TEMA 4. DESCRIPCIÓN NUMÉRICA DE DATOS UNIVARIANTES: MEDIDAS DE POSICIÓN

Curso OCW de “Estadística Descriptiva con Excel para Grados de Ciencias Sociales”

Grupo Decisión Multicriterio Zaragoza (GDMZ)

1. DESCRIPCIÓN NÚMERICA DE DATOS UNIVARIANTES

La distribución de frecuencias de una variable nos ofrece toda la información posible, pero en muchas ocasiones es imposible interpretar toda esa extensa información. Por lo tanto, nos vemos obligados a resumirla en una serie de medidas descriptivas que nos permiten entender de forma rápida su contenido. Estas medidas nos permiten:

- Describir los aspectos notables de una distribución de frecuencias destacando algunas propiedades o características importantes de la variable analizada.
- Comparar diferentes conjuntos de datos o distribuciones mediante el análisis de las medidas numéricas adecuadas.

Para ello será necesario elegir y calcular las medidas adecuadas para resumir los aspectos notables de los datos, e interpretar correctamente los valores de dichas medidas y evaluar su representatividad.

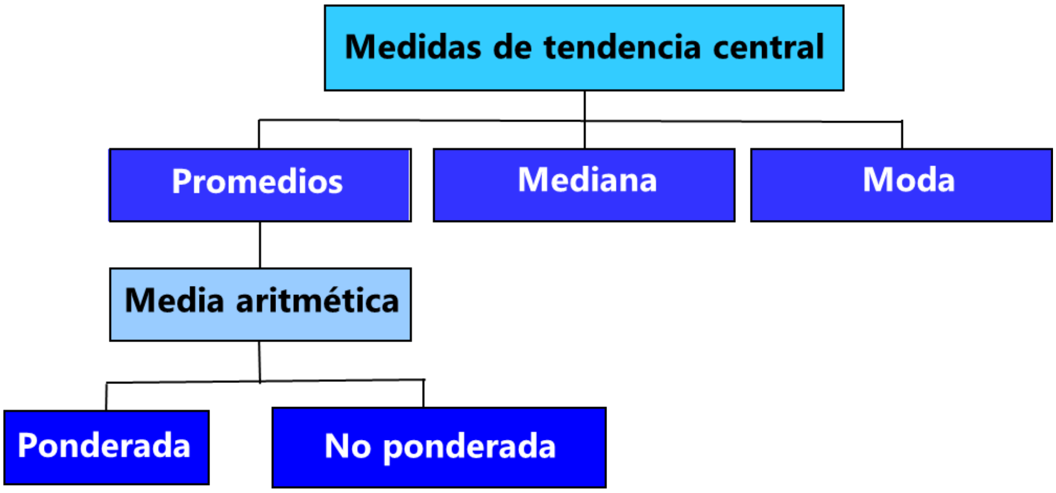
Este proceso de síntesis puede tener distintos objetivos de comprensión de la información, así pues, hablamos de medidas de posición, dispersión y forma.

Las medidas de posición nos permiten tener una idea clara de la situación de la variable en su escala de medida. Las medidas de dispersión tienen un doble fin, por un lado, nos permiten averiguar si las medidas de posición son representativas de la distribución y, por otro lado, nos dan una clara idea de la separación, variabilidad o dispersión de los distintos valores que toma la variable estadística. Las medidas de forma nos permiten conocer otras características más generales de la distribución y que están estrechamente ligadas a la forma que tendría la distribución si realizáramos una representación gráfica. Así pues, intentaremos conocer la posible simetría o asimetría de la distribución y su apuntamiento con respecto a un modelo de referencia. Dicha forma también se puede analizar, de forma gráfica, mediante los diagramas de caja que nos permiten, además, comparar distribuciones de frecuencias diferentes.

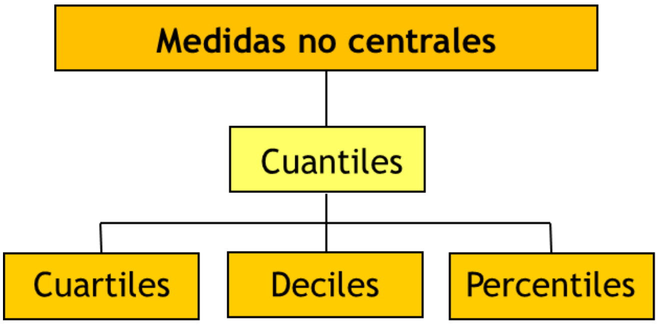
2. MEDIDAS DE POSICIÓN

Las medidas de posición indican un valor de la variable en torno al cual se sitúa un grupo de observaciones o datos. Sirven para estudiar las características de los valores que indican la posición en la que se sitúa un grupo de valores. Su objetivo es describir y sintetizar la información contenida en un conjunto de datos, usualmente con el propósito de compararlos con respecto de otros. Pueden clasificarse en las llamadas medidas de tendencia central y las medidas de tendencia no central. Una medida de tendencia central trata de ubicar el centro de la distribución que se puede considerar como representativo de la misma, mientras que una medida de tendencia no central permite conocer otros puntos característicos de la distribución que no son los valores centrales.

En lo que sigue, vamos a considerar las siguientes medidas de tendencia central[1]:



Y las siguientes medidas de tendencia no central:



2.1. Medidas de tendencia central

Una medida de tendencia central es un valor que intenta describir el conjunto de datos de una variable identificando la posición central dentro de dicho conjunto de datos. Las medidas de tendencia central más utilizadas son: media (ponderada o no ponderada), mediana y moda.

2.1.1. Moda

Se define la **moda** (**Mo**) como el valor o dato de la distribución que se observa en más individuos, que más se repite. Es decir, la moda corresponde a la modalidad cuya frecuencia es más alta.

[1] Dentro de los promedios, también podemos encontrar la media geométrica y la media armónica.

Su cálculo depende del tipo de variable analizada.

a) Para datos cualitativos o cuantitativos no agrupados. Se identifica la modalidad cuya frecuencia es mayor:

$$Mo = x_j \quad \text{si} \quad n_j = \max_i\{n_i\}$$

b) Para datos agrupados en intervalos, primero hay que determinar el **intervalo modal** que es aquél que tiene mayor densidad de frecuencia (recordar que la densidad de frecuencia de un intervalo viene dado por $d_i = \frac{n_i}{a_i}$).

Una vez obtenido este intervalo, hay que determinar el valor preciso de la moda, para lo cual se pueden utilizar diferentes criterios. Una posibilidad es aproximarlo por la expresión:

$$Mo = L_{m-1} + \frac{d_{m+1}}{d_{m-1} + d_{m+1}} \cdot a_m$$

O, de una forma más sencilla, quedarnos con la marca de clase del intervalo modal:

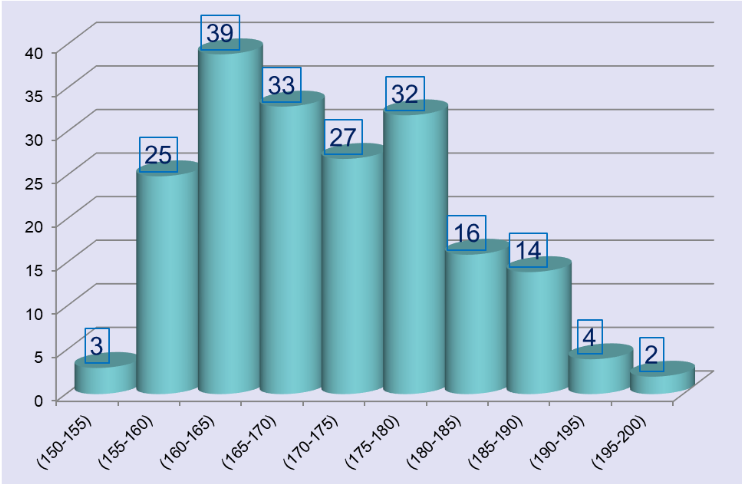
$$Mo = x_m = \frac{L_{m-1} + L_m}{2}$$

Propiedades:

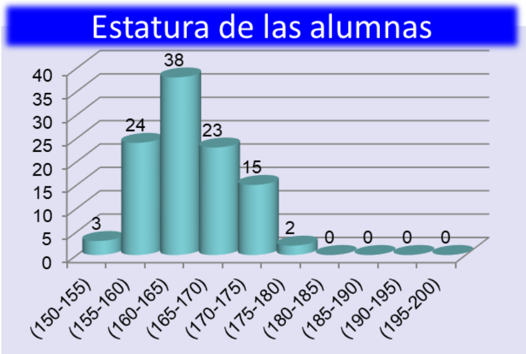
- Fácil de interpretar y sencilla de calcular.
- No se ve afectada por valores extremos
- Se puede calcular para cualquier tipo de variable.
- No es invariante frente a cambios de escala y origen. Formalmente, se cumple que si $Y = a + b \times X$, entonces $Mo(Y) = a + b \times Mo(X)$.

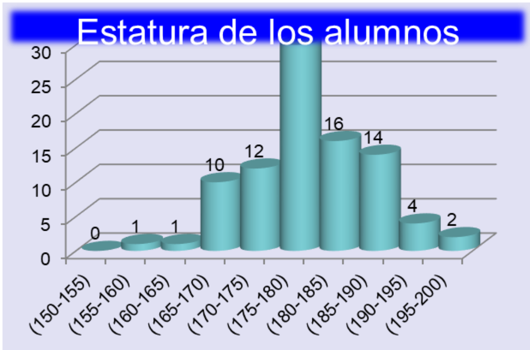
Pueden darse varias modas y, dependiendo del número de modas, las distribuciones se clasifican como **unimodales**, **bimodales** o **multimodales**. En caso de no ser única, la Moda pierde representatividad. La presencia de dos o más modas se debe, generalmente, a una mezcla de dos o más grupos heterogéneos de modo que conviene estudiar los grupos por separado. En el siguiente ejemplo observamos una distribución bimodal correspondiente a las estaturas de un conjunto de estudiantes:

(150-155)	3
(155-160)	25
(160-165)	39
(165-170)	33
(170-175)	27
(175-180)	32
(180-185)	16
(185-190)	14
(190-195)	4
(195-200)	2



Realizando el estudio por separado:





2.1.2. Media aritmética

La media aritmética se denota por $\overline{x}$ y se define como la suma de todos los valores de la distribución dividida por el número total de observaciones:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{k} x_i \times n_i = \sum_{i=1}^{k} x_i \times f_i$$

En el caso de variables continúas agrupadas en intervalos, su cálculo se aproxima utilizando las marcas de clase.

Propiedades:

- Es una medida única
- Su cálculo es sencillo y en él intervienen todos los valores de la distribución.
- No tiene sentido calcularla en variables cualitativas, excepto si son binarias codificadas como 0/1, en cuyo caso la media aritmética es una proporción
- No se puede calcular en distribuciones agrupadas que presentan algún intervalo de clase infinito

- La media aritmética es el centro de gravedad de la distribución, es decir, es el punto que establece la posición respecto de la cual los datos que se encontrarán por encima y por debajo de la media, compensan sus distancias (con signo) respecto de ésta. Formalmente: $\sum_{i=1}^{k}(x_i - \bar{x}) \times n_i = 0$
- No es robusta
- Es una medida descomponible. La media aritmética de todo un conjunto de datos se puede calcular promediando las medias aritméticas de los diferentes subconjuntos que constituyen una partición. Formalmente, para dos subconjuntos de tamaños N_A y N_B se expresa:

$$\bar{x}_T = \frac{N_A \times \bar{x}_A + N_B \times \bar{x}_B}{N_A + N_B}$$

- No es invariante frente a cambios de escala y origen. Formalmente, se cumple que si $Y = a + b \times X$, entonces $\bar{y} = a + b \times \bar{x}$

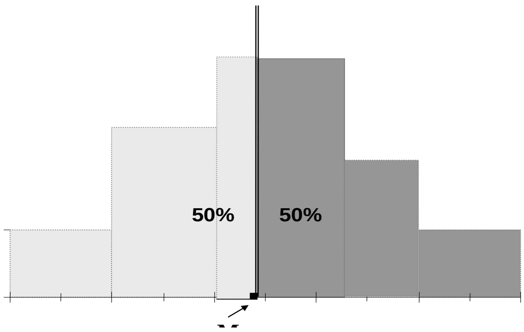
Las medias ponderadas se utilizan cuando no todos los valores de la variable tienen la misma importancia. La media aritmética ponderada, se calcula como:

$$\overline{x}_w = \frac{\sum_{i=1}^{k} x_i \times w_i}{\sum_{i=1}^{k} w_i}$$

Donde $\{w_i; i = 1, \ldots, k\}$ un conjunto de pesos asignando la importancia de cada dato.

2.1.3. Mediana

La **mediana** es el valor de la distribución, supuesta ordenada de menor a mayor, que deja a su izquierda y a su derecha el mismo número de datos, es decir, es el valor que ocupa el lugar central, supuesto un número impar de observaciones. En otras palabras, la mediana es el valor que **divide a la distribución de frecuencias en dos partes iguales**, dejando la mitad de los datos por debajo (y la otra mitad) por encima. Por ello, también puede definirse como el valor de la distribución cuya frecuencia acumulada es N/2 ó alternativamente, cuya frecuencia relativa acumulada es 50%.



Su cálculo depende del tipo de variable analizada.

a) En datos no agrupados –repetidos ó no– se ordenan de forma creciente y se identifica el valor o dato que ocupa la posición central. Formalmente, se detecta el entero m que cumple $m - 1 < N/2 \leq m$

- Si N es impar, $Me = x_m$, correspondiendo con el valor central
- Si N es par, puede observarse que hay dos valores centrales, x_m y x_{m+1} y, si la variable es cuantitativa, se tomará como mediana la media de ambos:

$$Me = \frac{x_m + x_{m+1}}{2}$$

Si N es par y la variable no es cuantitativa, se sigue la regla anterior.

b) Para datos agrupados en intervalos determinamos, en primer lugar, el **intervalo mediano**, que es aquél donde la frecuencia acumulada alcanza el 50%. Esto es:

$$N_{m-1} < \frac{N}{2} \leq N_m \ \text{ o } \ F_{m-1} < 0{,}5 \leq F_m$$

El valor concreto se puede determinar dentro de dicho intervalo por interpolación lineal:

$$Me = L_{m-1} + \frac{{}^{N}\!/_{2} - N_{m-1}}{n_m} \cdot a_m \ \ o \ \ Me = L_{m-1} + \frac{0.5 - F_{m-1}}{f_m} \cdot a_m$$

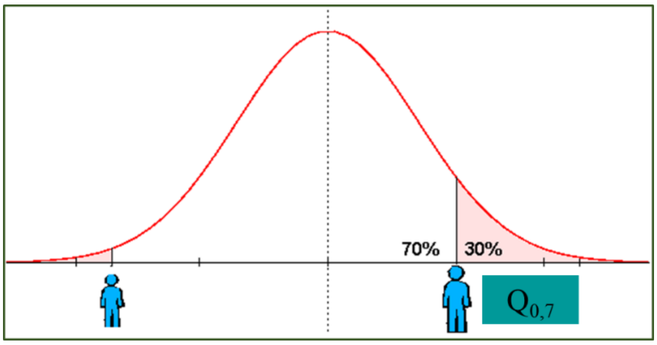
Propiedades:

- Es única, si bien puede aproximarse de formas diferentes.
- No utiliza todos los datos
- No tiene sentido para variables cualitativas con escala de medida nominal
- Es más robusta que la media
- No es invariante frente a cambios de escala y origen. Formalmente, se cumple que si $Y = a + b \times X$, entonces $Me(Y) = a + b \times Me(X)$

2.2. Medidas de tendencia no central

Las medidas de tendencia no central, denominadas **cuantiles**, son medidas que no van a reflejar ninguna tendencia de tipo central del conjunto de los datos. Extendiendo el concepto de la mediana, dividen a la distribución de frecuencias en varias partes, todas ellas con idéntica frecuencia; es decir dividen a la distribución en diversos intervalos que contienen todos ellos un mismo número de datos. De esta forma, reflejan los valores superiores, medios e inferiores.

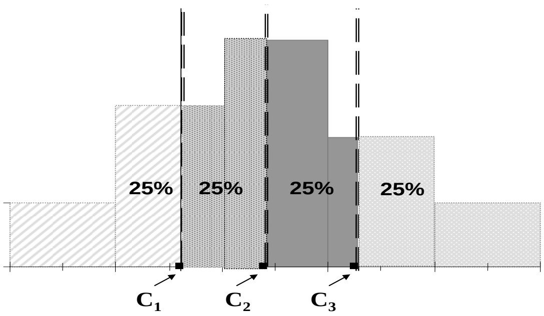
Se define **cuantil de orden *p*, *Q*$_p$**, como el valor de la variable que divide la distribución en dos partes cuyas frecuencias son ***p*** y ***1-p***, es decir, hay una frecuencia igual a *p* por debajo de él y una frecuencia *1-p* por encima de él. Es una extensión del concepto de mediana posibilitando que el porcentaje acumulado del 50% pueda ser cualquier otro valor *p*, tal que $0 < p < 1$.



Entre los **cuantiles** destacan, por ser de uso más frecuente, los llamados **cuartiles**, los **deciles** y los **percentiles**. Otros cuantiles menos empleados son los quintiles y octiles.

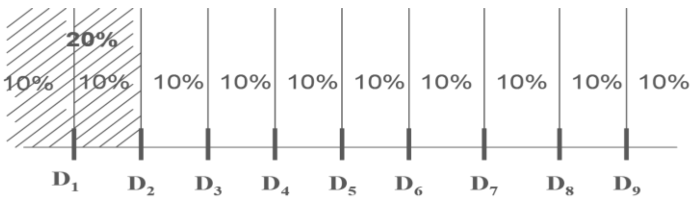
2.2.1. Cuartiles

Valores que dividen a la distribución de frecuencias en cuatro partes con idéntica frecuencia; es decir, en cuatro intervalos dentro de cada cual están incluidos el 25% de los valores. Los cuartiles son tres, y los denotamos por C_1, C_2, C_3



2.2.2. Deciles

Valores que dividen a la distribución de frecuencias en diez partes iguales, es decir, en diez intervalos dentro de cada uno de los cuales están incluidos el 10% de los de los datos o frecuencias. Son 9 los denotamos por $\{D_i; i = 1,...,9\}$



2.2.3. Percentiles

Son los valores de la distribución que la dividen en cien partes iguales, es decir, en cien intervalos dentro de cada cual están incluidos el 1% de los valores de la distribución. Los percentiles (P_i =1,...,99) son 99 valores que dividen la distribución en 100 partes iguales.

2.2.4 Cálculo de los cuantiles

Su identificación o cálculo, que difiere en función del tipo de datos, es esencialmente similar al caso de la mediana:

- Calcular las frecuencias acumuladas N_i o F_i
- Identificar $\boldsymbol{m}$ tal que $\boldsymbol{F_{m-1} < 100 \times p\% \leq F_m}$ o $\boldsymbol{N_{m-1} < N \times p \leq N_m}$
- Si la variable es discreta y **no agrupada** en intervalos:
    - Si $\boldsymbol{F_m > 100 \times p}\%$: $\boldsymbol{Q_p = x_m}$
    - Si $\boldsymbol{F_m = 100 \times p}\%$: $\boldsymbol{Q_p = \frac{x_m + x_{m+1}}{2}}$
- Si la variable es continua **agrupada** en intervalos:

$$\boldsymbol{Q_p = L_{m-1} + \frac{N \cdot p - N_{m-1}}{n_m} \cdot a_m} \quad \text{o} \quad \boldsymbol{Q_p = L_{m-1} + \frac{p - F_{m-1}}{f_m} \cdot a_m}$$