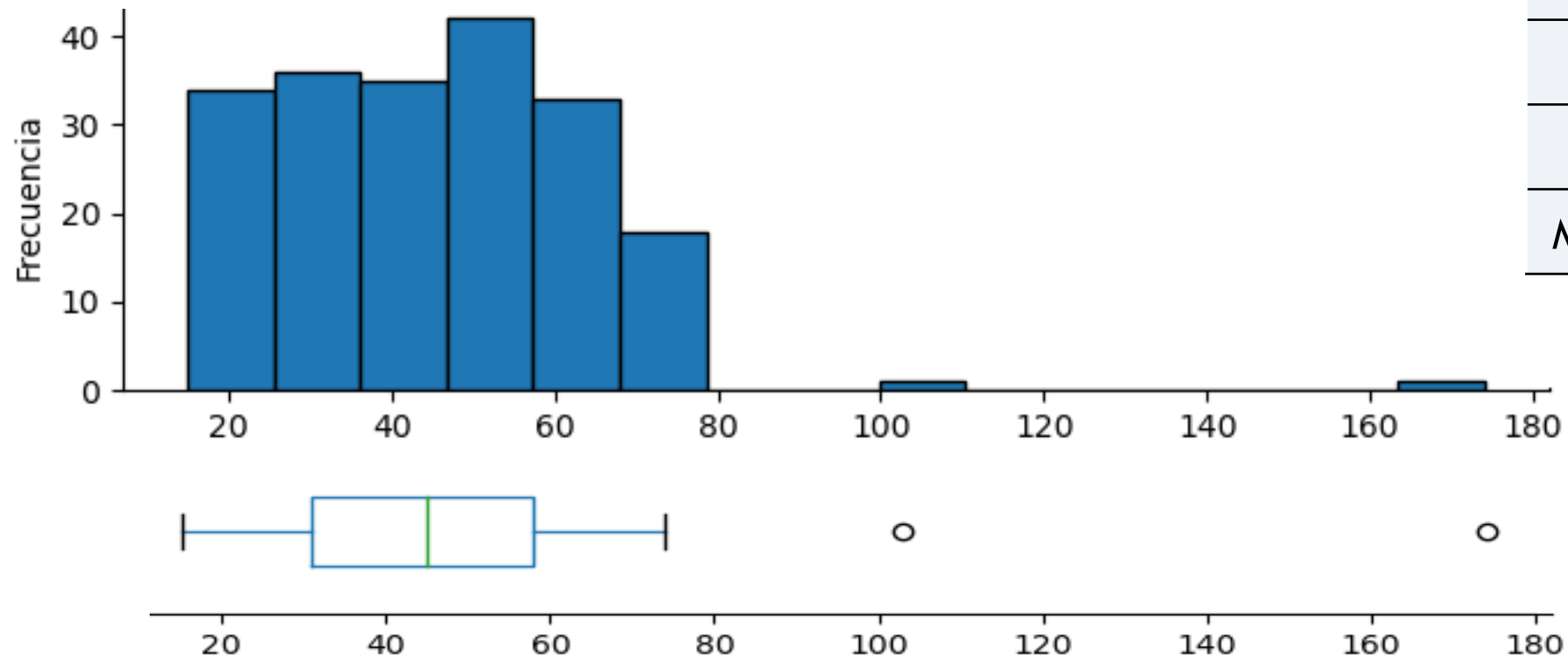
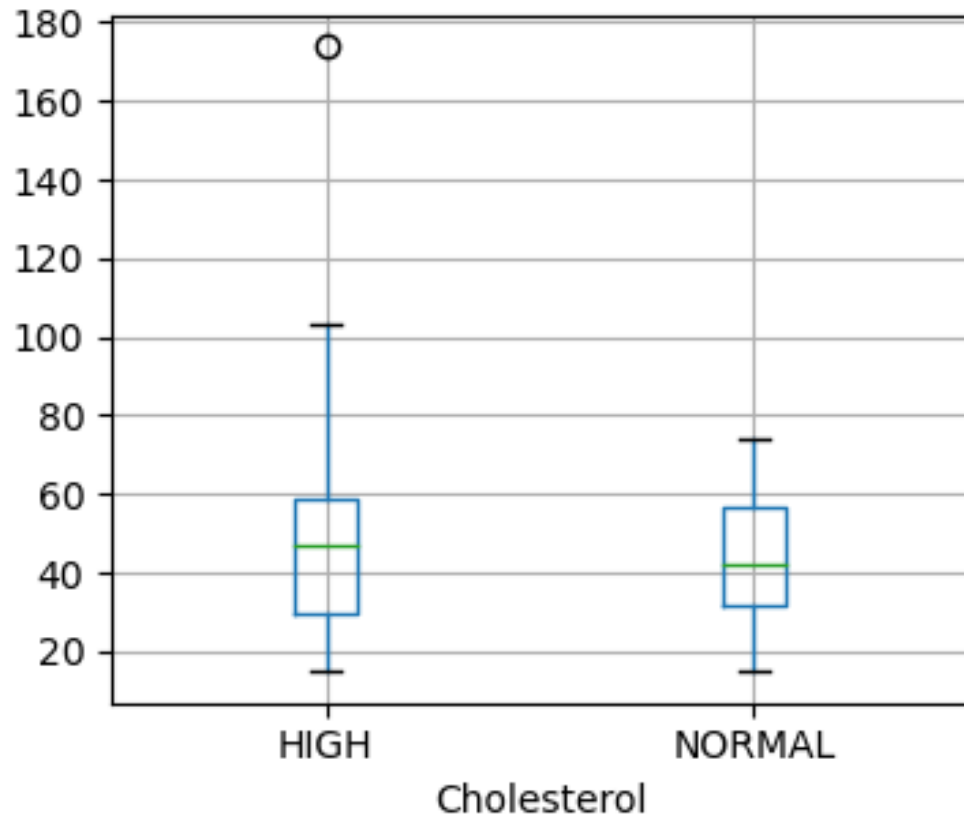


Diagrama de caja usando BY

```
df = pd.read_csv('Drug5_atipicos.csv')  
df.boxplot(column=['Age'], by='Cholesterol')
```



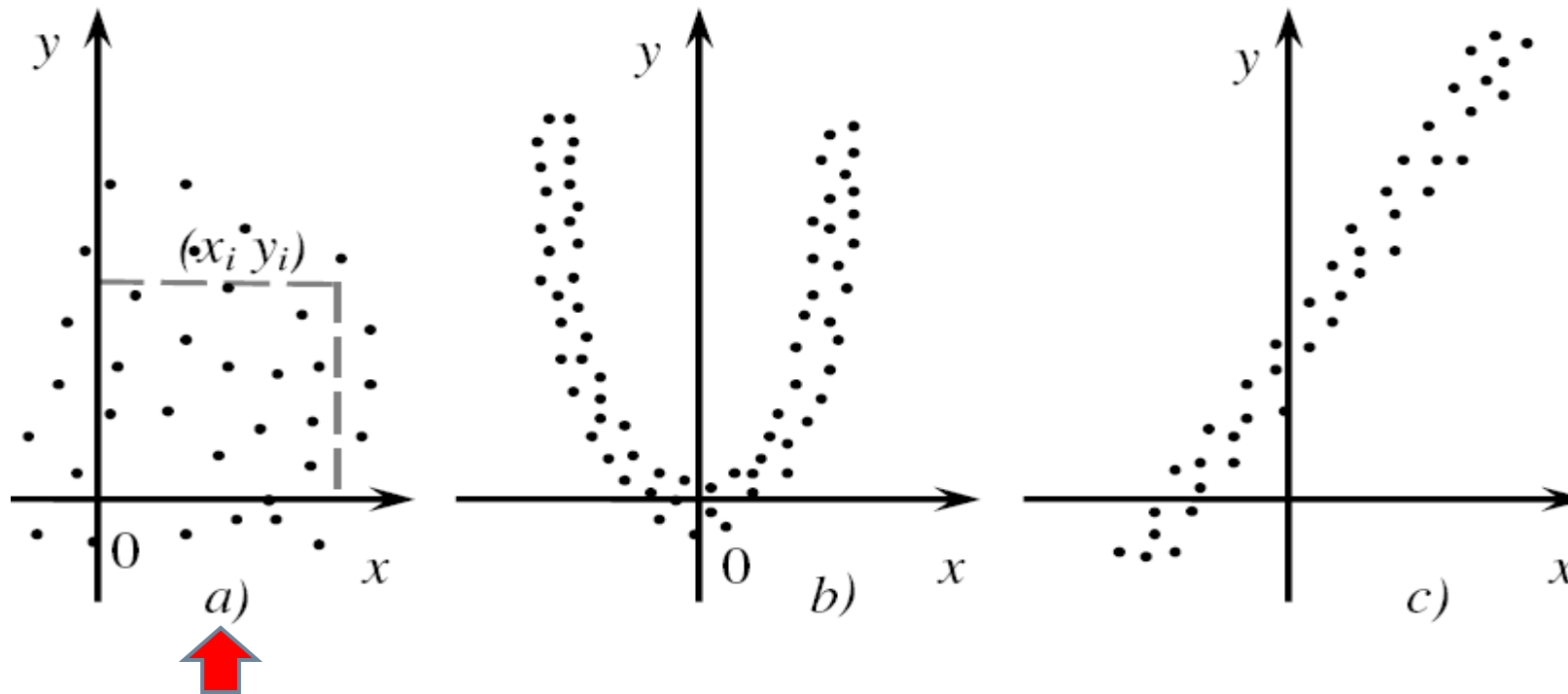
CUARTILES - Edades c/Colesterol NORMAL
[32. 42. 57.]

CUARTILES - Edades c/Colesterol HIGH
[29.5 47. 59.]

Análisis_Drug5.ipynb

Diagrama de Dispersión

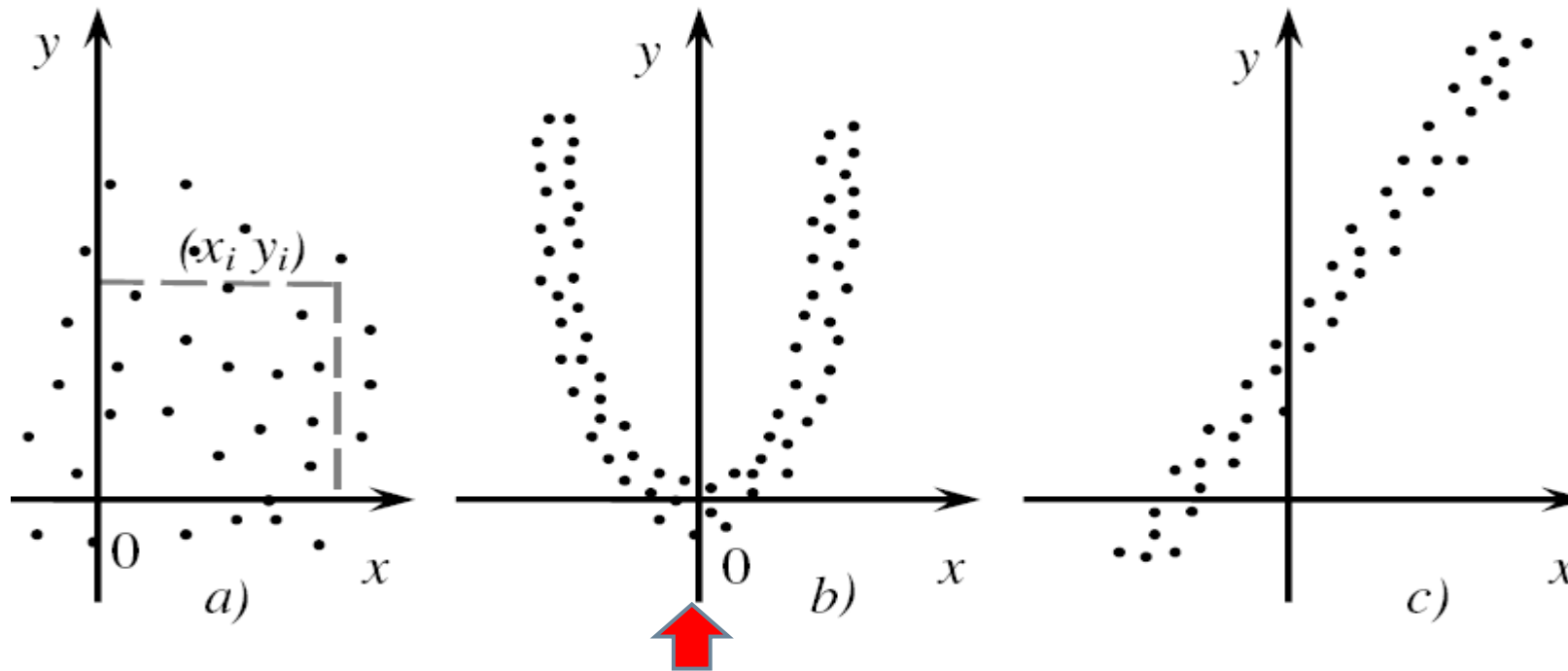
- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y no hay ninguna relación funcional

Diagrama de Dispersión

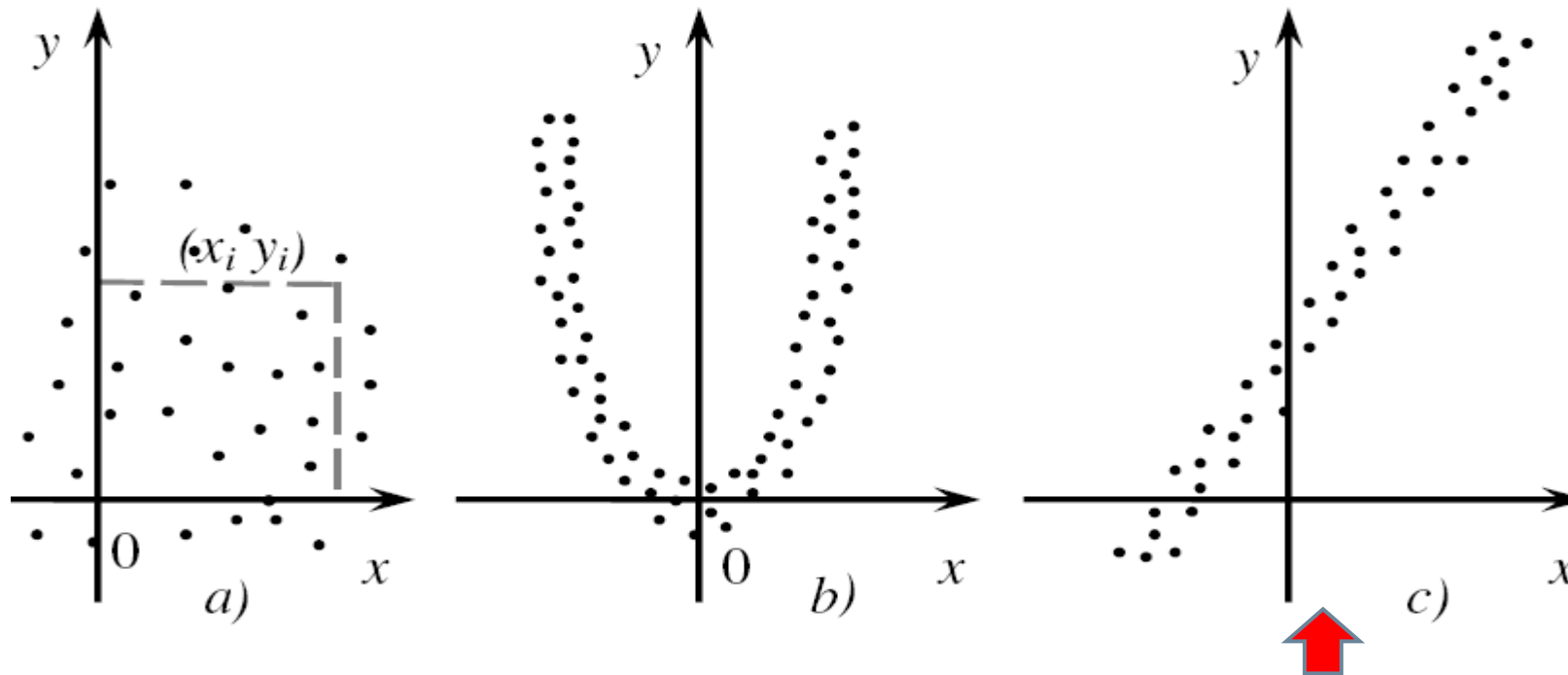
- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X, Y) en un sistema de coordenadas



Entre X e Y podría existir un relación funcional que corresponde a una parábola

Diagrama de Dispersión

- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y existe una **relación lineal**. Este es el tipo de relación que nos interesa

Relación entre atributos numéricos

- Al momento de construir un modelo resulta de interés saber si dos atributos numéricos se encuentran linealmente relacionados o no. Para ello se usa el **coeficiente de correlación lineal**.

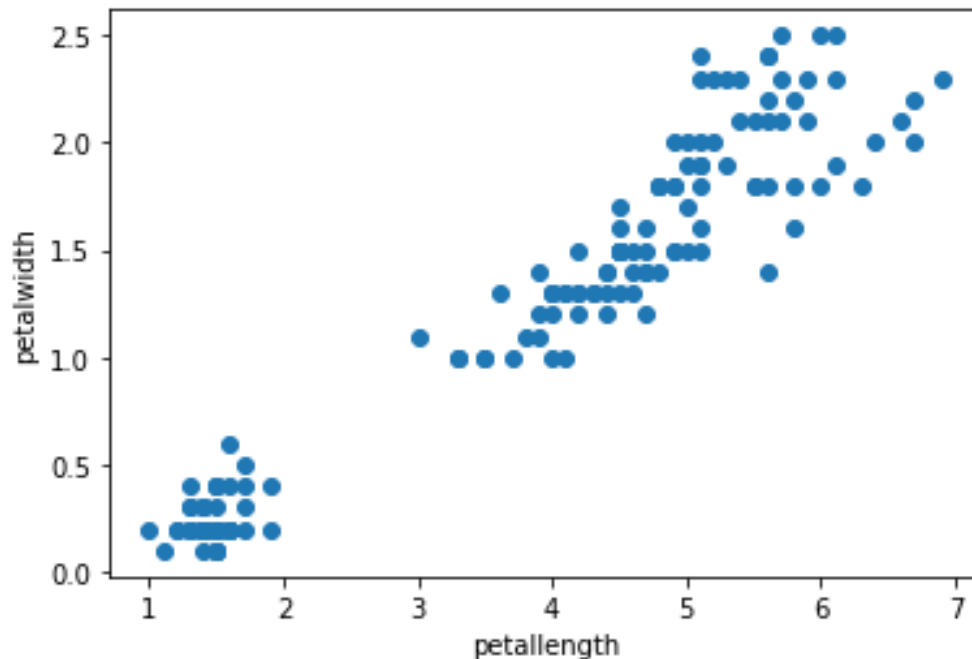


Diagrama de dispersión entre la longitud y el ancho del pétalo de una flor.

Coeficiente de correlación lineal

- Dados dos atributos X e Y el coeficiente de correlación lineal entre ellos se calcula de la siguiente forma

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

siendo $\text{Cov}(X, Y)$ la covarianza entre X e Y y σ_X y σ_Y los desvíos de cada variable.

Covarianza y desvío estándar

- Dadas dos variables X y Y

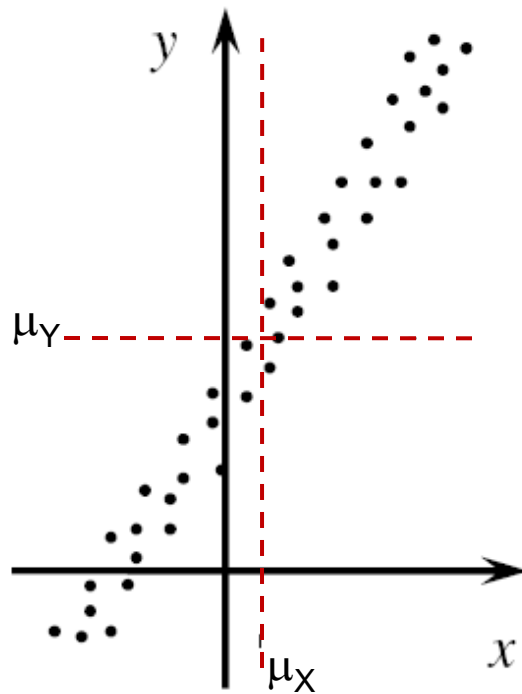
$$\text{Cov}(X, Y) = \left[\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right] / N$$

$$\sigma_X = \sqrt{\left[\sum_{I=1}^N (x_i - \mu_X)^2 \right] / N}$$

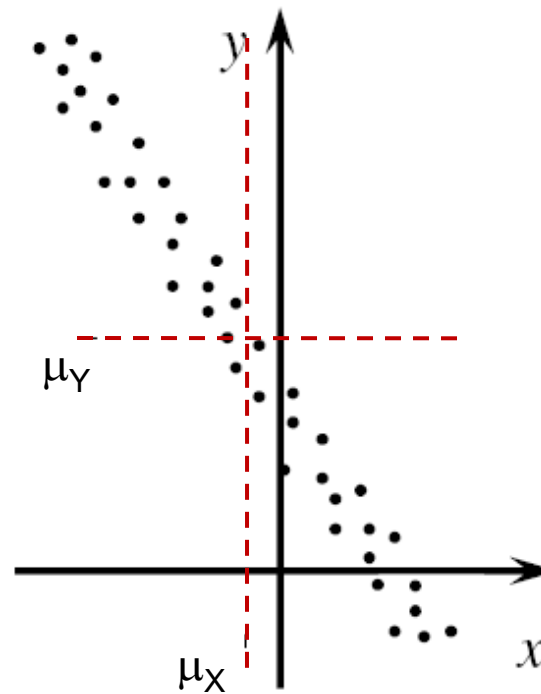
Covarianza

$$\text{Cov}(X, Y) = \left[\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right] / N$$

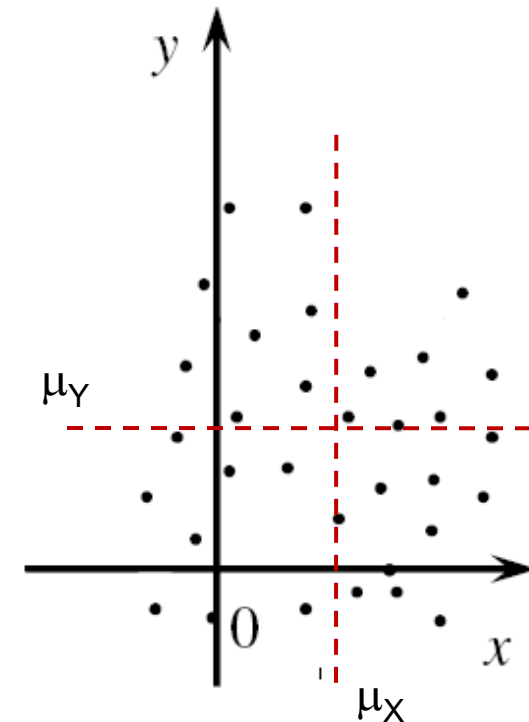
- La **covarianza** es un valor que indica el grado de variación conjunta de dos **variables aleatorias** respecto a sus medias.



Covarianza Positiva



Covarianza Negativa



Covarianza cercana a cero

Coeficiente de correlación lineal

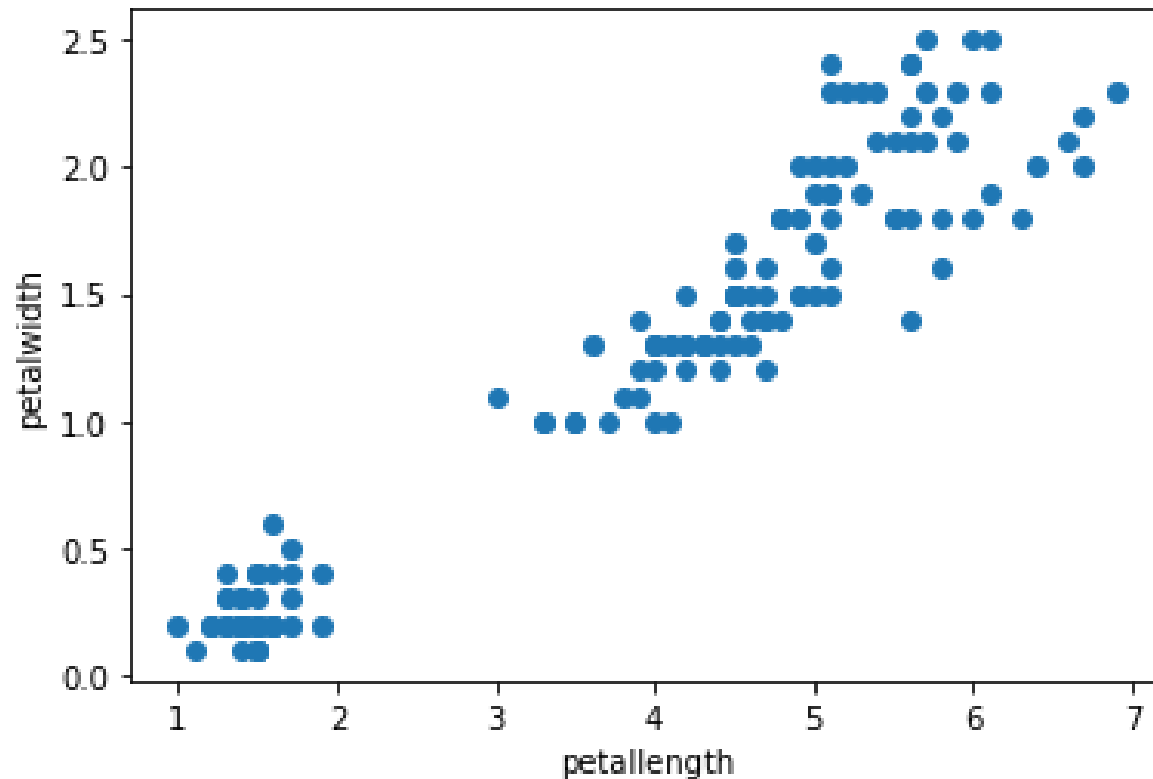
INTERPRETACION

- Si $0.5 \leq \text{abs}(\text{Corr}(A,B)) < 0.8$ se dice que A y B tienen una correlación lineal débil.
- Si $\text{abs}(\text{Corr}(A,B)) \geq 0.8$ se dice que A y B tienen una correlación lineal fuerte
- Si $\text{abs}(\text{Corr}(A,B)) < 0.5$ se dice que A y B no están correlacionados linealmente. Esto NO implica que son independientes, sólo que entre ambos no hay una correlación lineal.

Ejemplo

Correlacion_Iris.ipynb

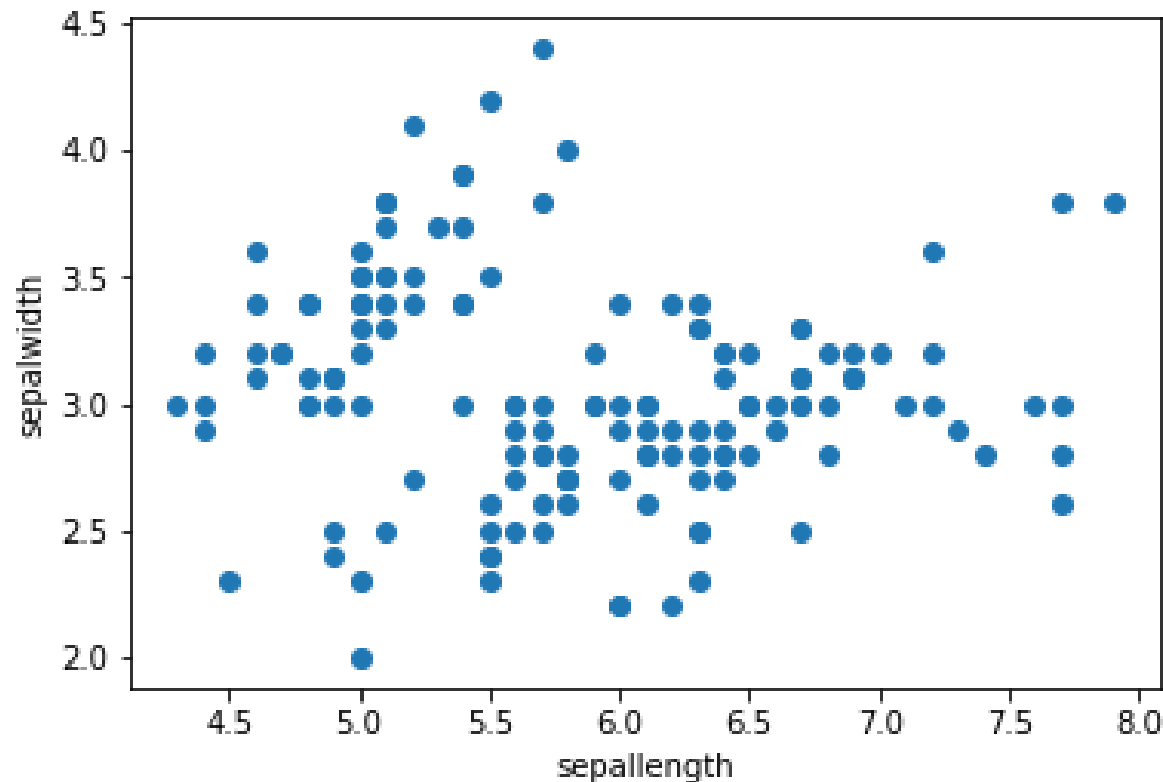
- El valor del **coeficiente de correlación lineal** entre los atributos PETALLENGTH y PETALWIDTH es **0.96**



Ejemplo

Correlacion_Iris.ipynb

- El valor del **coeficiente de correlación lineal** entre los atributos SEPALLENGTH y SEPALWIDTH es **-0.11**



Resumen

□ Tipos de Variables

- ▣ Cuantitativas y cualitativas

□ Descripciones estadísticas

- ▣ Medidas de tendencia central
 - Media, moda, mediana, rango medio
- ▣ Medidas de dispersión
 - Varianza, desviación estándar
 - Rango
 - Cuartiles, Rango intercuartil

□ Gráficos

- ▣ Diagrama de barras
- ▣ Diagrama de torta
- ▣ Histograma
- ▣ Diagrama de caja
- ▣ Diagrama de dispersión
Coeficiente de correlación lineal