TEMA 2

MÓDULO:
FUNDAMENTOS DE ESTADÍSTICA

ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL II

DOLORES LORENTE

Diplomada en Estadística y Graduada en Estadística aplicada por la UCM. Responsable científica de datos en Big Data Analytics e Innovación.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

Objetivos Específicos

2.2. Medidas de centralización, dispersión, posición, concentración y forma. Métodos gráficos.

2.2.1. Medidas de Centralización

2.2.2. Medidas de Dispersión

2.2.3. Medidas de Posición

2.2.4. Medidas de Concentración y forma

2.2.5. Métodos gráficos

Actividad: Titanic

Ideas clave

OBJETIVOS ESPECÍFICOS

- Profundizar en estadística descriptiva y controlar la información, realizando análisis y extrayendo las primeras conclusiones de los datos.
- Elegir, correctamente, el gráfico adecuado para su correcta visualización e interpretación.
- Aplicar la estadística descriptiva a situaciones de la vida real y profesional.

2.2. MEDIDAS DE CENTRALIZACIÓN, DISPERSIÓN, POSICIÓN, CONCENTRACIÓN Y FORMA. MÉTODOS GRÁFICOS.

Este apartado consta de cuatro partes importantes, en cada una de ellas se tratan unas medidas que sirven para determinar una cualidad de una variable estadística.

Un **parámetro estadístico** es un número que se obtiene a partir de los datos de una distribución estadística. Los parámetros sirven para sintetizar la información dada por una tabla o una gráfica.

Los tipos de parámetros usados son: de centralización, de dispersión, de posición y de forma.

2.2.1. MEDIDAS DE CENTRALIZACIÓN

Las medidas de centralización nos indican en torno a **qué** valor del centro se distribuyen los datos. Las medidas de centralización son:

Media aritmética

Es el valor **promedio** de la distribución. Es por tanto, el valor medio obtenido al sumar todos los datos y dividir el resultado entre el número total de datos. A la media aritmética se la denomina también *centro de gravedad* de la distribución. El símbolo para representar la media aritmética en notación griega es μ aunque muchas veces se opta por la notación simplificada $\bar{X}$, entre otras formas de escribirla.

A. Cálculo de la media aritmética para **datos sin agrupar :**

Su fórmula es:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{N}$$

EJEMPLO

Los pesos de seis amigos son: 84, 91, 72, 68, 87 y 78 kg. Hallar el peso medio.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{N} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = 80Kg$$

B. Cálculo de la media aritmética para **datos agrupados:**

Su fórmula es:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \cdots + x_n n_n}{N}$$

EJEMPLO

En un test realizado a un grupo de 42 personas, se han obtenido las puntuaciones que muestra la tabla. Calcula la puntuación media.

Intervalo	x_i	n_i	$x_i \cdot n_i$
[10, 20)	15	1	15
[20, 30)	25	8	200
[30,40)	35	10	350
[40, 50)	45	9	405
[50, 60)	55	8	440
[60,70)	65	4	260
[70, 80)	75	2	150
Σ		42	1820

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{N} = \frac{15 \cdot 1 + 25 \cdot 8 + 35 \cdot 10 + 45 \cdot 9 + 55 \cdot 8 + 65 \cdot 4 + 75 \cdot 2}{42} = \frac{1820}{42} = 43.33$$

En conclusión, la fórmula varía dependiendo de si se trata de valores para datos sin agrupar (tipo discreto) o para datos agrupados (tipo continuo).

Propiedades de la media aritmética

1. La suma de las desviaciones de todas las puntuaciones de una distribución respecto a la media es igual a cero, esto es:

$$\sum (x_i - \bar{x}) = 0$$

2. La media aritmética de los cuadrados de las desviaciones de los valores de la variable con respecto a un número cualquiera se hace mínima cuando dicho número coincide con la media aritmética. Esto traducido a una expresión matemática es:

$$\sum (x_i - \bar{x})^2 \text{ Mínimo}$$

3. Si a todos los valores de la variable se les suma un mismo número, la media aritmética queda aumentada en dicho número.

4. Si todos los valores de la variable se multiplican por un mismo número, la media aritmética queda multiplicada por dicho número.

RECUERDA

1. La media se puede hallar **sólo para variables cuantitativas.**

2. La media es **independiente de las amplitudes** de los intervalos.

3. **La media es muy sensible a las puntuaciones extremas.** Si se tiene una distribución con los siguientes pesos: 65 Kg., 69 Kg., 65 Kg., 72 Kg., 66 Kg., 75 Kg., 70 Kg. y 110 Kg. La media es igual a 74 Kg., que es una medida de centralización poco representativa de la distribución.

4. La media **no se puede calcular si hay un intervalo con una amplitud indeterminada.** Es decir, si la marca de clase del último intervalo no está definida o si tiene un valor infinito, en la última clase modal no se puede calcular la media.

5. Existen otras medias estadísticas además de la media aritmética, que es la más conocida. Estas son:

   - Media Armónica.
   - Media Geométrica.
   - Media Cuadrática.

SABÍAS QUE...

A continuación, te damos acceso a un documento para profundizar en estas medias estadísticas. El paper explica y desarrolla, de manera increíble, las distintas medias estadísticas. En él encontrarás demostraciones, ejemplos y las aplicaciones que tienen cada una de ellas. Además, contiene la explicación de las relaciones entre ellas.

Uno de los autores, Venancio Tomeo Perucha es autor de muchos libros, y es muy recomendable la lectura de cualquiera de sus trabajos.

Las medias estadísticas

Mediana

Es el **valor central** de la distribución, dicho de otra manera, es aquel que ocupa el lugar de todos los datos cuando éstos están ordenados de menor a mayor. Es, por tanto, el valor de la variable que separa la mitad superior de la distribución de la mitad inferior. Así divide la serie de datos en dos partes iguales.

La mediana se puede hallar sólo para variables cuantitativas. La mediana se representa por ***Me.***

A. Cálculo de la mediana para **datos sin agrupar.**

Los pasos a seguir son:

1. Ordenar los datos de menor a mayor.

2. Si la serie tiene un número impar, de medidas la mediana es la puntuación central de la misma. Si tenemos: 2, 3, 4, 4, 5, 5, 5, 6, 6 entonces claramente el valor central es 5, esto es: ***Me* = 5**

3. Si la serie tiene un número par de puntuaciones la mediana es la media entre las dos puntuaciones centrales. Si tenemos: 7, 8, 9, 10, 11, 12 entonces el valor medio entre esos valores es 9.5. Luego Me = 9.5.

B. Cálculo de la mediana para **datos agrupados.**

La mediana se encuentra en el intervalo donde la frecuencia acumulada llega hasta la mitad de la suma de las frecuencias absolutas. O sea, tenemos que buscar el intervalo en el que se encuentre N/ 2. Cuya fórmula matemática es:

- Donde:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot a_i$$

  - n_i es la frecuencia absoluta de la clase modal.
  - $N/2$ es la semisuma de las frecuencias absolutas.
  - N_{i-1} es la frecuencia acumulada anterior a la clase mediana.
  - a_i es la amplitud de la clase.
  - L_{i-1} es el límite inferior de la clase donde se encuentra la mediana.

Dentro de este apartado, existen **dos opciones** a su vez, que son:

- Los intervalos tienen **la misma amplitud.**
- Los intervalos tienen **distinta amplitud.**

EJEMPLO

Misma amplitud

Calcula la mediana de una distribución estadística que viene dada por la tabla:

Intervalo	n_i	N_i
[60, 63)	5	5
[63, 66)	18	23
[66, 69)	42	65
[69, 72)	27	92
[72, 75)	8	100
Σ	100	

Primero, observa que los datos del intervalo ya están ordenados de menor a mayor.

Después, si consideramos que tenemos 100 datos en total (es decir, $N = 100$). La mitad de estos son 50, por tanto, se puede escribir de la forma: $N/2 = 100/2 = 50$

El siguiente paso sería preguntarse: ¿en qué intervalo está el valor hallado (50)? La respuesta a esta pregunta el intervalo de la clase modal [66, 69).

Por tanto, situándose en dicho intervalo, al intervalo anterior le corresponderá $N_i - 1 = 23$ y el límite $L_i - 1 = 66$. Siendo el límite superior el $L_i + 1 = 69$. Con lo que ya se puede obtener el valor de la amplitud: $ai = 69 - 66 = 3$

En definitiva, con todos los datos obtenidos hasta el momento, se puede aplicar la fórmula vista anteriormente de la mediana:

Así es que, como era de esperar, **el valor central (67.93) pertenece al intervalo central [66,69).**

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot a_i = 66 + \frac{50 - 23}{42} \cdot 3 = 67.93$$

EJEMPLO

Distinta amplitud

Calcula la mediana de una distribución estadística que viene dada por la tabla:

Intervalo	ni	Ni	ai
[0, 20)	8	8	20
[20, 30)	9	17	10
[30, 40)	12	29	10
[40, 45)	10	39	5
[45, 50)	9	48	5
[50, 60)	10	58	10
[60, 80)	8	66	20
[80, 100)	4	70	20

Observa que los datos del intervalo ya están ordenados de menor a mayor. Al mismo tiempo, puedes apreciar, rápidamente, que tienen distinta amplitud.

Al igual que antes, debes calcular el valor central, luego: $N/2 = 70/2 = 35$

Al preguntarte: *¿en qué intervalo está el valor 35?* Si s la tabla, verás que el intervalo de la clase modal [30, 40) contiene el valor buscado.

Luego si $N_i = 29$ implica necesariamente que $N_i - 1 = 17$, $L_i - 1 = 30$ y, en consecuencia, $a_i = 10$. En definitiva, simplemente queda sustituir todos estos valores en la fórmula de la mediana para hallarla de la siguiente forma:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot a_i = 30 + \frac{35 - 17}{12} \cdot 10 = 45$$

Moda

Es el valor que más se repite en una distribución. La moda es el **valor más frecuente,** que tiene mayor frecuencia absoluta.

Se puede hallar la moda tanto para variables cualitativas como para variables cuantitativas. Se representa por ***Mo*.**

A. Cálculo de la moda para **datos sin agrupar.**

Analicemos los siguientes ejemplos:

1. Halla la moda de la distribución: 2, 3, 3, 4, 4, 4, 5 y 5. $\rightarrow$ $Mo = 4$

   El valor 4 es el número que más se repite, es decir, es el valor de x_i que corresponde con el **máximo valor de n_i.**

2. Halla la moda de la distribución: 1, 1, 1, 4, 4, 5, 5, 5, 7, 8, 9, 9 y 9. $\rightarrow$ $Mo = 1, 5, 9$

   Si en un grupo hay dos o varias puntuaciones con la misma frecuencia y esa frecuencia es la máxima, la distribución es bimodal o multimodal, es decir, **puede tener varias modas.** Esto sucede en este caso, en el que hay tres modas el 1, el 5 y el 9.

3. Halla la moda de la distribución: 2, 2, 3, 3, 6, 6, 9 y 9.

   Cuando todas las puntuaciones de un grupo **tienen la misma frecuencia, no hay moda.**

4. Halla la moda de la distribución: 0, 1, 3, 3, 5, 5, 7 y 8. $\rightarrow$ $Mo = 4$

   Si dos puntuaciones adyacentes tienen la frecuencia máxima, **la moda es el promedio de las dos puntuaciones adyacentes.** En este caso, serian el 3 y el 5, que se repiten dos veces cada uno de ellos, donde el valor medio entre ambos es 4.

B. Cálculo de la moda para **datos agrupados.**

Hay dos maneras de calcularla, la fórmula general y la fórmula aproximada. Veamos las dos:

Fórmula General:

$$Mo = L_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \cdot a_i$$

Fórmula Aproximada:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot a_i$$

- Donde (para ambas fórmulas):
  - n_i es la frecuencia absoluta de la clase modal.
  - n_{i-1} es la frecuencia absoluta inmediatamente inferior de la clase modal.
  - n_{i+1} es la frecuencia absoluta inmediatamente posterior de la clase modal.
  - a_i es la amplitud de la clase.
  - L_{i-1} es el límite inferior de la clase modal.

Dentro de este apartado, existen **dos opciones** a su vez, que son:

- Los intervalos tienen **la misma amplitud.**
- Los intervalos tienen **distinta amplitud.**

EJEMPLO

Misma amplitud

Calcula la moda de los salarios mensuales de 200 trabajadores de una empresa que se han recogido en la siguiente distribución de frecuencias:

C_i	n_i
[75-125)	25
[125-175)	100
[175-225)	50
[225-275)	25
Σ	200

Donde: $C_i = [L_i - 1, L_i)$

Observa que la amplitud de los intervalos es constante e igual a $a_i = 50$.

La mayor frecuencia absoluta es $n_i = 100$, por lo que el intervalo modal será: (125, 175]

Entonces, el valor de la moda, con la fórmula aproximada, será:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot a_i = 125 + \frac{50}{25 + 50} \cdot 50 = 158.33 \quad u.m.$$

Esto implica que el salario que más se repite en la empresa es de 158.33 unidades monetarias.

EJEMPLO

Distinta amplitud

Los salarios mensuales de 100 trabajadores de un hotel se reflejan en la siguiente distribución de frecuencias:

C_i	n_i
[75-200)	50
[200-250)	40
[250-300)	7
[300-400)	3
Σ	100

Observa que los intervalos no tienen una amplitud constante, por lo que para obtener la moda, debemos calcular la densidad de frecuencias, esto es:

$$h_i = \frac{n_i}{a_i}$$

Se añade la altura a la tabla de frecuencias:

C_i	n_i	h_i
[75-200)	50	50/125 = 0,40
[200-250)	40	40/50 =0,80
[250-300)	7	7/50 = 0,14
[300-400)	3	3/100 = 0,03
Σ	100	

La mayor densidad de frecuencias es hi=0.80, por lo que el intervalo modal será [200-250) y entonces la moda será:

$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i = 200 + \frac{0.14}{0.40 + 0.14} \cdot 50 = 213 \quad u.m.$$

En el hotel, el salario que más se repite es de 213 unidades monetarias.

2.2.2. MEDIDAS DE DISPERSIÓN

Las medidas de dispersión nos informan sobre **cuánto** se alejan del centro los valores de la distribución.

Rango o Recorrido

Es la diferencia entre el mayor y el menor de los datos de una distribución estadística. Se representa por *R*.

Su fórmula es:

$$R = X_{max} - X_{min}$$

EJEMPLO

Halla el rango si el peso de 11 alumnos/as es: 74, 67, 72, 41, 60, 66, 44, 75, 42, 79 y 45 (expresando en Kg.).

Máximo (Dato mayor) = 79 y Mínimo (dato menor) = 41 → *Rango* = 79– 41 = 38

Varianza

Es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística. La varianza se representa por la notación griega σ^2 o se opta por una notación más sencilla como *V* o *Var*.

A. Cálculo de la varianza para **datos sin agrupar.**

Su fórmula tiene varias formas de representarse:

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot f_i \quad \text{o bien} \quad \sigma^2 = \sum_{i=1}^{N} (x_i^2 \cdot f_i) - \bar{x}^2 \quad \text{o bien} \quad \sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N}$$

Te animamos a leer el siguiente documento donde se explica la equivalencia entre las igualdades de la varianza:

Demostración Equivalencia Igualdades de la Varianza

EJEMPLO

Calcula la varianza de la distribución: 9, 3, 8, 8, 9, 8, 9 y 18.

Para el cálculo de la varianza necesitas saber previamente el valor de la media aritmética, por tanto:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{N} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{N}$$

$$= \frac{(9-9)^2 + (3-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (18-9)^2}{8} = 15$$

B. Cálculo de la varianza para **datos agrupados.**

Varias fórmulas o caminos para hallar el valor.

A continuación, mostramos las dos más frecuentes a la hora de conocer el dato específico:

$$\sigma^2 = \sum_{i=1}^{N} \frac{x_i^2 \cdot n_i}{N} - \bar{x}^2 \quad \text{o bien} \quad \sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2 \cdot n_i}{N}$$

EJEMPLO

Con los siguientes datos agrupados, calculamos la varianza de la distribución de la tablaque se muestra a continuación:

Intervalo	n_i
[10, 20)	1
[20, 30)	8
[30,40)	10
[40, 50)	9
[50, 60	8
[60,70)	4
[70, 80)	2

Podemos calcular la varianza o bien aplicando la fórmula, o bien con la tabla de frecuencias. En esta ocasión, hallaremos el valor de la segunda manera:

Intervalo	x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[10, 20)	15	1	15	225
[20, 30)	25	8	200	5000
[30,40)	35	10	350	12250
[40, 50)	45	9	405	18225
[50, 60	55	8	440	24200
[60,70)	65	4	260	16900
[70, 80)	75	2	150	11250
Σ		42	1820	88050

Ahora, podemos aplicar la fórmula de manera más simplificada, tal y como mostramos a continuación:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \cdot n_i = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{N} = \frac{1820}{42} = 43.33$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i^2 \cdot n_i)}{N} - \bar{x}^2 = \frac{\sum_{i=1}^{N} x_i^2 \cdot n_i}{N} - \bar{x}^2 = \frac{88050}{42} - 43.33^2 = 218.94$$

Propiedades de la varianza

1. En el caso de que las puntuaciones sean iguales, la varianza será siempre un valor positivo o cero.

2. Si a todos los valores de la variable se les suma un número, la varianza no varía.

3. Si todos los valores de la variable se multiplican por un número, la varianza queda multiplicada por el cuadrado de dicho número.

4. Si tenemos varias distribuciones con la misma media y conocemos sus respectivas varianzas, se puede calcular la varianza total.

   - Si todas las muestras tienen el mismo tamaño:

$$\sigma^2 = \frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{n}$$

   - Si todas las muestras tienen distinto tamaño:

$$\sigma^2 = \frac{k_1\sigma_1^2 + k_2\sigma_2^2 + \ldots + k_n\sigma_n^2}{k_1 + k_2 + \ldots + k_n}$$

RECUERDA

1. La varianza, al igual que la media, es un índice **muy sensible a las puntuaciones extremas.**
2. En los casos que **no se pueda hallar la media, tampoco será posible hallar la varianza.**
3. **La varianza no viene expresada en las mismas unidades que los datos,** ya que las desviaciones están elevadas al cuadrado.

Desviación típica

También llamada **desviación estándar,** es la raíz cuadrada positiva de la varianza. La desviación típica se representa por σ o bien por su notación más simple *Dt.*

Tiene, también, fórmulas para datos sin agrupar y fórmulas para datos agrupados, aunque lo habitual es calcular previamente la varianza y, a partir de esta, hallar la desviación típica.

Su forma de cálculo queda muy simplificada:

$$\sigma = \sqrt{\sigma^2}$$

Veamos, ahora, los ejemplos vistos antes en la varianza, pero añadiendo el cálculo de la desviación típica.

EJEMPLO

En el **primer ejemplo anterior para datos sin agrupar** se tenía que la varianza era:

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N} = 15$$

Entonces, la desviación típica quedaría de la siguiente forma:

$$\sigma = \sqrt{\sigma^2} = \sqrt{15} = 3.87$$

En el **segundo ejemplo anterior para datos sin agrupar** se tenía que la varianza era:

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i^2 \cdot n_i)}{N} - \bar{x}^2 = 218.94$$

Luego la desviación típica correspondiente será:

$$\sigma = \sqrt{\sigma^2} = \sqrt{218.94} = 14.797$$

Propiedades de la desviación típica

1. En el caso de que las puntuaciones sean iguales, la desviación típica será siempre un valor positivo o cero.
2. Si a todos los valores de la variable se les suma un número, la desviación típica no varía.
3. Si todos los valores de la variable se multiplican por un número, la desviación típica queda multiplicada por dicho número.
4. Si tenemos varias distribuciones con la misma media y conocemos sus respectivas desviaciones típicas se puede calcular la desviación típica total.
   - Si todas las muestras tienen el mismo tamaño:

$$\sigma = \frac{\sigma_1 + \sigma_2 + \ldots + \sigma_n}{n}$$

   - Si todas las muestras tienen el distinto tamaño:

$$\sigma = \frac{k_1\sigma_1 + k_2\sigma_2 + \ldots + k_n\sigma_n}{k_1 + k_2 + \ldots + k_n}$$

RECUERDA

1. La desviación típica, al igual que la media y la varianza, es un índice **muy sensible a las puntuaciones extremas.**
2. En los casos que **no se pueda hallar la media, tampoco será posible hallar la desviación típica.**
3. **La desviación típica es la raíz cuadrada de la varianza.**

4. Cuanto **más pequeña sea la desviación típica, mayor será la concentración** de datos alrededor de la media.

5. Se llama **tipificación** de una variable al proceso de restar la media y dividir por su desviación típica. De esta manera, las observaciones tienen media igual a cero y la desviación típica igual a uno. Se representa por z_i. Cuya fórmula es:

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Cuasivarianza

La **cuasivarianza** de una población o muestra se obtiene al multiplicar la varianza por *N*/(N − 1). La **cuasivarianza muestral** es un estimador centrado (no sesgado) de la varianza poblacional. Se suele representar por:

$$\sigma^2_{N-1} \text{ o bien } S^2_{N-1}.$$

Su fórmula general se puede simplificar, por tanto se puede hallar de las siguientes dos formas:

$$\sigma^2_{N-1} = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2 \cdot n_i}{N-1} \quad \text{o bien} \quad \sigma^2_{N-1} = \frac{N}{N-1} \cdot \sigma^2$$

Es semejante a la Varianza, excepto que la división es por ***N-1*** (tamaño de la muestra) y no por ***N*** (tamaño del grupo de datos). Este estadístico es **apropiado para obtener estimaciones de la varianza** de la población en el análisis inferencial de datos.

SABÍAS QUE...

Te recomendamos la lectura de este paper para profundizar en el conocimiento de **su utilidad:**

El porqué de la cuasivarianza

También, recomendamos la lectura de este otro paper para entender el **desarrollo matemático** de la simplificación de la fórmula:

Desarrollo fórmula simplificada

Cuasidesviación típica

Se trata de un estimador centrado (no sesgado) de la varianza poblacional. Se calcula como la raíz cuadrada de la cuasivarianza, y se denotará por σ_{N-1} o bien S_{N-1}.

Tiene una fórmula general y otra fórmula simplificada, por tanto, la podemos hallar de ambas formas, esto es:

$$\sigma_{N-1} = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2 \cdot n_i}{N-1}} \quad \text{o bien} \quad \sigma_{N-1} = \sqrt{\frac{N}{N-1} \cdot \sigma^2}$$

Al igual que antes, normalmente, la hallaremos de la manera más simple, es decir, como la raíz cuadrada de la cuasivarianza, que expresado matemáticamente es de la forma siguiente:

$$\sigma_{N-1} = \sqrt{\sigma^2_{N-1}}$$

Desviación absoluta con respecto a la mediana

Es la suma de los valores absolutos de las diferencias de los valores de la variable a la mediana. Se denotará por D_{Me}.

Las fórmulas equivalentes para poder hallarla son:

$$D_{Me} = \sum_{i=1}^{N} |x_i - Me| \cdot f_i \quad \text{o bien} \quad D_{Me} = \sum_{i=1}^{N} \frac{|x_i - Me| \cdot n_i}{N}$$

RECUERDA

La desviación absoluta con respecto a la mediana se usa en lugar de la desviación estándar **cuando es necesario que los valores extremos afecten menos** al valor de la desviación. Esto se debe al hecho de que los valores extremos afectan menos a la mediana que a la media.

También hay una **desviación absoluta con respecto a la media,** en el que se sustituye la mediana por la media y se representa por Dx^- . En este aspecto no vamos a profundizar aunque dejamos el siguiente enlace ya que leyéndolo podrás ahondar en un mayor conocimiento:

Desviación media

La desviación absoluta con respecto a la mediana, siempre cumple la siguiente propiedad:

$$D_{Me} < D_{\bar{x}} < \sigma$$

SABÍAS QUE...

Para ampliar información sobre la desviación media, puedes leer:

Desviación respecto a la media y desviación media

EJEMPLO

Calcula la desviación media de la siguiente distribución de frecuencias:

x_i	n_i
0	2
3	4
5	6
7	5
9	3
Σ	20

El primer cálculo consiste en hallar la mediana, como ya está ordenado en la tabla y es un valor impar, entonces puedes apreciar que: *Me* = 5.

Ahora ya puedes hallar la desviación media:

$$Dm = \sum_{i=1}^{N} \frac{|x_i - Me| \cdot n_i}{N} = |0-5|\frac{2}{20} + |3-5|\frac{4}{20} + |5-5|\frac{6}{20} + |7-5|\frac{5}{20} + |9-5|\frac{3}{20} = 2$$

Coeficiente de Variación de Spearman

También llamado **Coeficiente de Variación o *CV:*** es la relación entre la desviación típica y su media. Esto es, es una medida estadística que **ofrece información respecto de la dispersión** relativa a un conjunto de datos.

Su fórmula es:

$$CV = \frac{\sigma}{\bar{x}} \quad \text{suponiendo que} \quad \bar{x} \neq 0$$

Su **interpretación** es relativa al **grado de variabilidad, independiente de la escala de la variable,** a diferencia de la desviación típica o estándar. Se puede interpretar el CV de la siguiente manera:

- A mayor valor del *CV,* mayor heterogeneidad de los valores de la variable.
- A menor valor del *CV,* mayor homogeneidad en los valores de la variable.

Por ejemplo, si el C.V es menor o igual al 80%, significa que la media aritmética es representativa del conjunto de datos y, por tanto, el conjunto de datos es homogéneo. Por el contrario, si el C.V supera al 80%, el promedio no será representativo del conjunto de datos y será heterogéneo.

Una de las utilidades del coeficiente de variación es que sirve como un indicador que **permite establecer comparaciones entre distintos casos o poblaciones,** además de poder establecer una relación entre el tamaño de la media aritmética y la variabilidad de la variable, pero ¡ojo! hay que tener cuidado a la hora de hacer comparativas.

Veamos el siguiente ejemplo para comprenderlo mejor.

EJEMPLO

Se han recogido el número de clientes que han recibido dos sucursales de un mismo banco. La sucursal situada en Madrid tiene una media de 140 visitantes y una desviación típica de 28.28. En cambio la sucursal situada en Barcelona tiene una media de 150 visitas y una desviación típica de 25. ¿Cuál de las dos presenta menor dispersión?

$$CV_{Madrid} = \frac{\sigma}{\bar{x}} = \frac{28.28}{140} = 0.2020 \quad \rightarrow \quad CV_{Madrid} = 20.2\%$$

$$CV_{Barcelona} = \frac{\sigma}{\bar{x}} = \frac{25}{150} = 0.1667 \quad \rightarrow \quad CV_{Barcelona} = 16.67\%$$

Así, en la población de Barcelona el coeficiente de variación es de un 16%, mientras que en la población de Madrid el *CV* es de un 20%. De acuerdo con estos datos, la población con mayor dispersión es la de Madrid.

En general, a igualdad de medias, la homogeneidad o heterogeneidad dependerá de la desviación típica. Como en este caso, ambas ciudades presentan una media similar y es la desviación típica la que afecta al coeficiente de variación.

Propiedades del coeficiente de correlación lineal de Pearson

1. El coeficiente de variación no tiene unidades. El *CV* es **invariante frente a cambios de escala.**

2. El coeficiente de variación se expresa en porcentaje, pues es como mejor se expresa. Aunque también podemos encontrarlo en cifras de 0 a 1, si bien es cierto que, en ciertas distribuciones de probabilidad, este coeficiente puede ser 1 o incluso mayor que 1. En general, el *CV* se suele **expresar en porcentajes.**

3. El coeficiente de variación **depende de la desviación típica y de la media aritmética.**

4. Permite **comparar las dispersiones de dos distribuciones distintas** (variables que aparecen en unidades distintas o que toman magnitudes muy diferentes o para comparar las variabilidades de varios conjuntos de datos, muestras o poblaciones), siempre que sus medias sean positivas y distintas de 0.

5. Se calcula para cada una de las distribuciones, y los valores que se obtienen se comparan entre sí. **La mayor dispersión corresponderá al valor del coeficiente de variación mayor.**

RECUERDA

- Para su interpretación se puede expresar como porcentaje, teniendo en cuenta que **puede superar el valor 100%.**
- Depende de la desviación típica y, en mayor medida, de **la media aritmética,** dado que **cuando esta es 0 o muy próxima a este valor, el C.V. pierde significado,** ya que puede dar valores muy grandes que no necesariamente implican una gran dispersión de datos.
- Permite establecer la relación entre el tamaño de la media y la variabilidad de la variable.
- Permite establecer comparaciones entre distintas muestras que estén midiendo lo mismo.

2.2.3. MEDIDAS DE POSICIÓN

Las medidas de posición nos informan de **dónde** se alejan del centro los valores de la distribución. Para hacerlo, dividen un conjunto de datos en grupos con el mismo número de individuos. Para calcular las medidas de posición es necesario que los datos estén ordenados de menor a mayor. Las medidas de posición son:

Cuartiles

Son los que **dividen la serie de datos en cuatro partes iguales.** Es decir, son los tres valores de la variable que dividen un conjunto de datos ordenados en cuatro partes iguales. Determinan los valores correspondientes al 25%, al 50% y al 75% de los datos. Hay que fijarse en que **el cuartil segundo coincide con la mediana.**

Se representan por:

$$Q_k \ \ con \ \ k = 1,\ 2 \text{ ó } 3$$

A. Cálculo de los cuartiles para **datos sin agrupar.**

En este caso no es necesario aplicar ninguna fórmula, basta con seguir una serie de pasos:

1. Primero ordenamos los datos de menor a mayor.

2. Después, buscamos el lugar que ocupa cada cuartil mediante la siguiente expresión:

$$CV = \frac{kN}{4} \quad \text{siendo} \quad k = 1, 2, 3$$

Para los cálculos, hay que tener en cuenta si la **distribución es par o impar:**

- Número impar de datos: 2, 5, 3, 6, 7, 4 y 9 → La distribución ordenada es: 2, **3** , 4, **5**, 6, **7**, 9

Luego los cuartiles son:

$$Q_1 = 3 \quad Q_2 = 5 \quad Q_3 = 7$$

- Número par de datos: 2, 5, 3, 4, 6, 7, 1 y 9 → La distribución ordenada es: 1, **2**, **3**, **4**, **5**, **6**, **7**, 9

Luego, los cuartiles son:

$$Q_1 = \frac{2+3}{2} = 2.5 \quad Q_2 = \frac{4+5}{2} = 4.5 \quad Q_3 = \frac{6+7}{2} = 6.5$$

B. Cálculo de la media aritmética para **datos agrupados.**

En este caso, se cuenta con una fórmula para poder hallarlos, que es:

$$Q_k = L_{i-1} + \frac{\frac{k \cdot N}{4} - N_{i-1}}{n_i} \cdot a_i \quad \text{donde} \quad k = 1, 2, 3$$

EJEMPLO

Calcula los cuartiles de la distribución de la tabla siguiente:

Intervalo	n_i	N_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	Σ 65	

En general los intervalos suelen estar ordenados, se divide la recta en 4 partes iguales para, posteriormente, analizar los resultados para ver en qué parte de la recta está la frecuencia absoluta.

Una vez elegido el intervalo se podrá aplicar la fórmula general, tal y como se muestra a continuación:

$$Q_1 = \frac{k \cdot N}{4} = \frac{1 \cdot 65}{4} = 16.25 \quad \rightarrow \quad Q_1 = L_{i-1} + \frac{\frac{k \cdot N}{4} - N_{i-1}}{n_i} \cdot a_i = 60 + \frac{16.25 - 8}{10} \cdot 10 = 68.25$$

$$Q_2 = \frac{k \cdot N}{4} = \frac{2 \cdot 65}{4} = 32.5 \quad \rightarrow \quad Q_2 = L_{i-1} + \frac{\frac{k \cdot N}{4} - N_{i-1}}{n_i} \cdot a_i = 70 + \frac{32.5 - 18}{16} \cdot 10 = 79.0625$$

$$Q_3 = \frac{k \cdot N}{4} = \frac{3 \cdot 65}{4} = 48.75 \quad \rightarrow \quad Q_3 = L_{i-1} + \frac{\frac{k \cdot N}{4} - N_{i-1}}{n_i} \cdot a_i = 90 + \frac{48.75 - 48}{10} \cdot 10 = 90.75$$

Rango intercuartílico

Es la diferencia entre el tercer y el primer cuartil. Matemáticamente se representa por: RQ. En ocasiones se denota por su notación inglesa *(IQR).*

Su fórmula es:

$$R_Q = Q_3 - Q_1$$

EJEMPLO

Continuando con el ejemplo anterior, el rango intercuartílico se hallaría de la siguiente forma:

$$R_Q = Q_3 - Q_1 = 90.75 - 68.25 = 22.5$$

Deciles

Son los que **dividen la serie de datos en diez partes iguales.** Es decir, son los nueve valores que dividen la serie de datos en diez partes iguales. Los deciles dan los valores correspondientes al 10%, al 20%... y al 90% de los datos. **El decil quinto coincide con la mediana.** Se representan por D_k.

Al igual que antes, primero buscamos la clase modal donde se encuentra *(kN)*/ 10 en la tabla de las frecuencias acumuladas para, inmediatamente después, aplicar la siguiente fórmula:

$$D_k = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i \quad \text{donde} \quad k = 1, 2, 3, \ldots 9$$

EJEMPLO

Calcula los deciles de la distribución de la tabla vista anteriormente:

Intervalo	n_i	N_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
Σ	65	

Primero, analizaremos dónde está la clase modal para cada corte y, luego, aplicaremos la fórmula vista.

Por tanto, los deciles son:

$$D_1 = \frac{k \cdot N}{10} = \frac{1 \cdot 65}{10} = 6.5 \quad \rightarrow \quad D_1 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 50 + \frac{6.5 - 0}{8} \cdot 10 = 58.12$$

$$D_2 = \frac{k \cdot N}{10} = \frac{2 \cdot 65}{10} = 13 \quad \rightarrow \quad Q_2 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 60 + \frac{13 - 8}{10} \cdot 10 = 65$$

$$D_3 = \frac{k \cdot N}{10} = \frac{3 \cdot 65}{10} = 19.5 \quad \rightarrow \quad D_3 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 70 + \frac{19.5 - 18}{16} \cdot 10 = 70.94$$

$$D_4 = \frac{k \cdot N}{10} = \frac{4 \cdot 65}{10} = 26 \quad \rightarrow \quad D_4 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 70 + \frac{26 - 18}{16} \cdot 10 = 75$$

$$D_5 = \frac{k \cdot N}{10} = \frac{5 \cdot 65}{10} = 32.5 \quad \rightarrow \quad D_5 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 70 + \frac{32.5 - 18}{16} \cdot 10 = 79.06$$

$$D_6 = \frac{k \cdot N}{10} = \frac{6 \cdot 65}{10} = 39 \quad \rightarrow \quad D_6 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 80 + \frac{39 - 34}{14} \cdot 10 = 83.57$$

$$D_7 = \frac{k \cdot N}{10} = \frac{7 \cdot 65}{10} = 45.5 \quad \rightarrow \quad D_7 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 80 + \frac{45.5 - 34}{14} \cdot 10 = 88.21$$

$$D_8 = \frac{k \cdot N}{10} = \frac{8 \cdot 65}{10} = 52 \quad \rightarrow \quad D_8 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 90 + \frac{52 - 48}{10} \cdot 10 = 94$$

$$D_9 = \frac{k \cdot N}{10} = \frac{9 \cdot 65}{10} = 58.5 \quad \rightarrow \quad D_9 = L_{i-1} + \frac{\frac{k \cdot N}{10} - N_{i-1}}{n_i} \cdot a_i = 100 + \frac{58.5 - 58}{5} \cdot 10 = 101$$

Percentiles

Son los que **dividen la serie de datos en cien partes iguales.** Es decir, son los 99 valores que dividen la serie de datos en 100 partes iguales. Los percentiles dan los valores correspondientes al 1%, al 2%...

y al 99% de los datos. Vemos que, con esta definición, **el percentil 50 coincide con la mediana.** Se representan por P_k.

Al igual que antes, primero se busca la clase modal donde se encuentra *(kN)*/10 en la tabla de las frecuencias acumuladas. Posteriormente, aplicaremos la expresión matemática formulada a continuación:

$$P_k = L_{i-1} + \frac{\frac{k \cdot N}{100} - N_{i-1}}{n_i} \cdot a_i \quad \text{donde} \quad k = 1, 2, 3, \ldots 99$$

En el siguiente gráfico, que analiza la altura, observa cómo la mediana dibujada con la línea verde coincide con el **percentil 50.** Además, el Q_1 coincide con el P_{25} y Q_3 coincide con P_{75}.

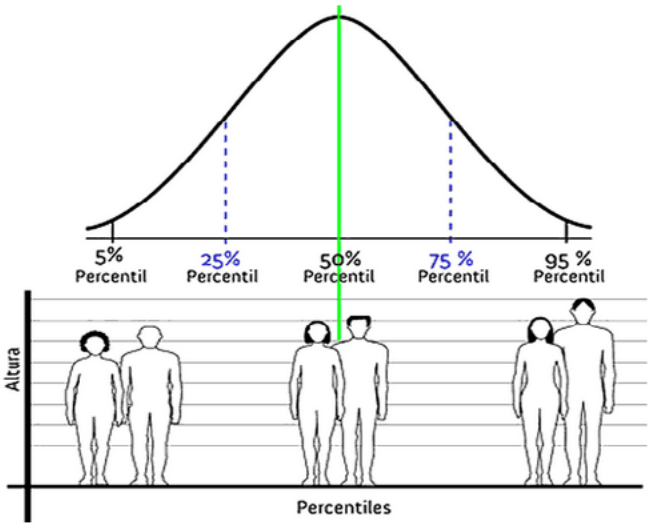


Imagen 5. Percentiles
Fuente de la imagen: *www.curiosoando.com*

EJEMPLO

Calcula el percentil 35 y 60 de la distribución de la misma tabla:

Intervalo	n_i	N_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
Σ	65	

Primero, analizaremos dónde está la clase modal para cada corte y, luego, aplicaremos la fórmula vista. Por tanto, los percentiles serán:

$$P_{35} = \frac{k \cdot N}{100} = \frac{35 \cdot 65}{100} = 22.75 \quad \rightarrow \quad P_{35} = L_{i-1} + \frac{\frac{k \cdot N}{100} - N_{i-1}}{n_i} \cdot a_i = 70 + \frac{22.75 - 18}{16} \cdot 10 = 72.97$$

$$P_{60} = \frac{k \cdot N}{100} = \frac{60 \cdot 65}{100} = 39 \quad \rightarrow \quad P_{60} = L_{i-1} + \frac{\frac{k \cdot N}{100} - N_{i-1}}{n_i} \cdot a_i = 80 + \frac{39 - 34}{14} \cdot 10 = 83.57$$

RECUERDA

- La mediana es igual a Q_2, D_5 y P_{50}. Se cumple siempre la igualdad $Me = Q_2 = D_5 = P_{50}$.
- De las medidas de posición, probablemente los percentiles son los estadísticos más usados ya que permiten tener una idea más pormenorizada de dónde están distribuidos los datos al dividirse más. Muy útiles y usados en medicina (en especial en pediatría para analizar el desarrollo de bebes y niñ@s). Por tanto, si tenemos que comunicarnos con un público no experto recomendamos el uso de percentiles en lugar de deciles o cuartiles.

2.2.4. MEDIDAS DE CONCENTRACIÓN Y FORMA

También llamadas medidas de distribución. Estas medidas nos informan sobre **cómo** se alejan del centro los valores de la distribución. Cada medida de concentración genera una "forma" característica de la distribución.

Asimetría o Sesgo

Este estadístico indica **cómo de simétrica es la gráfica.** Se halla a través del Coeficiente de Asimetría de Fisher **(CAF)** también se puede **ver la forma que tiene la distribución gráficamente,** observando el comportamiento que tiene el eje de la distribución con respecto a la media aritmética. Esto es, evaluando la proximidad o lejanía de los datos con respecto a la media.

Su fórmula es:

$$CA_F = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N \cdot \sigma^3}$$

Cuando los datos están agrupados, hay que añadir n_i a la fórmula, quedando de la siguiente manera:

$$CA_F = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3 \cdot n_i}{N \cdot \sigma^3}$$

La distribución puede ser:

- **Asimétrica negativa** o a la izqda. (coeficiente negativo o $CA_F < 0$).
- **Simétrica** (coeficiente de Fisher igual a cero o $CA_F = 0$).
- **Asimétrica positiva** o a la dcha. (coeficiente positivo o $CA_F > 0$).

Gráficamente la distribución presenta las siguientes formas:

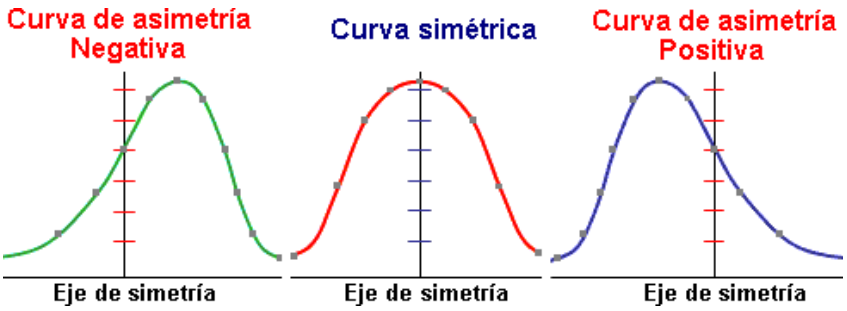


Imagen 6. Asimetría o Sesgo
Fuente de la imagen: *www.spssfree.com*

Además del *CAF*, que es el estimador común para el cálculo de la concentración y forma de la distribución, existen otros coeficientes que son:

- El Coeficiente de asimetría de Bowley.
- El Coeficiente de asimetría de Pearson.

PARA SABER MÁS

Para ahondar en estos dos últimos coeficientes, además de poder compararlo con el de **Fisher** os proporcionamos el siguiente enlace:

Coeficiente de asimetría de Pearson y Bowley

Curtosis o Apuntamiento

Este estadístico indica **cómo están de concentrados los valores en la gráfica.** En él se estudia la distribución de frecuencias en la zona central. El apuntamiento se mide respecto de la campana de Gauss, que es unimodal y simétrica. Dicho de otra manera, **el apuntamiento se compara con la campana de Gauss,** por tanto se compara con una distribución que cumple la propiedad: *Media = Mediana = Moda.*

La curtosis se puede hallar a través de la siguiente expresión:

$$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N \cdot \sigma^4} - 3$$

Al igual que antes, cuando los datos están agrupados, hay que añadir n_i a la fórmula, quedando como mostramos a continuación:

$$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4 \cdot n_i}{N \cdot \sigma^4} - 3$$

La distribución puede ser de 3 tipos:

- **Platicúrtica** (menos concentración en los valores centrales de la variable o Curtosis < 0.
- **Mesocúrtica** (igual concentración o Curtosis = 0.
- **Leptocúrtica** (más concentración en los valores centrales o Curtosis > 0.

Gráficamente sería:

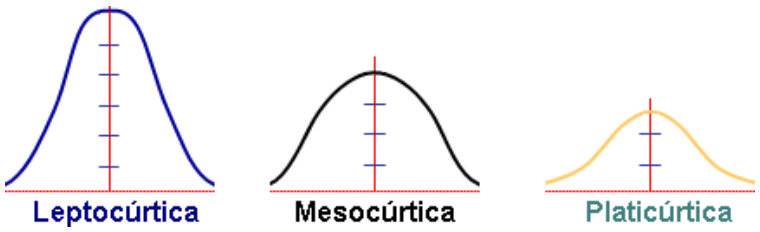


Imagen 7. Curtosis o Apuntamiento
Fuente de la imagen: *www.spssfree.com*

PARA SABER MÁS

Para complementar y profundizar en las medidas de distribución puedes acceder al siguiente paper:

Medidas de asimetría y curtosis

2.2.5. MÉTODOS GRÁFICOS

Para analizar las medidas de centralización, dispersión y posición hay un gráfico con el que se recogen casi todas las medidas vistas y ofrece una visión muy completa de un análisis estadístico unidimensional.

Box-Plot

El Box-plot o *Diagrama de Cajas y Bigotes* o simplemente *Diagrama de Caja* es un gráfico en el cual se recogen las medidas de dispersión y de centralización.

El diagrama de cajas es un gráfico basado en los cuartiles, contiene información sobre la simetría de la

distribución y permite definir la idea de un dato atípico.

Se representan en él y por este orden:

- $X_{Mín}$
- Q_1, Q_2, Q_3
- X_{Max}

Esto es:

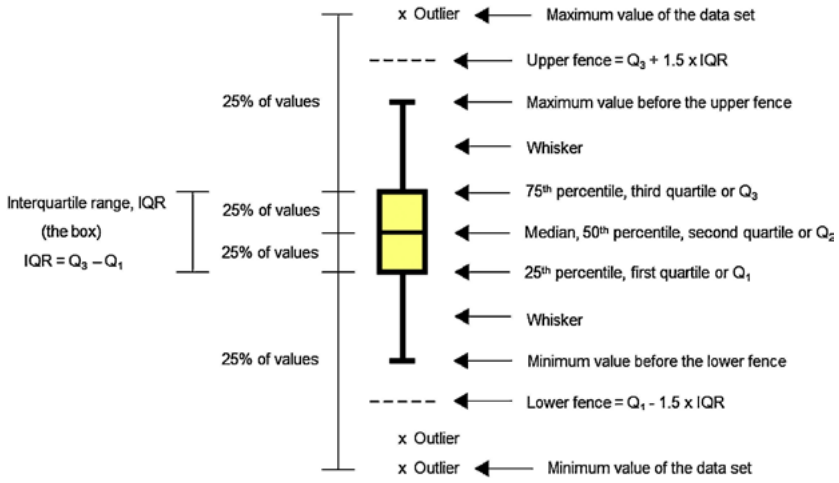


Imagen 8. Box-Plot
Fuente de la imagen: *www.elsevier.es*

SABÍAS QUE...

El concepto de ***outlier*** o valor atípico que aparece en el gráfico es importante. Es un valor que no representa a la "mayoría de la distribución" y presenta una gran dispersión.

Como ofrece mucha información, en ocasiones, se simplifica el gráfico. Otra forma de verlo habitualmente es de la siguiente forma:

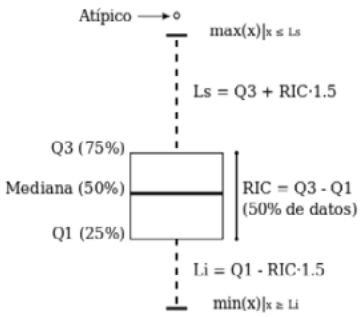


Imagen 9. Box-Plot (simplificado)
Fuente de la imagen: *www.cajaybigotes.blogspot.com*

Los diagramas de cajas y bigotes no solamente son usados para analizar una distribución aislada, también se usan para realizar **análisis comparativos** como mostramos en el siguiente gráfico:

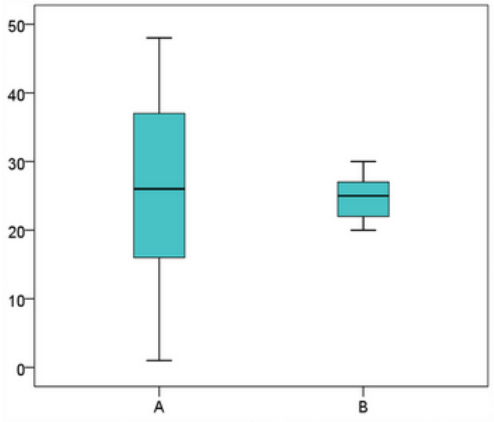


Imagen 10. Box-Plot (comparativa)
Fuente de la imagen: *www.analisisdedatos.org*

La otra información que podemos extraer de un box-plot es sobre la dispersión de las observaciones de una distribución. Por ejemplo, en la imagen observamos la diferencia entre una variable con una gran dispersión de puntuaciones (A) y una variable con una dispersión de puntuaciones muy pequeña (B).

Observaciones de todos los gráficos

Hay más gráficos que aquí no aparecen, ya que sólo hemos explicado **los gráficos** más comunes. Pero todos ellos deber seguir una "misma estructura" o lógica, que es la siguiente:

1. Pueden tener o no títulos, pero si se ponen han de ser claros, no muy largos y concisos.
2. Se pueden representar en **horizontal o en vertical** (barras, box-plot, etc.).
3. Se puede dibujar en **dos o tres dimensiones.**
4. Se pueden sustituir por dibujos y, entonces, se llaman **pictogramas** (si son mapas **cartogramas**).
5. Pueden ser "mixtos", es decir, se representa en un mismo gráfico **dos o más series de datos** (cada serie se representa con un color diferente, para identificarlo más fácilmente).
6. Pueden llamarse de otra manera. Para identificar si se trata de un gráfico visto aquí, debemos fijarnos en qué valores aparecen y cómo se están representado.

ACTIVIDAD

Titanic

El **objetivo** de esta actividad es identificar y analizar el total pagado de un dataset de facturas que tiene una compañía.

Se pide:

1. Hallar la media, la desviación típica, los 3 cuartiles, el mínimo y el máximo.
2. ¿Cuál es el billete que tiene como mínimo el 5% del importe pagado del total de todos viajeros?
3. Realiza un análisis de la distribución y añade un box-plot para una mejor evaluación.

Solución

In [ ]:

```
#Import libraries import pandas as pd import numpy
as np
import matplotlib.pyplot as plt

# Carga del fichero desde el enlace web y creación del dataframe
url_data = 'https://raw.githubusercontent.com/md-
    lorente/data/master/titanic.csv'

# Creación Dataframe
df = pd.read_csv(url_data, sep=',')

# Visualización del dataframe (lacabecera)
df.head()
```

Out[ ]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Recordemos que se trata de una variable cuantitativa de tipo continuo. Para el primer apartado basta con usar "describe" de *python* que tiene los principales estadísticos para "ubicar" la variable en la distribución.

In [ ]:

```
# Análisis rápido: resumen de estadística descriptiva
df["Fare"].describe()
```

Out[ ]:

```
count 891.000000 mean   32.204208 std    49.693429 min     0.000000
25%        7.910400
50%       14.454200
75%       31.000000 max  512.329200
Name: Fare, dtype: float64
```

¿Cuál es el billete que tiene como mínimo el 5% del importe pagado del total de todos viajeros?

En esta pregunta, nos están realmente preguntando por el percentil 95, por tanto:

In [ ]:

```
# Percentil 95
p95 =

df["Fare"].quantile(0.95)


print('p95:' , round(p95, 2))
```

Out[ ]:

```
p95: 112.08
```

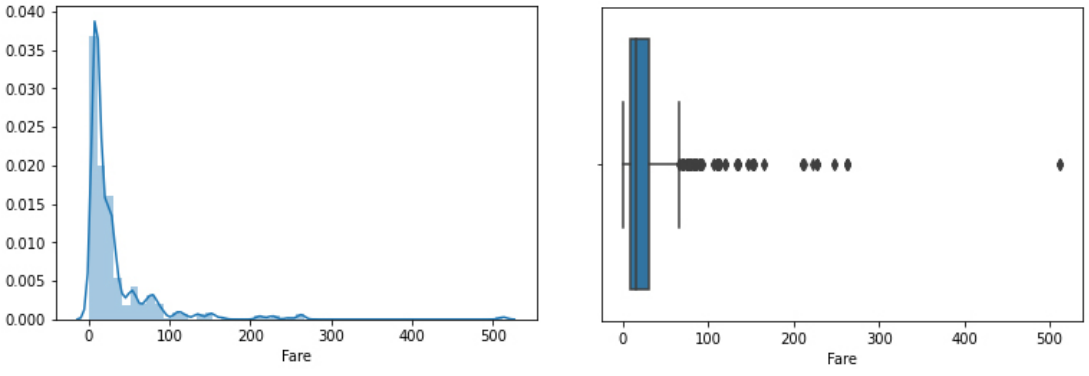
El resultado es que el 5% con el importe más elevado de los billetes fue de 112,08 unidades monetarias de todos los billetes pagados a bordo del Titanic.

In [ ]:

```
# Análisis de la distribución
import scipy.stats as ss

print("Asimetria: %f" % df["Fare"].skew())
print("Curtosis:  %f" % df["Fare"].kurt())


# Gráfico de la distribución sns.distplot(df["Fare"])
plt.show()

# Box-plot sns.boxplot(df["Fare"]) plt.show()
```

Out[ ]:

```
Asimetría: 4.787317
Curtosis:  33.398141
```



El resultado obtenido es una gráfica en la que se observa una curva de asimetría positiva y leptocúrtica.

En el gráfico de *boxplot* se observa que hay no hay mucha dispersión. Es especialmente carecterístico dentro de los outliers hay uno que es muy extremo.

IDEAS CLAVE

- Implica una recopilación de datos **teniendo como objetivo la inferencia o predicciones.**

- **La idea de resumir en unos pocos datos la información del comportamiento global del fenómeno** con las tablas estadísticas se recogen los estadísticos básicos de frecuencia absoluta, relativa y acumulada de ambas.

- Requiere una selección de un subconjunto de **una gran colección de datos,** con el propósito de hacer inferencias con respecto a las características del conjunto completo (población).

- Con las **medidas de centralización, dispersión, posición, concentración y forma se puede analizar cómo se comporta una observación en concreto y también cómo es la distribución de los datos.** Además se pueden detectar valores atípicos.