Medidas de dispersión

Las medidas de dipersión son el complemento a las medidas de tendencia central, y nos indican qué tan dispersos están los datos respecto de las medidas centrales.

Vamos a conocer tres conceptos que nos van a permitir trabajar esta información: **rango, rango intercuartil** y **desviación standard.**

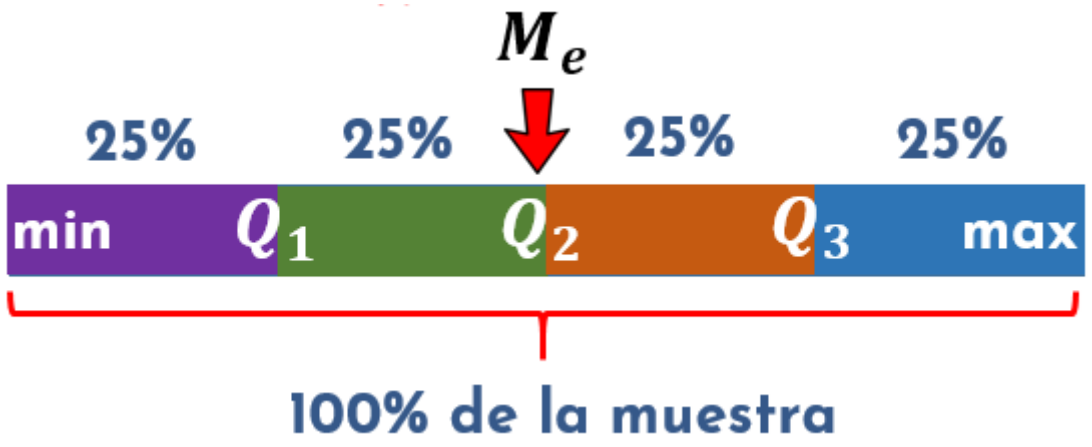
.1 Rango

El rango es el intervalo entre el valor máximo y el valor mínimo.

.2 Cuartiles, deciles y percentiles

En un conjunto de datos en el que éstos se hallan ordenados de acuerdo con su magnitud, el valor de en medio (o la media aritmética de los dos valores de en medio), que divide al conjunto en dos partes iguales, es la ***mediana***. Continuando con esta idea se puede pensar en aquellos valores que dividen al conjunto de datos en cuatro partes iguales. Estos valores, denotados Q_1, Q_2 y Q_3 son el ***primero, segundo y tercer cuartiles***, respectivamente; el valor Q_2 coincide con la mediana.

Fig –



Cuartiles

De igual manera, los valores que dividen al conjunto en diez partes iguales son los ***deciles*** y se denotan D_1, D_2, . . . , D_9, y los valores que dividen al conjunto en 100 partes iguales son los ***percentiles*** y se les denota P_1, P_2, . . . , P_{99}. El quinto decil y el percentil 50 coinciden con la mediana. Los percentiles 25 y 75 coinciden con el primero y tercer cuartiles, respectivamente.

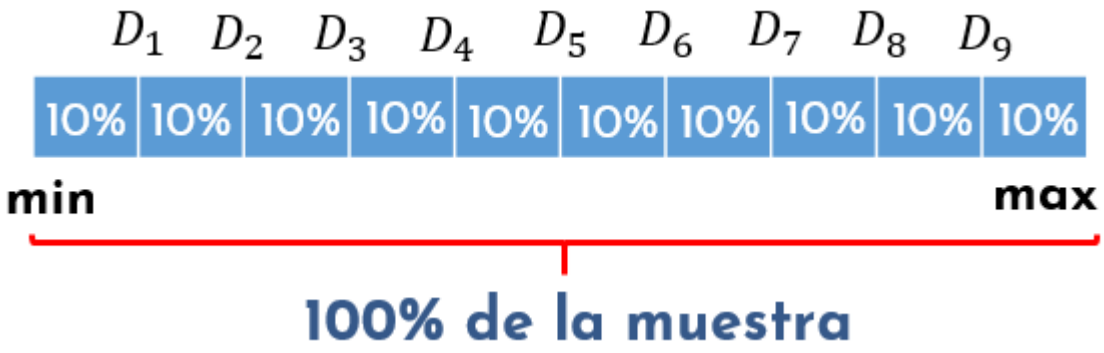


Fig – Deciles

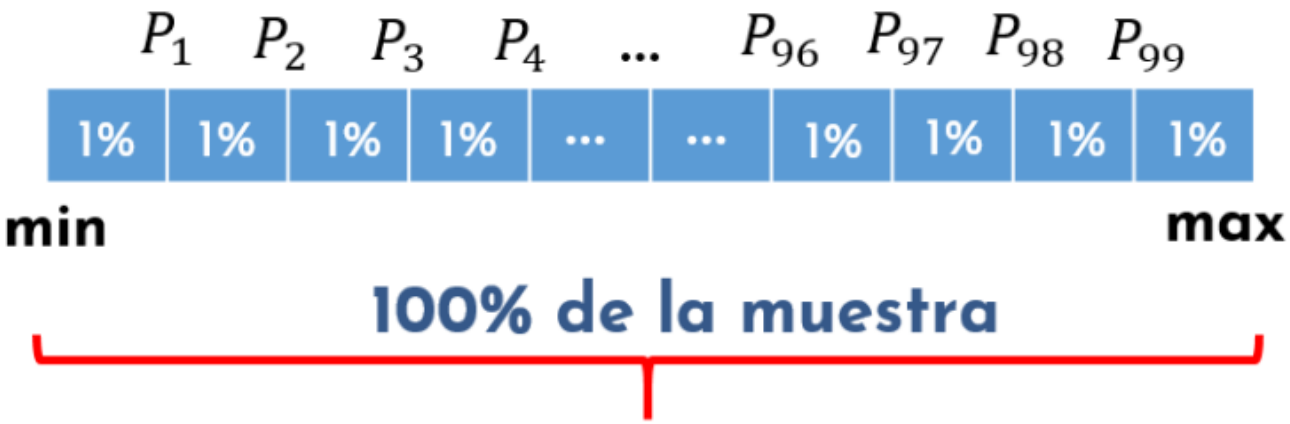
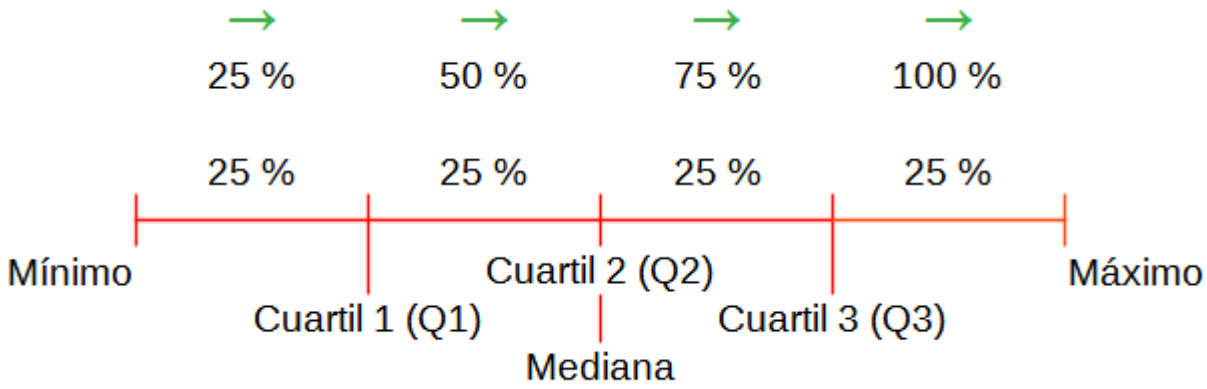


Fig – Percentiles

A los cuartiles, deciles, percentiles y otros valores obtenidos dividiendo al conjunto de datos en partes iguales se les llama en conjunto ***cuantiles***.

a) Los cuartiles

Los cuartiles son los tres elementos de un conjunto de datos ordenados que dividen el conjunto en cuatro partes iguales.



b) Características de los cuartiles

- El cuartil 1 (Q_1) es el percentil 25 (P_{25}). El 25 % de los datos son menores o iguales a Q_1.
- El cuartil 2 (Q_2) es la mediana y el percentil 50 (P_{50}). El 50 % de los datos son menores o iguales a Q_2.
- El cuartil 3 (Q_3) es el percentil 75 (P_{75}). El 75 % de los datos son menores o iguales a Q_3.

c) Cálculo de los cuartiles

Dado el siguiente set de datos con 20 elementos ordenados:

edad = [19, 21, 24, 28, 28, 29, 30, 32, 33, 34, 37, 40, 45, 45, 52, 53, 54, 56, 60, 63]

Hallamos el cuartil 2 (Q_2) calculando la mediana. Al ser una cantidad de elementos par, debemos tomar los dos centrales y promediarlos:

$i = N/2 = 20/2 = 10 \rightarrow Q2 = (x_{10} + x_{11})/2 = (34 + 37)/2 =$ **35, 5**

Para el cuartil 1 (Q_1) y cuartil 3 (Q_3) hallaremos su posición mediante las siguientes fórmulas:

$Q1 = (N+1)/4$

$Q3 = 3(N+1)/4$

Esto nos dará la posición del elemento que coincide con el cuartil correspondiente.

En este caso, el primer cuartil es $(N+1)/4 = (20+1)/4 = 21/4 = 5, 25$. Como el resultado tiene parte decimal, resulta que el primer cuartil se encuentra entre los elementos 28 y 29.

edad = [19, 21, 24, 28, x_5 =28, x_6 =29, 30, 32, 33, 34, 37, 40, 45, 45, 52, 53, 54, 56, 60, 63]

Si consideramos que nuestro cuartil está entre los datos i e i+1, vamos a utilizar la siguiente fórmula, sabiendo que x_i es la parte entera del dato obtenido, y d es la parte decimal:

$$Q_1 = x_i + d \cdot (x_{i+1} - x_i)$$

Entonces: $Q_1 = x_5 + 0, 25. (x_6 - x_5) = 28 + 0, 25. (29 - 28) =$ **28, 25**

Si el resultado no tiene parte decimal: elegimos ese mismo objeto. Por ejemplo, si el conjunto tiene 19 elementos, $(N+1)/4 = (19+1)/4 = 20/4 = 5$, el primer cuartil será $Q_1=X_5$, o sea… el elemento de la 5ª posición.

Ahora, $Q3 = 3(N+1)/4 = 63/4 = 15, 75$. Como el número es decimal, el cuartil estará entre $X_{15}=52$ y $X_{16}=53$.

edad = [19, 21, 24, 28, 28, 29, 30, 32, 33, 34, 37, 40, 45, 45, x_{15} =52, x_{16} =53, 54, 56, 60, 63]

El número decimal es el 15,75, por lo que i=15 y d=0, 75. El cuartil 3 es:

$Q_3 = x_{15} + 0, 75. (x_{16} - x_{15}) = 52 + 0, 75. (53 - 52) =$ **52, 75**

Los cuartiles quedarían distribuidos de la siguiente forma:

19 21 24 28 28 29 30 32 33 34 37 40 45 45 52 53 54 56 60 63

Q1 Q2 Q3

Medidas de dispersión (2)

d) Trabajando con datos agrupados

Cuando estamos trabajando con datos agrupados, debemos modificar la manera de calcular los cuartiles.

Previamente, necesitamos conocer la **frecuencia absoluta acumulada** (F_i): es el resultado de ir sumando las frecuencias absolutas de los valores de la muestra. Por lo tanto, previamente debemos calcular las frecuencias absolutas de la muestra.

Salarios (en miles de pesos)	Frecuencia absoluta (f_i)	Frecuencia acumulada (F_i)
$ 250,00 – $ 259,99	8	8
$ 260,00 – $ 269,99	10	18
$ 270,00 – $ 279,99	16	34
$ 280,00 – $ 289,99	15	49
$ 290,00 – $ 299,99	10	59
$ 300,00 – $ 309,99	5	64
$ 310,00 – $ 319,99	3	67
$ 320,00 y más	3	70
Total	70	

Como podemos ver en el cuadro del ejemplo, la columna F_i es resultado de sumar la frecuencia de ese fragmento de la población con la F_i del fragmento anterior.

Ya estamos listos para calcular los cuartiles. Primero, debemos calcular las posiciones en las que estarán los tres cuartiles:

$$c = 1, 2, 3$$

$$Q_c = c.(N+1)/4$$

$$Q_1 = (N+1)/4$$

$$Q_2 = (N+1)/2$$

$$Q_3 = 3.(N+1)/4$$

Ahora debemos determinar en qué intervalos (I_i) están ubicadas estas tres posiciones, viéndolo en la columna de frecuencias absolutas acumuladas de la tabla.

Si alguna de esas tres posiciones calculadas fuera un número entero, el dato correspondiente a esa posición será el cuartil buscado.

La fórmula para calcular los cuartiles en datos agrupados es:

$$Q_c = L_i + \frac{n°.Q_c - F_{i-1}}{f_i}.I_i$$

Donde:

Q_c es uno de los tres cuartiles

L_i es el límite inferior del intervalo I_i en que está cada cuartiles

$n°\ Q_c$ es la posición calculada del cuartil

N_{i-1} es la frecuencia absoluta acumulada del intervalo anterior

n_i es la frecuencia absoluta del intervalo en que está el cuartil

En el ejemplo de los salarios:

1. Calculamos las posiciones de los cuartiles, sabiendo que la cantidad de datos N es de 70:
2. Cada cuartil aparece en su intervalo, a partir de la columna de la frecuencia acumulada N_i. Aparecen sombreadas las frecuencias acumuladas de los intervalos anteriores N_{i-1}.

Salarios (en miles de pesos)	Frecuencia absoluta (f_i)	Frecuencia acumulada (F_i)	
$ 250,00 – $ 259,99	8	8	
$ 260,00 – $ 269,99	**10**	**18**	← Q_1 = 17,75
$ 270,00 – $ 279,99	16	34	
$ 280,00 – $ 289,99	**15**	**49**	← Q_2 = 35,50
$ 290,00 – $ 299,99	**10**	**59**	← Q_3 = 53,25
$ 300,00 – $ 309,99	5	64	
$ 310,00 – $ 319,99	3	67	
$ 320,00 y más	3	70	
Total	70		

3. Con estos dato se calculan los tres cuartiles, sabiendo que la amplitud del intervalo I_i es 9,99:

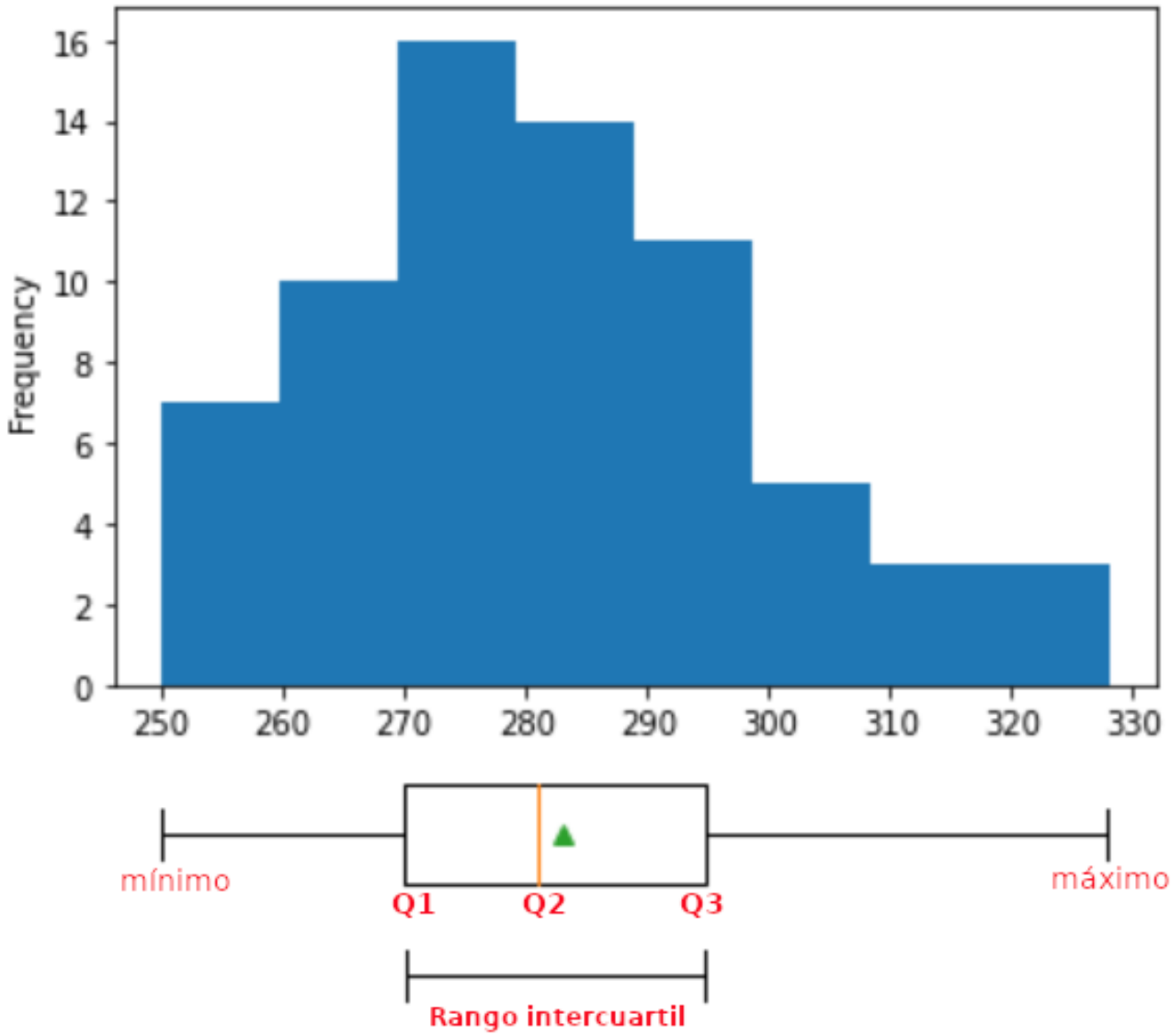
$$Q_1 = 260 + (17,75-8)/10.9,99 = 269,74$$

$$Q_2 = 280 + (35,50-34)/15.9,99 = 280,99$$

$$Q_3 = 290 + (53,25-49)/10.9,99 = 294,25$$

e) Diagramas de caja y bigote

Rango intercuartil o IQR: es la diferencia entre el tercer y primer cuartil ($r = Q_3 - Q_1$)


Frequency
16
14
12
10
8
6
4
2
0
250
260
270
280
290
300
310
320
330
mínimo
Q1
Q2
Q3
máximo
Rango intercuartil

Medidas de dispersión (3)

.3 Desviación standard

a) Cálculo

Es la medida de dispersión más utilizada en el área de la estadística descriptiva.

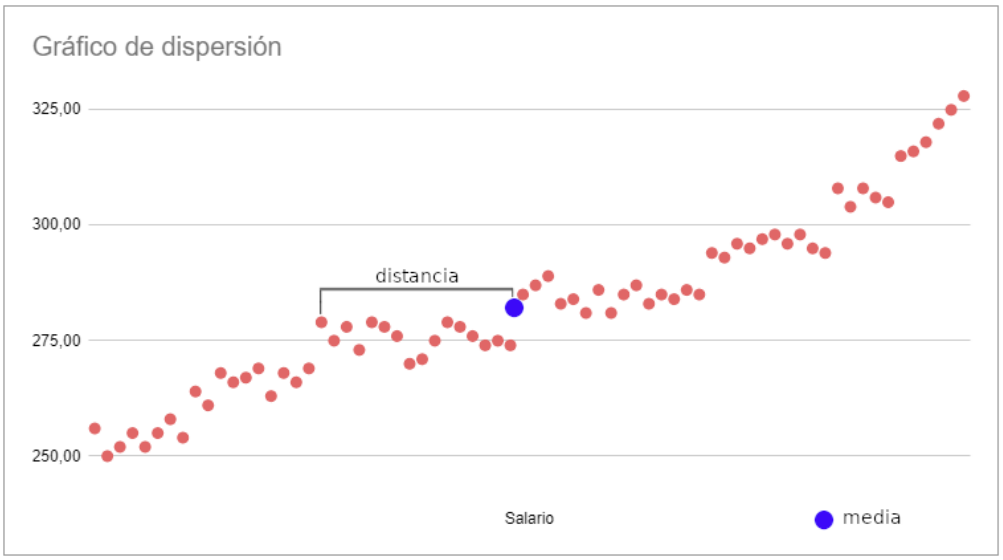
Dado un dataset = $\{x_1, \ x_2, \cdots \ x_n \ \}$, definimos la media $\bar{x}$ (también llamada μ)

$$\bar{x} = \frac{1}{N}.\sum_{i=1}^{N} \quad x_i$$

Conceptualmente, lo que estaremos haciendo es calcular la distancia entre la media y un punto cualquiera del set de datos (x_i). La fórmula es:

$$(x_i - \mu)^2 \rightarrow cuadradodeladistanciaentrelamediayunpuntox_i$$

Aplicamos el cuadrado para evitar que la diferencia nos dé un número negativo.



La desviación standard viene dada por un concepto llamado varianza (σ^2), que es la sumatoria de los cuadrados de las distancias entre todos los puntos y la media, dividido por la cantidad de elementos.

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

La desviación standard no es más que la raíz cuadrada de la varianza, o sea:

$$\sigma = \sqrt{\frac{1}{n}\sum_{n}^{i=1} \quad (x_i - \mu)^2} \rightarrow desviaciónstandard$$

Nota: cuando se trabaja con la desviación standard de sólo una parte de la población (lo llamamos **muestra**), se agrega el factor de corrección al denominador de la fórmula de cálculo de variación standard de la población, quedando

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

b) Significado

i. En una distribución de Gauss

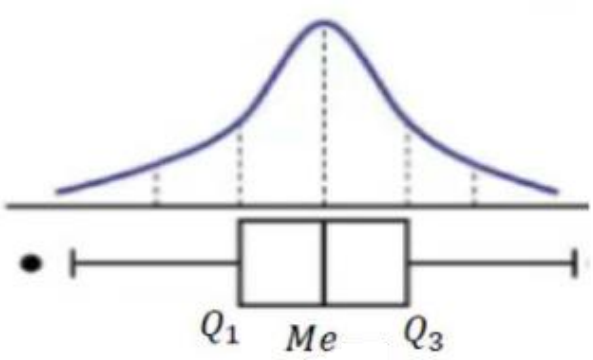
Dada una distribucion normal o gaussiana, se verifica que

$$\mu \pm 3\sigma \rightarrow contieneel99{,}73\%delosdatos$$

Esto significa que los datos que se encuentran más allá de 3 desviaciones standard de la media, se consideran anómalos o ***outliers***.

Otra manera de detectar outliers, es mediante la fórmula del método de detección de outliers con el rango intercuartil (IQR):

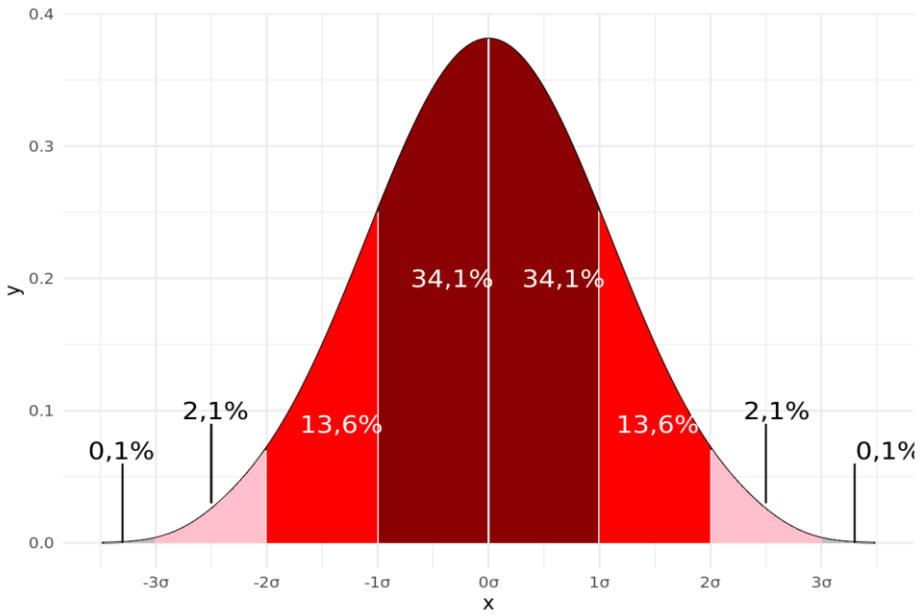
$$Datosnormales \rightarrow \{^{min = Q_1 - 1{,}5IQR}_{max = Q_3 + 1{,}5IQR}\}$$



¿Cómo se distribuyen los datos considerando desviaciones menores?

$\mu \pm 2\sigma \rightarrow contieneel95{,}45\%delosdatos$

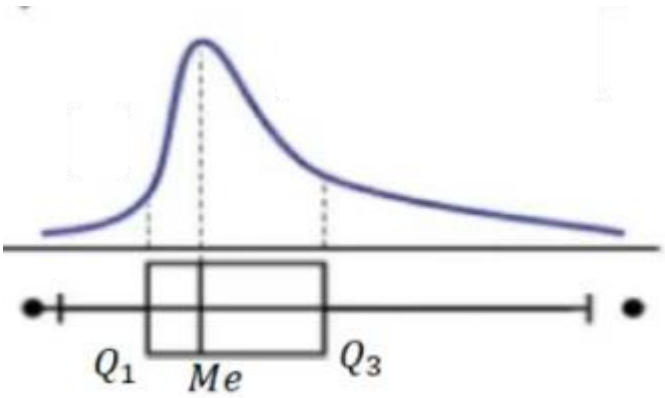
$\mu \pm \sigma \rightarrow contieneel68{,}27\%delosdatos$



ii. En distribuciones asimétricas o sesgadas

En este caso **no podemos** aplicar la fórmula del método de detección de outliers con el rango intercuartil. Lo que suele hacerse es generalizar la noción de cómo definir el criterio a partir del cual los datos se consideran anómalos por medio de una función que nos permita calcular perfectamente el sesgo de la distribución.

$$Datosnormales \rightarrow \{^{min = Q_1 - 1{,}5.\, f_{(x)}.\, IQR}_{max = Q_3 + 1{,}5.\, g_{(x)}.\, IQR}\}$$



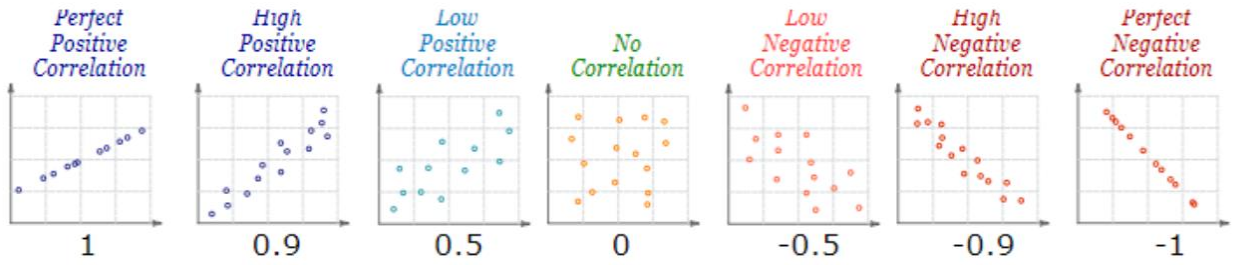
En conclusión: cuando tenemos una distribución simétrica se suele trabajar con la desviación standard, en cambio, con distribuciones sesgadas, es más conveniente utilizar los el rango intercuartil para determinar los datos válidos.

Correlaciones

.1 Correlación

La correlación es una medida estadística que expresa hasta qué punto dos variables están relacionadas linealmente (esto es, cambian conjuntamente a una tasa constante).

En la figura siguiente podemos ver algunos de los distintos tipos de correlaciones que podemos encontrar, variando desde $r=-1$ (correlación negativa perfecta), hasta $r=1$ (correlación positiva perfecta). Justo en medio podemos encontrar $r=0$ (no hay ningún tipo de correlación), y diversos valores $-1<r<0$ y $0<r<1$, que indican diferentes grados de correlación intermedios: altas negativas, bajas negativas, bajas positivas y altas positivas.



Esta información nos sirve para realizar reducción de datos: si dos variables están fuertemente correlacionadas, es posible que ambas estén aportando la misma información. Deberíamos eliminar una de ellas de nuestro análisis.

Este proceso no sólo funciona con variables numéricas. También podemos convertir nuestras variables categóricas a numéricas o booleanas, analizar si tienen fuertes correlaciones e ir reduciendo el número de variables a analizar.

Partiendo de la fórmula que hemos aprendido:

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

Vamos a calcular la **covarianza:** es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.

$$covarianza_{(x,y)} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu_x).(y_i - \mu_y)$$

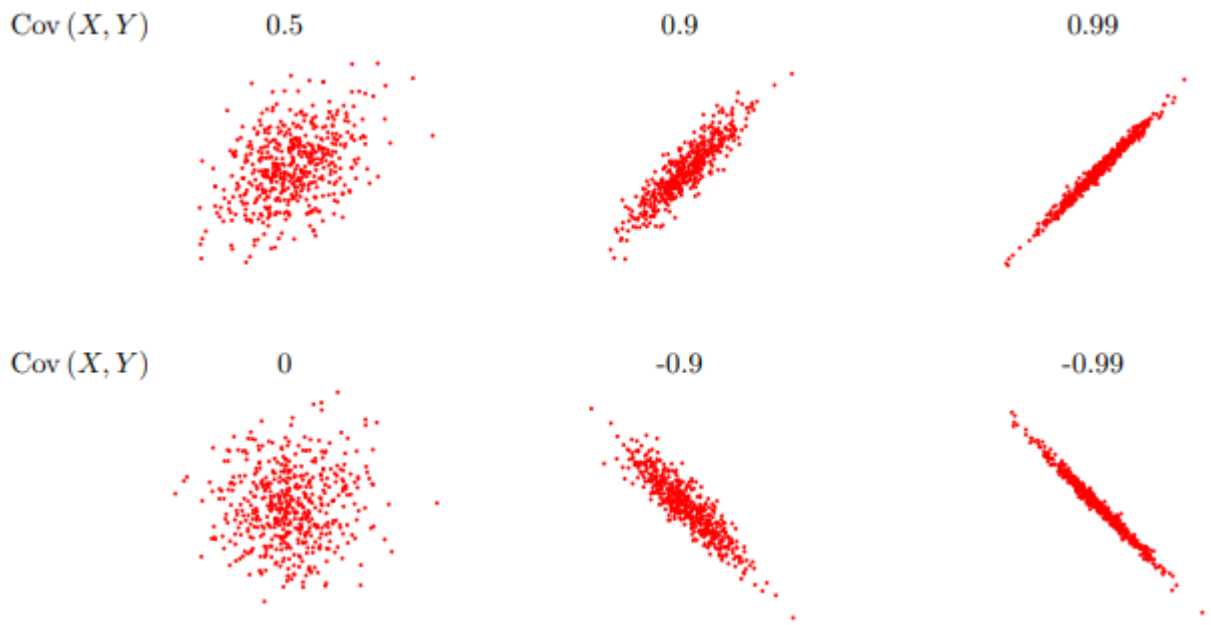
Es posible que x e y estén utilizando diferentes magnitudes o unidades, por lo que debemos buscar una forma más standard para normalizar el grado de variación de cada variable. Esto se hace dividiendo por la desviación standard cada una de las desviaciones de la variable. Esto “reduce“ ambas variables a la misma escala.

Esto nos introduce al llamado **coeficiente de correlación** ρ, siendo la covarianza dividida la desviación standard de x, multiplicada por la desviación standard de y:

$$\rho = \frac{covarianza}{\sigma_x . \sigma_y}$$

El coeficiente de correlación es la medida específica que cuantifica la intensidad de la relación lineal entre dos variables en un análisis de correlación.

Cuando tenemos un coeficiente de correlación alto, es porque las dos variables tienen un grado de correlación elevado, si el coeficiente se acerca a 0, indica que las variables no tienen correlación y si el coeficiente de correlación es cercano a -1, las variables tienen una relación inversa.



Debemos tener cuidado al decir que dos variables están fuertemente relacionadas: **una fuerte relación no implica una relación causa-efecto.** En otras palabras, puede que si una variable aumenta su valor, no sea la causa por la que otra tenga la misma tasa de variación.
Esta herramienta sirve simplemente para eliminar información redundante de nuestro set de datos a analizar.

.2 Matriz de covarianzas

La matriz de covarianzas, o matriz Σ, se utiliza cuando tenemos que evaluar la correlación de más de dos variables, contemplando todas las combinaciones de parejas de datos posibles. Esto facilita el **análisis de componentes principales (PCA)**, al agilizar el proceso de **reducción de datos**, eliminando alguna de las dos variables.

En la matriz de covarianzas, los elementos fuera de la diagonal contienen las covarianzas de cada par de variables. Los elementos de la diagonal contienen las varianzas de cada variable. Recordamos que la varianza mide qué tan dispersos se encuentran los datos alrededor de la media y es igual al cuadrado de la desviación estándar.

...	x	y	z

$$\Sigma\rightarrow \begin{array}{c|cccc} & x & y & z & \dots \\ x & \sigma^2_{(x)} & cov_{(x,y)} & cov_{(x,z)} & \dots \\ y & cov_{(y,x)} & \sigma^2_{(y)} & cov_{(y,z)} & \dots \\ z & cov_{(z,x)} & cov_{(z,y)} & \sigma^2_{(z)} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{array}$$