scidata

Medidas de Dispersión y de Forma

Varianza, desviación estándar, etc.

Instructor: Juan Luis Palacios Soto

palacios.s.j.l@gmail.com

Contenido

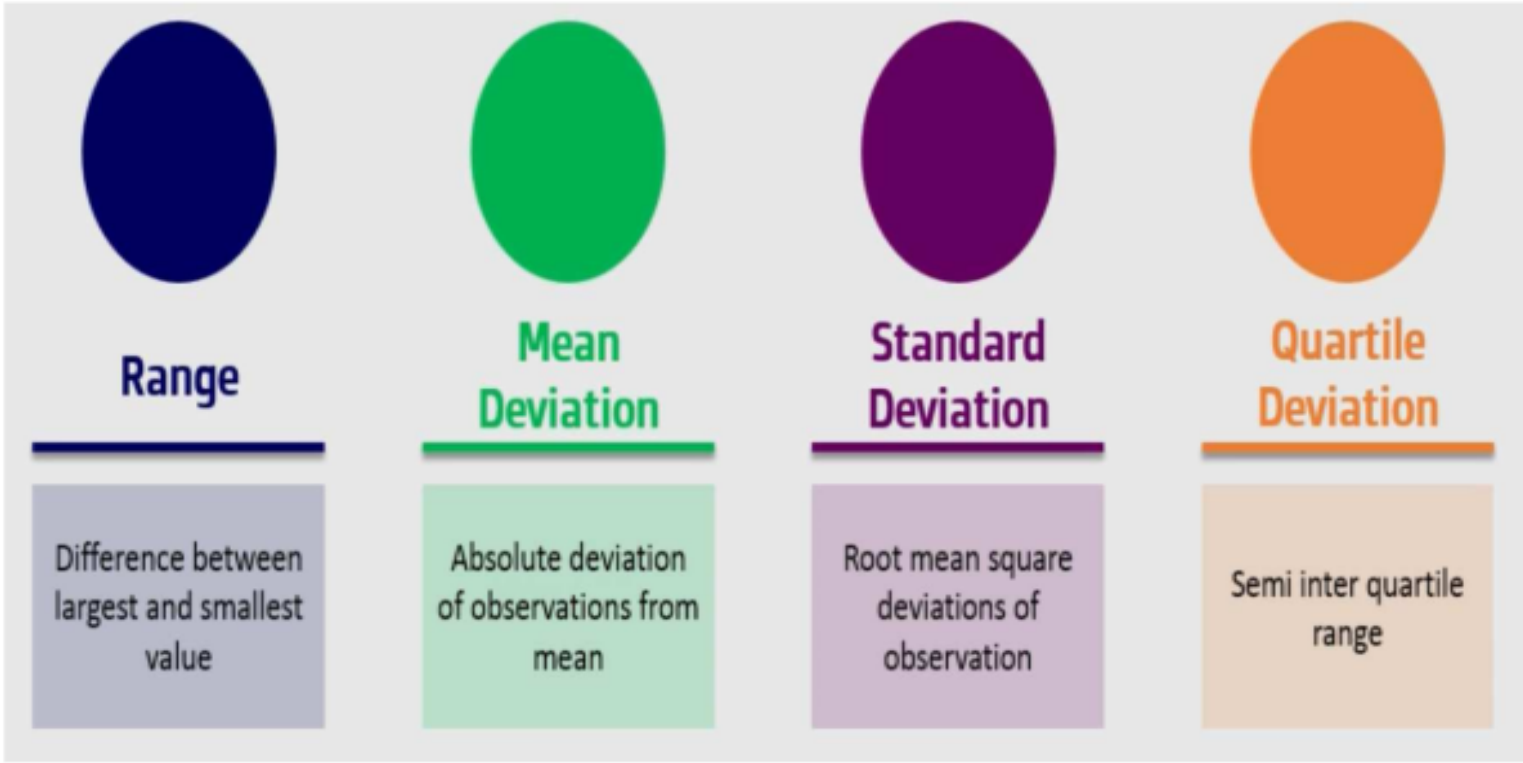
Medidas de dispersión

1. El rango
2. Desviación media absoluta
3. Desviación intercuartil
4. Desviación estándar
5. Varianza

Las dispersiones o desviaciones son una manera de medir qué tan alejados están los puntos de alguna medida de tendencia central. Se pueden calcular para cualquiera de las medidas de tendencia central que estudiamos en el capítulo anterior.

Básicamente, son medidas que utilizamos para conocer que tan dispersos se encuentran nuestros datos de algún valor central. Esto significa que si hay una desviación alta, entonces nuestros datos se alejan mucho de la medida de tendencia central y por lo tanto esta última no es un buen representante de nuestros datos.

Básicamente son cuatro: el rango, desviación absoluta media, la desviación intercuartil y la desviación estándar.



Rango

Se define simplemente como la diferencia entre el valor más alto y el valor más pequeño. Por lo tanto, se ve fuermente afectado por los valores atípicos. Es por esta razón que no se trata de la mejor medida de dispersión, además de que su cálculo no involucra ninguna medida de tendencia central.
Sean $x_{(1)}, x_{(2)}..., x_{(n)}$, n datos ordenados de manera ascendente, el rango es es:

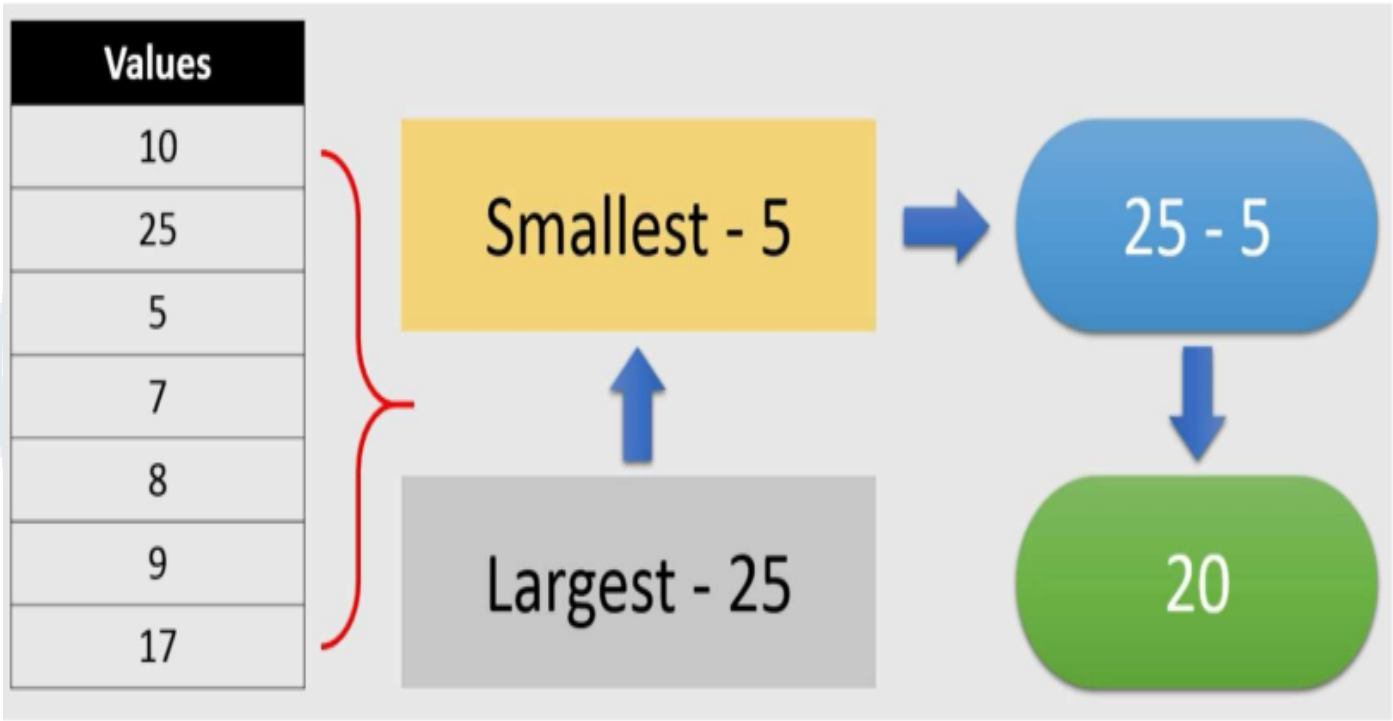
$$R = Valor_{max} - Valor_{min} = x_{(n)} - x_{(1)}$$

El coeficiente del rango se calcula como

$$CR = \frac{R}{x_{(n)} + x_{(1)}}$$

Definición (El rango en Excel)

No existe en Excel una manera de calcular el rango de los datos directamente; sin embargo, recuerda que Excel tiene las funciones MIN y MAX.



Definición (Desviación media absoluta)

Considera directamente todas los valores de los datos uno por uno. Se trata de medir cómo se aleja, en promedio, la "distancia" entre cada dato y alguna medida de tendencia central. Generalmente esta última se toma como la media aritmética o la mediana.

Sean $x_1, x_2, ..., x_n$, n *datos y sea* A *alguna medida de tendencia central. Entonces la desviación media absoluta se calcula como*

$$MD = \frac{1}{n} \sum_{i=1}^{n} |x_i - A|$$

Si $A = \bar{x}$, *entonces la fórmula queda*

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

El coeficiente de la desviación absoluta se calcula como

$$CMD = \frac{MD}{\bar{x}}$$

Definición (Desviación media absoluta en Excel)

En Excel, cuando la medida de tendencia central es la media aritmética (es decir, $A = \bar{x}$ en la fórmula de MD), se utiliza la función DESVPROM.

x	$\lvert x - \bar{x} \rvert$
4	8
10	2
15	3
12	0
13	1
18	6

$(4 + 10 + 15 + 12 + 13 + 18) = 72$

$$\frac{72}{6} = 12$$

$$\frac{20}{6}$$

Definición (Desviación intercuartil)

Representa qué tanto se desvía el 50 % de los datos en promedio con respecto de la mediana, por lo que se calcula utilizando Q1 y Q3. Como no toma en cuenta valores debajo de Q1 ni arriba de Q3, no se ve afectado por los valores extremos. Tampoco se ve afectado si se suma una constante a todos los datos, pero sí se afecta si se multiplican todos los datos por cualquier número.

Se calcula como:

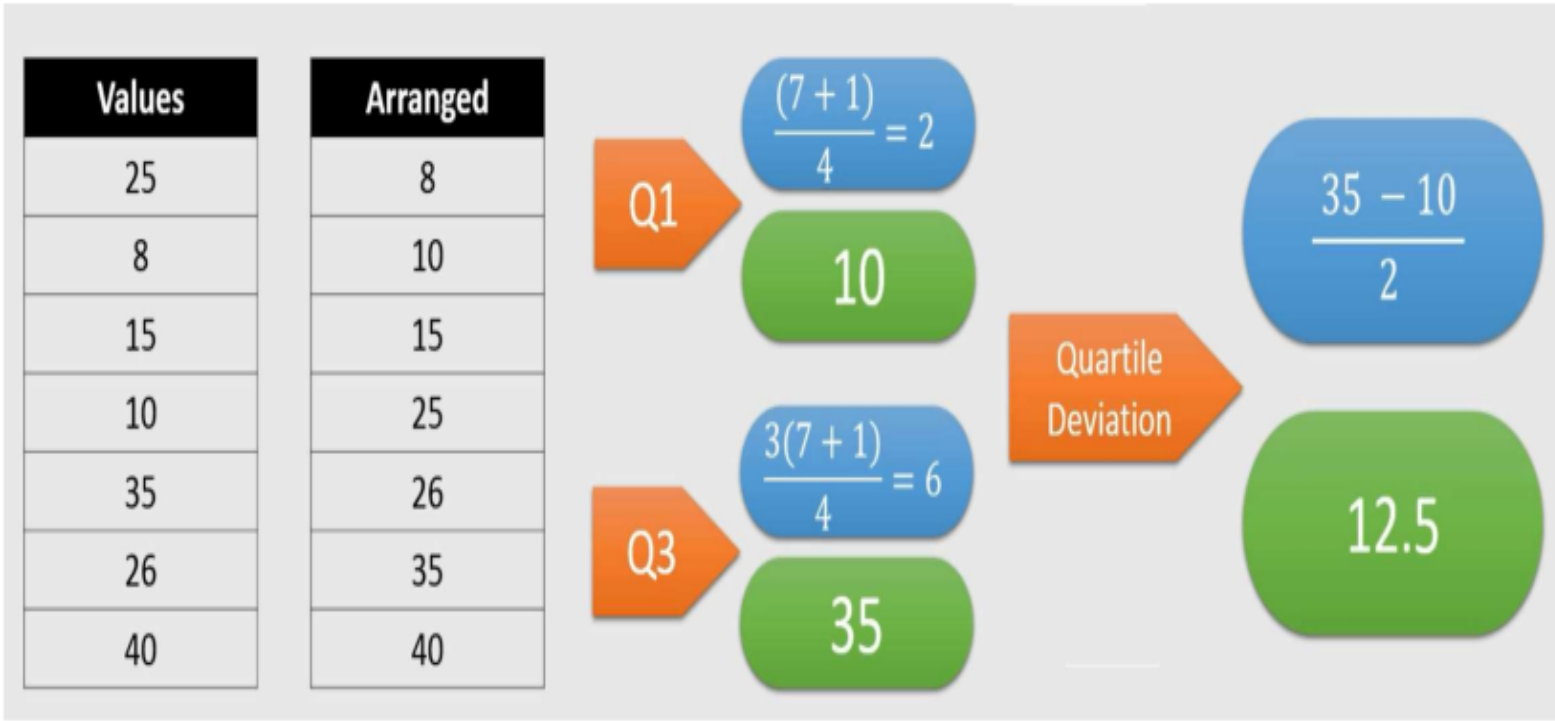
$$DQ = \frac{Q_3 - Q_1}{2}$$

Al numerador de la fórmula anterior se le conoce como rango intercuartil y se represente como RQ.
El coeficiente de la desviación intercuartil se calcula como

$$CDQ = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Definición (Desviación intercuartil en Excel)

No existe en Excel una función directa para calcularlo, pero recordemos que tenemos la función CUARTIL.EXC.



Definición (Desviación estándar)

Es considerada como la mejor medida de dispersión. Se utiliza, junto con la media aritmética, para dar reglas empíricas acerca de los valores poco comunes (como la regla de las desviaciones para distribuciones gaussianas). Al igual que la anterior, no se ve afectada por si sumamos un valor a todos los datos; sin embargo, se afecta (y en gran medida) si todos los datos se multiplican por un mismo número.

Sean $x_1, x_2, ..., x_n$ *un conjunto de datos numérico con media* $\bar{x}$*, su desviación estándar, denotada por* SD*, se calcula como*

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

El coeficiente de la desviación estándar se calcula como

$$CSD = \frac{SD}{\bar{x}}$$

Otras notaciones muy comunes para la desviación estándar en libros de texto es S.

Definición (Desviación estándar en Excel)

En Excel existen dos funciones para calcular desviación estándar: DESVEST.M y DESVEST.P. La diferencia entre ellas es que la primera calcula la desviación estándar MUESTRAL (y por lo tanto aplica la fórmula de la diapositiva anterior) y la segunda calcula la desviación estándar POBLACIONAL (que en lugar dividir por $n-1$ en la fórmula anterior, divide por el tamaño de la población).

Class (x)	x - $\bar{x}$	(x - $\bar{x}$)^2
25	10	100
10	- 5	25
15	0	0
12	- 3	9
13	- 2	4

(25 + 10 + 15 + 12 + 13) = 75

$$\frac{75}{5} = 15$$

(100 + 25 + 0 + 9 + 4) = 138

$$\sqrt{\frac{138}{4}}$$

5.25

Definición (Varianza)

Básicamente se trata de la desviación estándar elevada al cuadrado.

Sean $x_1, x_2, ..., x_n$ *un conjunto de datos numérico con media* $\bar{x}$*, su varianza, denotada por* Var*, se calcula como*

$$Var = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Definición (Varianza en Excel)

Nuevamente, en Excel existen dos funciones para calcular la varianza: VAR.M y VAR.P. La diferencia entre ellas es que la primera calcula la varianza MUESTRAL (sample, en inglés) y la segunda calcula la varianza POBLACIONAL.

Class (x)	$x - \bar{x}$	$(x - \bar{x})^2$
30	-11	121
22	- 19	361
78	37	1369
20	- 21	441
55	14	196

$(30 + 17 + 78 + 20 + 55) = 205$

$\frac{205}{5} = 41$

$(121 + 361 + 1369 + 441 + 196) = 2488$

$\frac{2488}{4}$

622

Definición (Coeficiente de asimetría o sesgo)

Sean $x_1, x_2, ..., x_n$ un conjunto de datos con media $\bar{x}$ y desviación estándar muestral S, entonces se llama **coeficiente de asimetría o coeficiente de sesgo** *a la medida que representa el grado de asimetría de la gráfica y lo denotaremos por CA. Una de las fórmulas para calcular este número se usa*

$$CA = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S} \right)^3.$$

El CA caracteriza el grado de alejamiento de los datos con respecto a su media y generalmente se encuentra entre -4 y 4.

$$CA = \begin{cases} 0, & \text{los datos son simétricos} \\ < 0, & \text{sesgo a la izquierda} \\ > 0, & \text{sesgo a la derecha.} \end{cases} \tag{1}$$

Si $CA \approx 0$ tiene sentido preguntarse qué tan concentrados están de la media.

Definición (Coeficiente de asimetría en Excel)

Nuevamente, en Excel existen dos funciones para calcular el coeficiente de asimetría: COEFICIENTE.ASIMETRIA y COEFICIENTE.ASIMETRIA.P. La diferencia entre ellas es que la primera calcula para una muestra y la segunda para una población.

X	$(X-\bar{X})^3$
6	1.49
2	-23.32
5	0.00
3	-6.41
8	31.04
7	9.84
3	-6.41

$$\bar{X} = 4.85714$$
$$S = 2.26778$$
$$n = 7$$
$$CA = 0.12493$$

Definición (Kurtosis)

Sean $x_1, x_2, ..., x_n$ un conjunto de datos con media $\bar{x}$ y desviación estándar muestral S, entonces se llama **kurtosis** *a la medida que representa el grado de achatamiento de la distribución de los datos con respecto a la distribución de la llamada distribución* **normal**.

No es práctico describir una de la fórmulas de la kurtosis, sin embargo usaremos los paquetes que vienen en Excel (o en R) para que se faciliten los cálculos. Denotaremos con K el valor de la kurtosis. De esta manera tenemos lo siguiente:

$$K = \begin{cases} 0, & \text{curva normal} \\ < 0, & \text{más achatada o colas más grandes} \\ > 0, & \text{más puntiaguda o más concentrados en la media.} \end{cases}$$