LECCIONES DE ESTADÍSTICA DESCRIPTIVA

4.1 Introducción.

Al construir la distribución de frecuencias de una determinada característica, se intenta representar mediante una tabla el numeroso conjunto de datos que originalmente se disponía. Este proceso de reducción puede continuarse tratando de buscar unas medidas, denominadas de ***posición***, o tendencia, que muestren, mediante algún criterio, cómo es la distribución de frecuencias.

Una primera clasificación de estas medidas sería en ***centrales***, si de alguna forma buscan el "centro" de la distribución de frecuencias, y ***no centrales***, en el caso de que sea otro el criterio que guía su obtención. En adelante se va a suponer que la característica objeto de estudio es cuantitativa, dejando para el último epígrafe el supuesto de que sea cualitativa.

De entre las medidas de tendencia central, tienen la consideración de ***promedios*** todas aquellas en las que para su determinación intervienen todos los valores de la variable objeto de estudio.

En cualquiera de los casos, una medida de posición debe de estar comprendida entre el valor más pequeño y el más grande que tome la variable de interés, sin más restricciones en este sentido, es decir, pudiendo ser un valor que pueda tomar la variable o no, como por ejemplo 1.2 hijos, y viniendo expresada en las mismas unidades en que venga dada dicha variable.

Tal y como tendremos ocasión de comprobar posteriormente, en el caso de que la variable se encuentre agrupada, será necesario establecer alguna hipótesis sobre la forma en la que se "reparten" las frecuencias dentro de los intervalos para poder especificar el valor de la medida de posición considerada.

A continuación se estudian las principales medidas de posición central, abordando en los dos epígrafes siguientes los principales promedios; la media aritmética, la media armónica, la media geométrica y la media cuadrática. Posteriormente se analizan otras medidas de posición central (mediana y moda). dejando para un epígrafe posterior el estudio de las de posición no central.

4.2 Media aritmética.

La media aritmética es la medida de posición central más utilizada, tanto por su fácil interpretación como por el conjunto de propiedades que de su expresión analítica podemos deducir. Así, la ***media aritmética*** de la variable X, que denotamos $\bar{x}$, es el número obtenido al dividir la suma de todos los valores de la variable entre el número total de observaciones, esto es,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{j=1}^{N} x_j}{N}$$

Centrándonos en las situaciones específicas que hemos estudiado, en el caso de que se disponga de una distribución de frecuencias donde la variable no está agrupada, esta expresión debe considerar la posibilidad de tener en cuenta el número de veces que se repite cada uno de los valores de la variable, esto es

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

Sin embargo, si la variable de interés se encuentra agrupada, únicamente podrá calcularse el verdadero valor de la media aritmética si se dispone de todos los datos originales de la mencionada variable, esto es, si se puede calcular dicho valor a través de la expresión:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{j=1}^{N} x_j}{N}$$

En el caso, más habitual, de que sólo se disponga de la distribución de frecuencias, y por lo tanto no sea posible conocer de forma exacta los valores que toma la variable, debemos establecer alguna ***hipótesis*** *sobre la forma en la que se reparten las frecuencias dentro de cada intervalo*. El supuesto que vamos a realizar en este caso es *agrupar todas las frecuencias de cada intervalo en su punto medio o marca de clase*, de forma que se considera que las n_i observaciones correspondientes al intervalo i-ésimo toman todas el valor x_i, i=1,2,...,k, y entonces, en la expresión

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

x_i denota la marca de clase del intervalo i-ésimo, i=1,2,...,k.

Obviamente, la media aritmética así calculada será una **aproximación** a la verdadera media de la variable X. En todo caso, si disponemos de los totales reales $x_i n_i$ para los distintos intervalos, éstos sí nos posibilitarán el cálculo de la verdadera media.

Ejemplo 4.2.1. Consideremos el número de hijos de cada una de las 15 personas a las que se refiere el ejemplo 3.1.2. Recordemos que los datos, tal y como se recogen en el mencionado ejemplo son los siguientes:

1, 0, 2, 2, 4, 0, 1, 2, 0, 0, 1, 1, 0, 1 y 1.

Si X representa el número de hijos de cada una de estas personas, podemos expresar nuevamente:

$x_1=1$, $x_2=0$, $x_3=2$, $x_4=2$, $x_5=4$, $x_6=0$, $x_7=1$, $x_8=2$, $x_9=0$, $x_{10}=0$, $x_{11}=1$, $x_{12}=1$, $x_{13}=0$, $x_{14}=1$ y $x_{15}=1$.

La media aritmética calculada a partir de estos 15 valores originales es precisamente:

$$\bar{x}=\frac{\sum_{j=1}^{N} x_j}{N}=\frac{1+0+2+...+1}{15}=\frac{16}{15}=1.0667$$

y por lo tanto, las 15 personas que forman el colectivo objeto de estudio tienen una media de 1.0667 hijos.

Obsérvese que 1.0667 es un número comprendido entre 0 y 4, que es un valor que la variable nunca puede tomar, pues necesariamente el número de hijos es un número entero positivo, pero que a pesar de ello es la media aritmética, y que ésta no es un número sin dimensiones, pues viene expresada en hijos.

Si en lugar de utilizar los 15 valores originales de la mencionada variable se emplea la distribución de frecuencias que aparece en el ejemplo 3.1.5, y que era:

NÚMERO DE HIJOS	
x_i	n_i
0	5
1	6
2	3
4	1

al considerar la expresión:

$$\bar{x}=\frac{\sum_{i=1}^{k} x_i n_i}{N}=\frac{\sum_{i=1}^{4} x_i n_i}{15}=\frac{0\cdot 5+1\cdot 6+2\cdot 3+4\cdot 1}{15}=\frac{16}{15}=1.0667$$

se obtiene el mismo valor que al utilizar los originales, pues obviamente

$$\sum_{j=1}^{N} x_j=1+0+2+...+1=0\cdot 5+1\cdot 6+2\cdot 3+4\cdot 1=\sum_{i=1}^{k} x_i n_i$$

Ejemplo 4.2.2. Consideremos ahora la altura, en centímetros, de cada una de las 15 personas a que se refiere el ejemplo 3.1.1, que recordemos eran:

168, 185, 160, 178, 193, 187, 172, 164, 175, 173, 195, 192, 166, 171 y 176.

Si X representa la altura, en centímetros, de cada una de estas personas, podemos expresar:

x_1=168, x_2=185, x_3=160, x_4=178, x_5=193, x_6=187, x_7=172, x_8=164, x_9=175, x_{10}=173, x_{11}=195, x_{12}=192, x_{13}=166, x_{14}=171 y x_{15}=176.

La media aritmética calculada a partir de estos 15 valores originales es:

$$\bar{x}=\frac{\sum_{j=1}^{N} x_j}{N}=\frac{168+185+160+...+176}{15}=\frac{2655}{15}=177$$

y por lo tanto, las 15 personas que forman el colectivo objeto de estudio tienen una altura media de 177 cm. Podemos observar nuevamente que 177 es un número comprendido entre 160 y 195, que es un valor que la variable puede tomar, pero que ninguna de estas 15 personas tiene, pero que a pesar de ello es la media aritmética, y que ésta no es un número sin dimensiones, pues viene expresada en centímetros. Si en lugar de utilizar los 15 valores originales, se emplea la distribución de frecuencias de la variable agrupada X que se propone en el ejemplo 3.2.9, que venía dada por la siguiente tabla:

ALTURA			
L_{i-1} - L_i	a_i	x_i	n_i
160 - 165	5	162.5	2
165 - 170	5	167.5	2
170 - 175	5	172.5	4
175 - 180	5	177.5	2
180 - 185	5	182.5	1
185 - 190	5	187.5	1
190 - 195	5	192.5	3

se tiene entonces que la media aritmética calculada a partir del supuesto de que las frecuencias de cada intervalo se agrupan en su marca de clase, es precisamente

$$\bar{x}=\frac{\sum_{i=1}^{k} x_i n_i}{N}=\frac{\sum_{i=1}^{7} x_i n_i}{15}=$$

$$=\frac{162.5\cdot 2+167.5\cdot+172.5\cdot 4+177.5\cdot 2+182.5\cdot 1+187.5\cdot 1+192.5\cdot 3}{15}=$$

$$=\frac{2652.45}{15}=176.83$$

que no coincide con el auténtico valor de la altura media de estas 15 personas.

Además, al considerar la agrupación en cuatro intervalos propuesta en el ejemplo 3.2.4, que recordemos era:

ALTURA			
L_{i-1} - L_i	a_i	x_i	n_i
160 - 170	10	165	4
170 - 180	10	175	6
180 - 190	10	185	2
190 - 195	5	192.5	3

se tiene que ahora

$$\bar{x}=\frac{\sum_{i=1}^{k} x_i n_i}{N}=\frac{\sum_{i=1}^{4} x_i n_i}{15}=$$

$$=\frac{165\cdot 4+175\cdot 6+185\cdot 2+192.5\cdot 3}{15}=\frac{2657.5}{15}=177.167$$

que no coincide ni con el auténtico valor de la altura media de estas 15 personas, ni con la media obtenida según la anterior agrupación de la variable.

Planteada conceptualmente la definición de media aritmética, podemos pasar seguidamente a estudiar las ***propiedades*** más relevantes de la misma, válidas tanto para la determinación de la media aritmética en distribuciones de frecuencias no agrupadas, como para el caso de las distribuciones de frecuencias agrupadas, con las acotaciones que sobre esta situación ya realizamos. Sin pérdida de generalidad, vamos a trabajar con las del primer tipo, entendiendo que cualquier referencia a los valores de la variable, puede serlo también a los intervalos de valores de la variable, y si fuera preciso, a la marca de clase de cada intervalo.

1) **La *suma de las desviaciones*, o diferencias, de los valores de la variable respecto a su media aritmética es cero**; esto es

$$\sum_{i=1}^{k}(x_i - \bar{x})n_i = 0$$

En efecto, como

$$\sum_{i=1}^{k}(x_i - \bar{x})n_i = \sum_{i=1}^{k}x_i n_i - \bar{x}\sum_{i=1}^{k}n_i$$

y puesto que

$$\bar{x} = \frac{\sum_{i=1}^{k}x_i n_i}{N}$$

se tiene

$$\bar{x}\,N = \sum_{i=1}^{k}x_i n_i$$

por lo que podemos concluir que

$$\sum_{i=1}^{k}(x_i - \bar{x})n_i = \bar{x}\,N - \bar{x}\,N = 0$$

Debido a esta propiedad, la media aritmética se suele denominar ***centro de gravedad***, haciendo referencia a que en este punto se encuentra el correspondiente al de equilibrio de la distribución de frecuencias.

2) **La media aritmética de una variable no varía si todas las *frecuencias* se multiplican, (o dividen) por una constante.** En efecto, consideremos la variable X cuya distribución de frecuencias se puede especificar mediante la tabla:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE X	
x_1	n_1
x_2	n_2
...	...
x_i	n_i
...	...
x_k	n_k

Al multiplicar todas las frecuencias de dicha distribución por una constante, r, donde $r \neq 0$, se tiene la siguiente distribución de frecuencias para la nueva variable X':

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE X'	
x_1	$r\,n_1$
x_2	$r\,n_2$
...	...
x_i	$r\,n_i$
...	...
x_k	$r\,n_k$

y entonces, la media de la variable X' será:

$$\overline{x}' = \frac{\sum_{i=1}^{k} x_i r n_i}{\sum_{i=1}^{k} r n_i} = \frac{r\sum_{i=1}^{k} x_i n_i}{rN} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \overline{x}$$

tal y como se pretendía demostrar.

Esta propiedad conduce a poder calcular la media aritmética a partir de las frecuencias relativas o del porcentaje de observaciones de cada valor de la variable, pues

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

y

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} x_i \frac{100 n_i}{N}}{\sum_{i=1}^{k} \frac{100 n_i}{N}} = \frac{\sum_{i=1}^{k} x_i p_i}{100\frac{N}{N}} = \frac{\sum_{i=1}^{k} x_i p_i}{100}$$

3) **Si a todos los valores de la variable les sumamos, (o restamos) una constante, la media de la nueva variable es igual a la media de la variable original más la constante.** Es decir, si Y=X+a, donde a es una constante, entonces $\overline{y} = \overline{x} + a$.

En efecto, consideremos la siguiente distribución de frecuencias de la variable X:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE X	
x_1	n_1
x_2	n_2
...	...
x_i	n_i
...	...
x_k	n_k

Al sumar a todos los valores de la variable una constante a, se tiene la siguiente distribución de frecuencias de la nueva variable Y=X+a:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE Y	
$y_1=x_1+a$	n_1
$y_2=x_2+a$	n_2
...	...
$y_i=x_i+a$	n_i
...	...
$y_k=x_k+a$	n_k

y entonces la media de la nueva variable Y viene dada por:

$$\overline{y}=\frac{\sum_{i=1}^{k} y_i n_i}{N}=\frac{\sum_{i=1}^{k}(x_i+a)n_i}{N}=\frac{\sum_{i=1}^{k} x_i n_i}{N}+\frac{a\sum_{i=1}^{k} n_i}{N}=\overline{x}+a$$

y por lo tanto,

$$\overline{y}=\overline{x}+a$$

tal y como se pretendía demostrar.

Téngase en cuenta que sumar una constante a la variable X equivale a realizar un ***cambio de origen*** en la variable, y puesto que la media de la nueva variable Y no es la misma que la de X, salvo que la constante sea 0, podemos afirmar que la media aritmética no es invariante frente a cambios de origen en la variable.

Por otra parte, se puede observar que la primera propiedad de la media aritmética, es un caso particular de esta que venimos desarrollando, pues si se considera $a=-\overline{x}$, se tiene que

$$\overline{y}=\overline{x}+a=\overline{x}+(-\overline{x})=0$$

y así, la media aritmética de la variable es cero, por lo que podemos concluir que

$$\sum_{i=1}^{k} y_i n_i=\sum_{i=1}^{k}(x_i-\overline{x})n_i=0$$

4) **Si todos los valores de la variable se multiplican, (o dividen) por una constante, la media de la nueva variable es igual a la media de la variable original multiplicada por la constante.** Es decir, si Y=bX, donde b es una constante, entonces $\overline{y}=b\overline{x}$.

En efecto, consideremos nuevamente la siguiente distribución de frecuencias de la variable X:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE X	
x_1	n_1
x_2	n_2
...	...
x_i	n_i
...	...
x_k	n_k

Al multiplicar todos los valores de la variable por una constante b, se tiene la siguiente distribución de frecuencias de la nueva variable Y=bX:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE Y	
$y_1=bx_1$	n_1
$y_2=bx_2$	n_2
...	...
$y_i=bx_i$	n_i
...	...
$y_k=bx_k$	n_k

y entonces la media de la nueva variable Y viene dada por:

$$\bar{y}=\frac{\sum_{i=1}^{k} y_i n_i}{N}=\frac{\sum_{i=1}^{k} bx_i n_i}{N}=\frac{b\sum_{i=1}^{k} x_i n_i}{N}=b\bar{x}$$

y por lo tanto,

$$\bar{y}=b\,\bar{x}$$

tal y como se pretendía demostrar.

Téngase en cuenta que multiplicar por una constante a la variable X equivale a realizar un ***cambio de escala o unidad*** en la misma, y puesto que la media de la nueva variable Y no es la misma que la de X, salvo que la constante sea 1, podemos afirmar que la media aritmética no es invariante frente a cambios de escala en la variable.

Conjugando esta propiedad con la anterior, se tiene que si se realiza simultáneamente una cambio de origen y de escala en la variable X, tal que Y=a+bX, entonces, la media de la nueva variable vendría afectada por ambos cambios, esto es,

$$\bar{y}=a+b\,\bar{x}$$

Ejemplo 4.2.3. Consideremos nuevamente la altura, en centímetros, de cada una de las 15 personas a que se refiere el ejemplo 4.2.2, cuya distribución de frecuencias se especifica en el mencionado ejemplo. Supongamos que nos interesa la altura media de estas 15 personas, no en centímetros sino en metros.

Una primera solución, sería transformar cada uno de los 15 valores de la variable en metros, esto es, construir una nueva variable Y, tal que Y=X/100. De esta forma se está realizando un cambio de escala o unidad en la variable X.

Si consideramos los 15 valores originales de la variable X, que se especifican en el ejemplo 3.1.1, y se transforman en metros se tendría que tales valores serían:1.68, 1.85, 1.60, 1.78, 1.93, 1.87, 1.72, 1.64, 1.75, 1.73, 1.95, 1.92, 1.66, 1.71 y 1.76.

Si Y representa la altura, en metros, de cada una de estas personas, se tiene que

$$\bar{y}=\frac{\sum_{j=1}^{N} x_j}{N}=\frac{1.68+1.85+1.60+...+1.76}{15}=\frac{26.55}{15}=1.77$$

y por lo tanto, las 15 personas que forman el colectivo objeto de estudio tienen una altura media de 1.77 m.

Ahora bien, si en lugar de utilizar los 15 valores originales, se emplea la distribución de frecuencias de la variable agrupada X según se ha especificado en el ejemplo 4.2.2, que recordemos venía dada por la siguiente tabla:

ALTURA			
$L_{i-1} - L_i$	a_i	x_i	n_i
160 - 165	5	162.5	2
165 - 170	5	167.5	2
170 - 175	5	172.5	4
175 - 180	5	177.5	2
180 - 185	5	182.5	1
185 - 190	5	187.5	1
190 - 195	5	192.5	3

se tiene que la transformación de dicha variable vendría dada por la conversión de los extremos de los intervalos, y por lo tanto de las marcas de clase, pudiendo especificar entonces:

ALTURA			
$L_{i-1} - L_i$	a_i	y_i	n_i
1.60 - 1.65	5	1.625	2
1.65 - 1.70	5	1.675	2
1.70 - 1.75	5	1.725	4
1.75 - 1.80	5	1.775	2
1.80 - 1.85	5	1.825	1
1.85 - 1.90	5	1.875	1
1.90 - 1.95	5	1.925	3

y de esta forma, la media aritmética calculada a partir de estos valores de la variable Y sería:

$$\bar{y} = \frac{\sum_{i=1}^{k} y_i n_i}{N} = \frac{\sum_{i=1}^{7} y_i n_i}{15} =$$

$$= \frac{1.625 \cdot 2 + 1.675 \cdot 2 + 1.725 \cdot 4 + 1.775 \cdot 2 + 1.825 \cdot 1 + 1.875 \cdot 1 + 1.925 \cdot 3}{15} =$$

$$= \frac{26.5245}{15} = 1.7683$$

lo que nos conduce a poder afirmar que las 15 personas que forman el colectivo objeto de estudio tienen una altura media de 1.7683 m.

Sin embargo, si conocemos la media de la variable original X, no es necesario obtener cada uno de los valores de la variable Y para que pueda ser calculada su media, pues si Y=X/100, se tiene que $\bar{y} = \bar{x}/100$, y entonces, $\bar{y} = 177/100 = 1.77$ si se consideran los valores originales, o $\bar{y} = 176.83/100 = 1.7683$ en el supuesto de que se hayan agrupados éstos tal y como se ha ofrecido anteriormente.

Por último, téngase en cuenta que si se consideran los valores originales de la variable X, entonces

$$\frac{\sum_{j=1}^{15} (x_j - \bar{x})}{N} = 0$$

siempre y cuando $\bar{x}$=177, mientras que si se utilizan los intervalos especificados anteriormente,

$$\frac{\sum_{i=1}^{7} (x_i - \bar{x}) n_i}{N} = 0$$

si $\bar{x}$=176.83.

Antes de la proliferación de los ordenadores personales, la utilidad práctica de los cambios de origen y escala en las variables era reducir sus valores, y así facilitar los cálculos, tal y como se propone en el siguiente ejemplo.

Ejemplo 4.2.4. Supongamos que a las 15 personas que vienen conformando nuestro colectivo objeto de estudio, se les solicita que indiquen su salario anual. Si X denota dicho salario, en pesetas, su distribución de frecuencias es la que se ofrece a continuación:

SALARIO	
x_i	n_i
2000000	3
2500000	4
3000000	2
3450000	1
3900000	3
4300000	2

El cálculo de la media a partir de los valores que toma la variable X, puede hacerse añadiendo a la tabla anterior la columna correspondiente al producto de cada valor de la variable por su frecuencia, y una última fila en la que se recoge la suma de las dos últimas columnas, esto es:

SALARIO			
	x_i	n_i	$x_i n_i$
	2000000	3	6000000
	2500000	4	10000000
	3000000	2	6000000
	3450000	1	3450000
	3900000	3	11700000
	4300000	2	8600000
TOTALES	- -	15	45750000

A partir de los datos que nos ofrece esta tabla, se tiene que

$$\overline{X} = \frac{\sum_{i=1}^{6} x_i n_i}{N} = \frac{45750000}{15} = 3050000$$

por lo que podemos concluir que el salario medio para las 15 personas es de 3050000 ptas.

Ahora bien, si construimos una nueva variable Y, tal que

$$Y = \frac{X - 3000000}{1000000}$$

podemos expresar entonces:

	Y		
	y_i	n_i	$y_i n_i$
	-1	3	-3
	-0.5	4	-2
	0	2	0
	0.45	1	0.45
	0.9	3	2.7
	1.3	2	2.6
TOTALES	--	15	0.75

De esta forma, la media de la variable Y es

$$\overline{y} = \frac{\sum_{i=1}^{6} y_i n_i}{N} = \frac{0.75}{15} = 0.05$$

y entonces, al ser

$$Y = \frac{X - 3000000}{1000000}$$

se tiene que

$$\overline{y} = \frac{\overline{x} - 3000000}{1000000}$$

por lo que

$$\overline{x} = \overline{y} \cdot 1000000 + 3000000 = 0.05 \cdot 1000000 + 3000000 = 30500000$$

y así se puede concluir que el salario medio para las 15 personas es de 3050000 ptas.

Obsérvese la gran diferencia que existe entre los valores de la columna $x_i n_i$ con la de $y_i n_i$.

Por otra parte, cuando se calcula la media aritmética de una variable X, se está atribuyendo la misma importancia a cada uno de los valores que toma dicha variable, pues incluso cuando se obtiene mediante la expresión:

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

sólo se está considerando el número de veces que aparece un determinado valor o el número de los que se encuentran en un determinado intervalo. Sin embargo hay ocasiones en las que debe asignarse un peso o ponderación que haga distinciones entre los valores de la variable, apareciendo lo que se denomina ***media aritmética ponderada***, que viene dada por:

$$\overline{x} = \frac{\sum_{j=1}^{N} x_j w_j}{\sum_{j=1}^{N} w_j}$$

donde w_j son los ***pesos o factores de ponderación***, tales que $w_j \geq 0, \forall j$.

Ejemplo 4.2.5. Un examen consta de tres partes distintas, un cuestionario tipo test, unas preguntas de teoría y un conjunto de ejercicios. En la calificación final no tiene la misma importancia cada una de las partes, y así, se le asigna al test un peso del 20%, a la teoría el 30%, y lógicamente a los ejercicios el 50% restante. Si un alumno ha obtenido 4 puntos en el test, 3 puntos en la teoría y 7 puntos en los ejercicios, su puntuación final es, si X es una variable que denota la puntuación de cada una de las partes:

$$\overline{x} = \frac{\sum_{j=1}^{3} x_j w_j}{\sum_{j=1}^{3} w_j} = \frac{4 \cdot 20 + 3 \cdot 30 + 7 \cdot 50}{100} = \frac{520}{100} = 5.2$$

y por lo tanto su calificación final es de 5.2 puntos.

Obsérvese que si no se hubiera asignado distintos pesos a cada una de las partes del examen, la calificación del alumno sería:

$$\overline{x} = \frac{\sum_{j=1}^{3} x_j n_j}{N} = \frac{4+3+7}{3} = \frac{14}{3} = 4.67$$

esto es, la media aritmética simple, en la que el término simple va a ser omitido en adelante, es 4.67 puntos.

Es muy habitual que los pesos se definan de forma que su suma sea la unidad, evitando de esta forma tener que dividir por dicha suma a la hora de calcular la media aritmética ponderada.

Ejemplo 4.2.6. En el supuesto especificado en el ejemplo anterior, consideremos los pesos, $w_1=0.2$, $w_2=0.3$ y $w_3=0.5$. Entonces se tiene que

$$\overline{x} = \frac{\sum_{j=1}^{3} x_j w_j}{\sum_{j=1}^{3} w_j} = \frac{4 \cdot 0.2 + 3 \cdot 0.3 + 7 \cdot 0.5}{1} = 5.2$$

lo que conduce al mismo valor que el obtenido anteriormente. Así, cualquier transformación en los pesos que mantenga de forma proporcional la importancia de cada uno de los valores de la variable conduce al mismo valor de la media aritmética ponderada.

Resulta muy interesante analizar cómo determinar la media aritmética de una variable X, cuando ésta se considera en un conjunto de poblaciones similares, (subpoblaciones), que de forma habitual se

denota ***composición de poblaciones***, tal y como se propone en el siguiente supuesto.

Ejemplo 4.2.7. Supongamos que las 15 personas a las que se refiere el ejemplo 3.1.2, han nacido en 4 Comunidades Autónomas distintas, de forma que las cinco primeras lo hicieron en Andalucía, las tres siguientes en Extremadura, las cinco siguientes en la Comunidad de Madrid, y las dos últimas en la Comunidad Valenciana. Entonces, si X representa el número de hijos de cada una de estas 15 personas, cuyos valores se ofrecen en el mencionado ejemplo, podemos formar la siguiente tabla, donde se tiene en cuenta el lugar de nacimiento de cada uno de ellos:

NÚMERO DE HIJOS

x_i	Nacidos en Andalucía	Nacidos en Extremadura	Nacidos en la Comunidad de Madrid	Nacidos en la Comunidad Valenciana	Total de personas con x_i hijos (n_i)
0	1	1	3	0	5
1	1	1	2	2	6
2	2	1	0	0	3
4	1	0	0	0	1
Total de personas por comunidad de nacimiento	5	3	5	2	- -

El número medio de hijos que tienen estas 15 personas viene dado por:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i n_i}{N} = \frac{0 \cdot 5 + 1 \cdot 6 + 2 \cdot 3 + 4 \cdot 1}{15} = \frac{16}{15} = 1.0667$$

donde n_i es lógicamente el total de personas con x_i hijos, i=1,2,3,4, y así, se tiene una media de 1.0667 hijos, tal y como se había obtenido en el ejemplo 4.2.1.

Ahora bien, ¿sería posible obtener esta media a partir del conocimiento de la media de la variable X en cada una de las comunidades autónomas?. Esto es, si conocemos el número medio de hijos que tienen las personas que han nacido en cada una de estas comunidades autónomas, ¿sería posible obtener la media para todas ellas?. A continuación, vamos a dar respuesta a esta cuestión, y tal y como vamos a comprobar, sí es posible calcular la media de la variable X a partir de las medias de dicha variable en cada comunidad, siempre y cuando sepamos también cuantas personas nacieron en cada una de ellas.

Sea una población P de tamaño N, formada por varias subpoblaciones P_1, P_2, ..., P_m, de tamaños N_1, N_2, ..., N_m, siendo entonces

$$N = \sum_{j=1}^{m} N_j$$

y sea n_{ij}, i=1,2,...,k, j=1,2,...,m, el número de elementos que en la subpoblación P_j alcanzan el valor x_i de la variable X. Así, para la población P, el número de elementos que toman el valor x_i, n_i, es precisamente

$$n_i = \sum_{j=1}^{m} n_{ij}$$

Bajo el supuesto de que la variable no se agrupa, simplificando entonces la presentación, podemos especificar su distribución de frecuencias mediante la siguiente tabla:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE X					
	P_1	P_2	... P_j ...	P_m	P (TOTALES)
x_1	n_{11}	n_{12}	... n_{1j} ...	n_{1m}	n_1
x_2	n_{21}	n_{22}	... n_{2j} ...	n_{2m}	n_2
...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	... n_{ij} ...	n_{im}	n_i
...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	... n_{kj} ...	n_{km}	n_k
TOTALES	N_1	N_2	... N_j ...	N_m	N

La media de la variable X en la subpoblación j, $\bar{x}_j$, j=1,2,..,m, se puede expresar

$$\bar{x}_j = \frac{\sum_{i=1}^{k} x_i n_{ij}}{N_j}$$

mientras que la media de dicha variable para la población P viene dada por

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

Ahora bien, dado que el total de valores observados de la población tiene que coincidir con los totales de los valores observados para todas las subpoblaciones, se puede decir que

$$\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i \sum_{j=1}^{m} n_{ij} = \sum_{j=1}^{m} \sum_{i=1}^{k} x_i n_{ij} = \sum_{j=1}^{m} \bar{x}_j N_j$$

y como la media de la población se puede definir como el cociente entre los valores observados y el total de observaciones, se tiene que

$$\overline{X} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{\sum_{j=1}^{m} \overline{x}_j N_j}{N}$$

y por lo tanto, la media aritmética de la variable X en la población P, esto es, la ***media aritmética de la composición de poblaciones***, puede expresarse mediante una media aritmética ponderada de las medias aritméticas de dicha variable en cada una de las subpoblaciones, donde las ponderaciones son precisamente el número de elementos de cada una de ellas.

Ejemplo 4.2.8. Consideremos nuevamente el supuesto especificado en el ejemplo 4.2.7, y supongamos que en lugar de ofrecer la información del número de hijos y comunidad de nacimiento de cada una de las 15 personas, se nos indica que las cinco personas que nacieron en Andalucía tienen una media de 1.8 hijos; que las 3 que lo hicieron en Extremadura tienen 1 hijo por término medio; que las cinco que nacieron en la Comunidad de Madrid tienen una media de 0.4 hijos; y que los dos que lo hicieron en la Comunidad Valenciana, tienen también 1 hijo por término medio.

En este caso, también es posible calcular el número medio de hijos que tienen estas 15 personas, pues:

$$\overline{x} = \frac{\sum_{j=1}^{m} \overline{x}_j N_j}{N} = \frac{1.8 \cdot 5 + 1 \cdot 3 + 0.4 \cdot 5 + 1 \cdot 2}{15} = \frac{16}{15} = 1.0667$$

y por lo tanto, el colectivo objeto de estudio tiene una media de 1.0667 hijos, que naturalmente coincide con la obtenida en el mencionado ejemplo.

Obsérvese que si no se conociera el número de personas que han nacido en cada comunidad autónoma, no se podría calcular exactamente la media solicitada, pues al considerar la media aritmética simple de las cuatro medias que se ofrecen, se tiene que

$$\overline{x} = \frac{\sum_{j=1}^{m} \overline{x}_j}{m} = \frac{1.8+1+0.4+1}{4} = \frac{15.75}{15} = 1.05$$

que como podemos comprobar no coincide con la verdadera media del número de hijos de estas 15 personas.

Un caso particular de la composición de poblaciones es aquel en el que una distribución de frecuencias se divide en dos o más ***subdistribuciones disjuntas***, tal y como se plantea en el siguiente supuesto.

Ejemplo 4.2.9. Supongamos que las 15 personas a que se refiere el ejemplo 3.1.2, trabajan en tres secciones distintas, denominadas A, B y C, de forma que la distribución de frecuencias de la variable X, que representa el número de hijos de cada una de éstas, donde se tiene en cuenta la sección en la que trabaja cada uno de ellas, es la siguiente:

NÚMERO DE HIJOS				
x_i	Sección A	Sección B	Sección C	n_i
0	5	0	0	5
1	0	6	0	6
2	0	0	3	3
4	0	0	1	1
TOTALES	5	6	4	15

Este supuesto, es, obviamente, un caso particular de una composición de poblaciones, y por lo tanto, la media de la variable X puede calcularse a partir del conocimiento de las medias y del total de personas que trabajan en cada una de las secciones, pues

$$\overline{X} = \frac{\sum_{j=1}^{3} \overline{x}_j N_j}{N} = \frac{0 \cdot 5 + 1 \cdot 6 + 2,5 \cdot 4}{15} = \frac{16}{15} = 1,0667$$

y así, las 15 personas tienen por término medio 1.0667 hijos, que naturalmente coincide con la obtenida en los ejemplos anteriores.

Por lo visto hasta este momento, parece que la media aritmética no presenta ningún inconveniente, salvo la aproximación que se produce en su valor cuando se agrupa una variable, y lógicamente, cuando *alguna marca de clase no se pueda obtener*, pues en este caso tampoco se puede calcular el valor de la media aritmética, tal y como ocurre con la distribución de frecuencias especificada en el ejemplo 3.2.3, que recordemos era

ALTURA	
Menos de 170	4 personas
Más de 170 pero menos de 180	6 personas
Más de 180 pero menos de 190	2 personas
Más de 190	3 personas

Sin embargo, el mayor inconveniente que puede presentar la media aritmética es su gran ***sensibilidad*** cuando en la distribución de frecuencias se presentan *valores anormalmente pequeños o grandes*, que denominamos habitualmente ***atípicos*** (***outliers***), pues en

cualquiera de estos casos, este promedio pierde en cierta medida su carácter de representatividad.

Ejemplo 4.2.10. Consideremos nuevamente la distribución de frecuencias correspondiente al número de hijos de las 15 personas que se especifica en la tabla que aparece en el ejemplo 4.2.1, que era:

NÚMERO DE HIJOS	
x_i	n_i
0	5
1	6
2	3
4	1

En este supuesto, las 15 personas tienen una media de 1.0667 hijos. Supongamos que la persona que tiene 4 hijos es reemplazada por otra que tiene 18. La nueva distribución de frecuencias de la variable Y que denota el número de hijos, se puede especificar:

NÚMERO DE HIJOS	
y_i	n_i
0	5
1	6
2	3
18	1

La media de la variable Y es:

$$\bar{y}=\frac{\sum_{i=1}^{4} y_i n_i}{15}=\frac{0\cdot 5+1\cdot 6+2\cdot 3+18\cdot 1}{15}=\frac{30}{15}=2$$

y por lo tanto, por término medio, estas 15 personas tienen 2 hijos.

Podemos observar entonces la diferencia, de cierta consideración, que existe entre la media aritmética de X y de Y, pues la de Y dobla prácticamente a la de X. Así, el hecho de que en la distribución de la variable Y exista un valor especialmente grande, hace que su media aritmética no parece que sea igual de representativa que la de la variable X, en donde no se presenta este problema.

Ahora bien, a pesar de estos últimos aspectos que se acaban de ofrecer, la media aritmética es el promedio por excelencia, esto es, el más comúnmente empleado. Sin embargo, en algunos casos, no es adecuado este promedio, apareciendo entonces otros como la media armónica, la media geométrica y la media cuadrática, que no van a ser presentados en el presente trabajo.

4.3 Mediana.

Abordamos a continuación otro tipo de medidas de tendencia central, de entre los que se van a considerar la mediana y la moda. Esta última será analizada en el siguiente epígrafe.

Supongamos que se ordenan en sentido creciente, esto es, de menor a mayor, todos los valores de la variable de interés. Pues bien, la ***mediana***, que vamos a denotar M_e, es **aquel valor de la variable que ocupa el lugar central.** Así pues, se busca el centro de la distribución sin promediar los valores de la variable, únicamente ordenándolos y a continuación determinando cual de ellos es el que

consigue formar dos grupos en los que haya el mismo número de observaciones.

Ejemplo 4.3.1. Consideremos el número de hijos de cada una de las 15 personas a las que se refiere el ejemplo 3.1.2. Los datos, tal y como se recogen en el mencionado ejemplo son:

1, 0, 2, 2, 4, 0, 1, 2, 0, 0, 1, 1, 0, 1 y 1.

Si ordenamos en orden ascendente estos valores, tenemos,

0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2 y 4,

y si X representa el número de hijos de cada una de estas personas, podemos expresar, entonces:

$x_1=0$, $x_2=0$, $x_3=0$, $x_4=0$, $x_5=0$, $x_6=1$, $x_7=1$, $x_8=1$, $x_9=1$, $x_{10}=1$, $x_{11}=1$, $x_{12}=2$, $x_{13}=2$, $x_{14}=2$ y $x_{15}=4$.

Obviamente x_8 es el valor de la variable que ocupa el lugar central, pues consigue formar dos grupos en los que hay siete valores en cada uno de ellos, de x_1 a x_7, y de x_9 a x_{15}. Por tanto, $M_e=1$, esto es, el valor mediano es 1 hijo.

Obsérvese que el mismo resultado se obtiene si en lugar de ordenar los valores de la variable en sentido ascendente, se ordenan en sentido descendente, y así, el hecho realmente importante para calcular la mediana, es ordenar los valores de la variable, sea cual sea el sentido en el que se haga.

Puesto que en el cálculo de la mediana no se promedian todos los valores de la variable, ésta *no será* ***sensible*** *cuando en su distribución de frecuencias se presenten valores anormalmente pequeños o grandes, atípicos,* por lo que en esos casos puede ser una medida de tendencia central preferible a la media aritmética.

Ejemplo 4.3.2. Supongamos, tal y como se recoge en el ejemplo 4.2.10, que la persona, de las 15 que se vienen considerando a lo largo de este trabajo, que tiene 4 hijos es reemplazada por otra que tiene 18. En el mencionado ejemplo, comprobamos que esta sustitución hacía que la media aritmética prácticamente se doblara, pues de 1.0667 pasaba a 2 hijos.

Sin embargo, la mediana, que tal y como se recoge en el ejemplo anterior es 1 hijo, no se modifica cuando se sustituye a la mencionada persona, pues en este caso, si X representa el número de hijos de cada una de estas personas, podemos expresar, entonces:

$x_1=0$, $x_2=0$, $x_3=0$, $x_4=0$, $x_5=0$, $x_6=1$, $x_7=1$, $x_8=1$, $x_9=1$, $x_{10}=1$, $x_{11}=1$, $x_{12}=2$, $x_{13}=2$, $x_{14}=2$ y $x_{15}=18$,

y así, x_8 sigue siendo el valor de la variable que ocupa el lugar central, pues consigue formar dos grupos en los que hay siete valores en cada uno de ellos, de x_1 a x_7, y de x_9 a x_{15}.

Cuando el número de observaciones es impar, no existe problema alguno para la determinación de la mediana, pues siempre es un valor que necesariamente toma la variable, lo que en muchos casos facilita su interpretación frente a otras medidas de tendencia central, como los promedios considerados en los epígrafes anteriores. Sin embargo, *cuando N es par, debemos realizar alguna* ***hipótesis*** *para poder especificar cuál es la mediana*, tal y como se muestra en el siguiente supuesto.

Ejemplo 4.3.3. Supongamos que a las 15 personas a que nos hemos referido en el ejemplo 4.3.1, se le añade otra que tiene 2 hijos. Entonces, una vez ordenados los valores de la variable X, se tiene que:

$x_1=0$, $x_2=0$, $x_3=0$, $x_4=0$, $x_5=0$, $x_6=1$, $x_7=1$, $x_8=1$, $x_9=1$, $x_{10}=1$, $x_{11}=1$, $x_{12}=2$, $x_{13}=2$, $x_{14}=2$, $x_{15}=2$ y $x_{16}=4$.

Es claro que no hay ningún valor de la variable que ocupe el lugar central, pues de x_1 a x_8 hay 8 valores, y de x_9 a x_{16} hay otros 8.

Para resolver este problema, vamos a considerar como mediana el valor medio de x_8 y x_9, esto es:

$$M_e=\frac{x_8+x_9}{2}=\frac{1+1}{2}=1$$

y así, consideramos que la mediana es 1 hijo.

Esta forma de resolver el problema planteado cuando N es par, puede ocasionar que la mediana no sea un valor que tome la variable, tal y como se ofrece a continuación.

Ejemplo 4.3.4. Supongamos que a las 15 personas a que se refiere el ejemplo 4.3.1, se les añaden otras 7 que tienen 2 hijos.

Entonces, una vez ordenados los valores de la variable X, se tiene que:

$x_1=0$, $x_2=0$, $x_3=0$, $x_4=0$, $x_5=0$, $x_6=1$, $x_7=1$, $x_8=1$, $x_9=1$, $x_{10}=1$, $x_{11}=1$, $x_{12}=2$, $x_{13}=2$, $x_{14}=2$, $x_{15}=2$, $x_{16}=2$, $x_{17}=2$, $x_{18}=2$, $x_{19}=2$, $x_{20}=2$, $x_{21}=2$ y $x_{22}=4$.

Nuevamente no existe ningún valor de la variable que ocupe el lugar central, pues de x_1 a x_{11} hay 11 valores, y de x_{12} a x_{22} hay otros 11. Entonces, manteniendo que la mediana es el valor medio de x_{11} y x_{12}, se tiene que

$$M_e=\frac{x_{11}+x_{12}}{2}=\frac{1+2}{2}=1.5$$

y así, consideramos que la mediana es 1.5 hijos, que es un valor que no puede tomar la variable.

2º Pues bien, a pesar de este inconveniente, el criterio que vamos a utilizar para especificar la mediana **cuando el número de observaciones es par es calcular la media aritmética de los dos valores centrales de la variable**, aunque hay autores que proponen la consideración de dos medianas, o de ninguna, o incluso de cualquier valor comprendido entre dichos valores centrales.

Ahora bien, normalmente no se dispone de los datos originales, sino de la distribución de frecuencias de la variable de interés. En este sentido, si la variable no ha sido agrupada, la mediana será la misma que la que se podría obtener a partir de los datos originales, pudiendo ofrecer la siguiente regla de actuación:

1. Ordenar los valores de la variable.
2. Calcular las frecuencias acumuladas N_i, i=1,2,...,k.
3. Obtener el valor de N/2.
4. Si no hay ningún valor de N_i que coincida con N/2, determinar el primero tal que $N_i>N/2$, y si éste es N_h, entonces $M_e=x_h$.
5. Si hay un valor de N_i que coincida con N/2, si éste es N_h, entonces $M_e=(x_h+x_{h+1})/2$.

Ejemplo 4.3.5. Consideremos la distribución de frecuencias de la variable número de hijos de las 15 personas a que se refiere el ejemplo 4.3.1.

NÚMERO DE HIJOS		
x_i	n_i	N_i
0	5	5
1	6	11
2	3	14
4	1	15

Puesto que N/2=15/2=7.5, es imposible al ser N impar que exista algún N_i, i=1,2,3,4, que coincida con N/2, y por lo tanto únicamente debemos buscar cuál es la primera frecuencia acumulada que sobrepasa a 7.5; como quiera que esta es N_2=11, se tiene que $M_e=x_2$, y por lo tanto la mediana es 1 hijo.

Si consideramos ahora la distribución de frecuencias que se propone en el ejemplo 4.3.3, que recordemos era:

NÚMERO DE HIJOS		
x_i	n_i	N_i
0	5	5
1	6	11
2	4	15
4	1	16

Puesto que N/2=16/2=8, cabe la posibilidad, al ser N par, de que algún N_i, i=1,2,3,4, coincida con N/2. Sin embargo, no es este el caso, pues N_1=5 y N_2=11, y por lo tanto, únicamente nos interesa cuál es el primer N_i que sobrepasa a N/2, y puesto que éste es N_2, se tiene que $M_e=x_2$ y por lo tanto la mediana es 1 hijo.

Si consideramos ahora la distribución de frecuencias que se propone en el ejemplo 4.3.4, que venía dada por:

NÚMERO DE HIJOS		
x_i	n_i	N_i
0	5	5
1	6	11
2	10	21
4	1	22

Puesto que $N/2=22/2=11$, cabe nuevamente la posibilidad, al ser N par, de que algún N_i, $i=1,2,3,4$, coincida con $N/2$. En este caso, $N_2=N/2=11$, y por lo tanto, la mediana será la media aritmética de x_2 y x_3, esto es $M_e=(x_2+x_3)/2=(1+2)/2$ y por lo tanto el valor mediano es 1.5 hijos.

Ahora bien, en el caso de que la variable haya sido agrupada necesitamos, al igual que al calcular cualquiera de los promedios considerados en los epígrafes anteriores, *realizar alguna* ***hipótesis*** *sobre la forma en la que las frecuencias se reparten en los intervalos*, tal y como se plantea en el siguiente supuesto.

Ejemplo 4.3.6. Consideremos la altura, en centímetros, de las 15 personas a las que se refiere el ejemplo 3.1.1. Los datos, tal y como se recogen en el mencionado ejemplo son:

168, 185, 160, 178, 193, 187, 172, 164, 175, 173, 195, 192, 166, 171 y 176.

Si ordenamos en orden ascendente estos valores, tenemos,

160, 164, 166, 168, 171, 172, 173, 175, 176, 178, 185, 187, 192, 193 y 195,

y si X representa la altura, en centímetros, de cada una de estas personas, podemos expresar, entonces:

$x_1=160$, $x_2=164$, $x_3=166$, $x_4=168$, $x_5=171$, $x_6=172$, $x_7=173$, $x_8=175$, $x_9=176$, $x_{10}=178$, $x_{11}=185$, $x_{12}=187$, $x_{13}=192$, $x_{14}=193$ y $x_{15}=195$.

Fácilmente podemos concluir que el valor mediano para estas 15 personas es 175 cm., pues x_8 es el valor de la variable que ocupa el lugar central dado que consigue formar dos grupos en los que hay siete valores en cada uno de ellos, de x_1 a x_7, y de x_9 a x_{15}, y así, $M_e=x_8=175$.

Supongamos ahora que únicamente se dispone de la distribución de frecuencias que se propone en el ejemplo 3.2.4, que recordemos era:

ALTURA				
L_{i-1} - L_i	a_i	x_i	n_i	N_i
160 - 170	10	165	4	4
170 - 180	10	175	6	10
180 - 190	10	185	2	12
190 - 195	5	192,5	3	15

Es claro que la mediana es un valor comprendido entre 170 y 180 cm., pues hay 4 observaciones inferiores a 170 y otras 5 que son superiores a 180. Sin embargo, como no se conocen exactamente cuáles son los 6 valores de la variable que se encuentran en dicho intervalo, no se puede especificar cuál es el valor central que denominamos mediana.

Se hace imprescindible pues considerar alguna hipótesis sobre el reparto de las frecuencias en dicho intervalo, no siendo necesario efectuar dicha hipótesis sobre el resto de los intervalos, esto es, necesitamos distribuir los 6 valores correspondientes al intervalo (170-180]. Pues bien, a la hora de calcular esta medida de tendencia central, **vamos a suponer que dicho reparto es uniforme**, esto es, se va a utilizar la misma hipótesis que la considerada al construir el diagrama acumulativo.

Obsérvese entonces, que para calcular cualquiera de los promedios especificados en los epígrafes anteriores, se supone que las frecuencias de un determinado intervalo se ubican en la marca de clase correspondiente, mientras que para calcular la mediana, se va a suponer que dichas frecuencias se reparten por igual dentro del intervalo en el que se encuentra la mediana. Esta hipótesis es la que permite especificar un valor, siempre aproximado, de la misma.

Si retomamos el supuesto planteado en el ejemplo 4.3.6, vamos a representar en el diagrama acumulativo de frecuencias de la variable X, el punto que vamos a considerar como la mediana de la distribución de frecuencias, tal y como se ofrece en la figura 4.3.1.

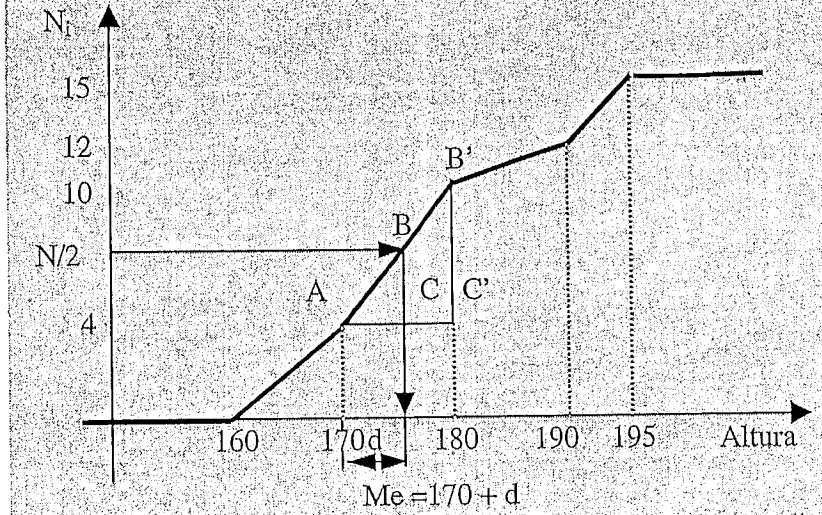

Figura 4.3.1: Determinación de la mediana correspondiente al ejemplo 3.2.4

Evidentemente, tal y como muestra la figura 4.3.1, M_e=170+d, y por lo tanto el único problema estriba en calcular el valor de d. Ahora bien, dada la semejanza de los triángulos ACB y AC'B', podemos expresar

$$\frac{\overline{AC}}{\overline{A'C}}=\frac{\overline{CB}}{\overline{C'B'}}$$

y como

$$\overline{AC}=d;\ \overline{A'C}=10;\ \overline{CB}=\frac{15}{2}-4;\ \overline{C'B'}=10-4$$

se tiene que

$$\frac{d}{10}=\frac{\frac{15}{2}-4}{10-4}$$

y por lo tanto,

$$M_e = 170 + d = 170 + \frac{\frac{15}{2} - 4}{10 - 4} 10 = 170 + \frac{\frac{15}{2} - 4}{6} 10 = 175{,}83$$

que no coincide con la verdadera mediana de este conjunto de observaciones, que es 175 cm.

Con carácter general, dada la distribución de frecuencias de la variable agrupada X:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE AGRUPADA X				
L_0-L_1	a_1	x_1	n_1	$N_1=n_1$
L_1-L_2	a_2	x_2	n_2	N_2
...	...	...	...	...
L_{i-2}-L_{i-1}	a_{i-1}	x_{i-1}	n_{i-1}	N_{i-1}
L_{i-1}-L_i	a_i	x_i	n_i	N_i
L_i-L_{i+1}	a_{i+1}	x_{i+1}	n_{i+1}	N_{i+1}
...	...	...	...	...
L_{k-1}-L_k	a_k	x_k	n_k	$N_k=N$

donde $L_0<L_1<...<L_k$, supongamos que el intervalo en el que encuentra la mediana; esto es, el primer intervalo en el que la frecuencia acumulada es superior a N/2, es el intervalo i-ésimo. Entonces, el diagrama acumulativo de la variable X, donde incluimos nuevamente el punto que vamos a considerar como la mediana de la distribución de frecuencias, es el que se ofrece en la figura 4.3.2.

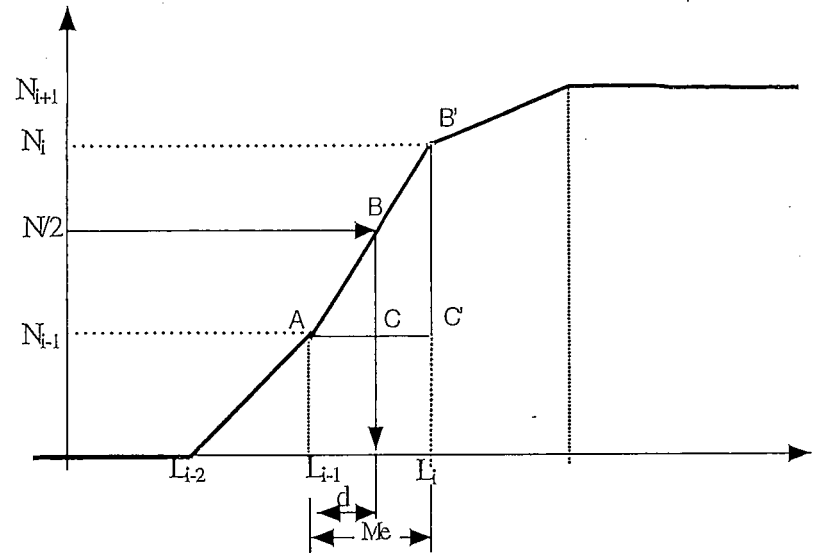

Figura 4.3.2: Determinación de la mediana para variables agrupadas.

En este caso, tal y como muestra la figura 4.3.2, $M_e=L_{i-1}+d$, y nuevamente, dada la semejanza de los triángulos ACB y AC'B', podemos expresar

$$\frac{\overline{AC}}{\overline{A'C}} = \frac{\overline{CB}}{\overline{C'B'}}$$

y como

$$\overline{AC} = d;\ \ \overline{A'C} = a_i;\ \ \overline{CB} = \frac{N}{2} - N_{i-1};\ \ \overline{C'B'} = N_i - N_{i-1}$$

se tiene que

$$\frac{d}{a_i} = \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}$$

y por lo tanto,

$$M_e = L_{i-1} + d = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} a_i = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

que es la expresión que vamos a utilizar para calcular la mediana de una distribución de frecuencias cuya variable está agrupada.

Ejemplo 4.3.7. Nótese que la aplicación de esta expresión al cálculo de la mediana correspondiente al supuesto especificado en el ejemplo 4.3.6, conduce, a partir de su distribución de frecuencias, que recordemos era

ALTURA				
$L_{i-1} - L_i$	a_i	x_i	n_i	N_i
160 - 170	10	165	4	4
170 - 180	10	175	6	10
180 - 190	10	185	2	12
190 - 195	5	192.5	3	15

a poder especificar, una vez determinado que el intervalo donde se encuentra la mediana es (170-180], que:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 170 + \frac{\frac{15}{2} - (4)}{(6)} 10 = 175.83$$

y por lo tanto que el valor mediano es 175.83 cm.

Obsérvese que cuando la variable se encuentra agrupada, es imprescindible para poder especificar la mediana de la distribución, que el intervalo en el que se encuentra dicha medida de tendencia central esté bien definido, pudiendo plantearse entonces como *una alternativa a los promedios* en distribuciones en las que esto no sucede para todos los intervalos, pero sí para el que engloba a la mediana. Así en la distribución especificada en el ejemplo 3.2.3, que recordemos era:

ALTURA	
Menos de 170	4 personas
Más de 170 pero menos de 180	6 personas
Más de 180 pero menos de 190	2 personas
Más de 190	3 personas

no se pueden obtener los promedios, mientras que el valor mediano es 175.83 hijos, tal y como se ha ofrecido en los ejemplos 4.3.6 y 4.3.7.

Sin embargo, hemos de mencionar que el valor de la mediana depende, cuando la variable está agrupada, de la forma en la que se agrupen los valores en los distintos intervalos. Así, consideremos el siguiente supuesto.

Ejemplo 4.3.8. Supongamos que en lugar de considerar la agrupación propuesta en el ejemplo anterior sobre la altura de las 15 personas, utilizamos la especificada en el ejemplo 3.2.9, que venía dada por:

ALTURA				
$L_{i-1} - L_i$	a_i	x_i	n_i	N_i
160 - 165	5	162.5	2	2
165 - 170	5	167.5	2	4
170 - 175	5	172.5	4	8
175 - 180	5	177.5	2	10
180 - 185	5	182.5	1	11
185 - 190	5	187.5	1	12
190 - 195	5	192.5	3	15

Puesto que el primer valor que excede a N/2=15/2=7.5 es N_3=8, el intervalo donde se encuentra la mediana es (170-175], y así,

$$M_e = L_{(i-1)} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 170 + \frac{\frac{15}{2} - 4}{4} 5 = 174.375$$

y por lo tanto que el valor mediano es 174.375 cm., diferente al obtenido al calcular dicho valor mediante la distribución de frecuencias:

ALTURA				
$L_{i-1} - L_i$	a_i	x_i	n_i	N_i
160 - 170	10	165	4	4
170 - 180	10	175	6	10
180 - 190	10	185	2	12
190 - 195	5	192.5	3	15

puesto que en este caso, la mediana era 175.83 cm., tal y como se ha obtenido en ejemplos anteriores.

4º También debemos hacer notar que cuando en una distribución de frecuencias de una variable agrupada, existe una frecuencia acumulada que coincide con N/2, entonces, el valor de la mediana es precisamente el extremo superior del intervalo para el cual dicha frecuencia acumulada coincide con N/2. A continuación se ofrece un supuesto en este sentido.

Ejemplo. 4.3.9. Supongamos que a las 15 personas a que se refiere el ejemplo 4.3.6, se le añade otra cuya altura es 186 cm. La distribución de frecuencias de la variable agrupada X que denota la altura de cada una de estas 16 personas se puede especificar de la siguiente forma:

ALTURA				
$L_{i-1} - L_i$	a_i	x_i	n_i	N_i
160 - 165	5	162.5	2	2
165 - 170	5	167.5	2	4
170 - 175	5	172.5	4	8
175 - 180	5	177.5	2	10
180 - 185	5	182.5	1	11
185 - 190	5	187.5	2	13
190 - 195	5	192.5	3	16

siendo su diagrama acumulativo el que se ofrece en la figura 4.3.3.

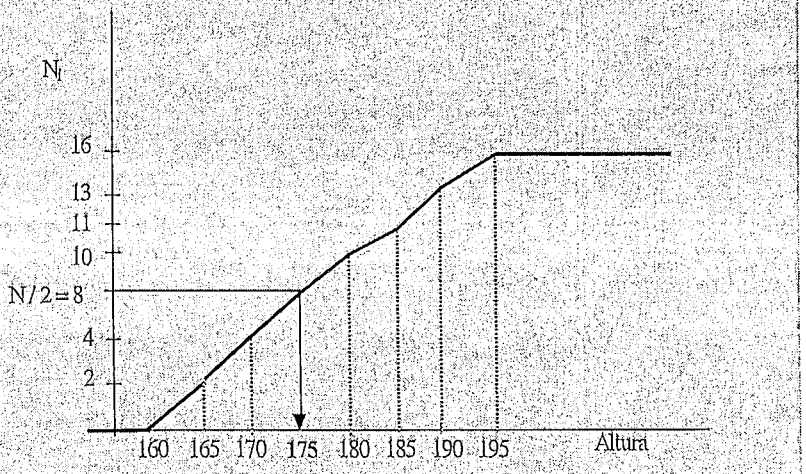

Figura 4.3.3: Determinación de la mediana correspondiente al ejemplo 4.3.9

La representación gráfica que aparece en la figura 4.3.3, pone de manifiesto, sin tener que recurrir al cálculo del punto que denominamos mediana, que $M_e=L_3=175$, pudiendo llegar al mismo resultado al considerar la expresión general considerada en este epígrafe, pues al ser N_4 la primera frecuencia acumulada que excede a $N/2=16/2=8$, se tiene que

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 175 + \frac{\frac{16}{2} - 8}{2} 5 = 175$$

También debemos resaltar que cuando se considera una ***composición de poblaciones***, es imposible calcular la mediana de la población a partir del conocimiento de la mediana de cada una de las subpoblaciones. En este sentido, sólo se puede probar que la mediana de la población será un valor que necesariamente estará comprendido entre el más pequeño y el más grande del conjunto formado por todas y cada una de las medianas de las subpoblaciones.

4.4 Moda.

La ***moda***, que denotamos M_o, es otra medida de tendencia central que podemos definir como **el valor más frecuente de la variable**, esto es, el que más se repite. Así, al igual que la mediana, no promedia los valores de la variable, sino que busca aquel valor de la variable que se presenta más veces, por lo que las frecuencias se configuran de alguna forma como el elemento básico para su determinación.

Ejemplo 4.4.1. Consideremos el número de hijos de cada una de las 15 personas a las que se refiere el ejemplo 3.1.2. Los datos, tal y como se recogen en el mencionado ejemplo son:

1, 0, 2, 2, 4, 0, 1, 2, 0, 0, 1, 1, 0, 1 y 1.

Es evidente que al haber 6 personas que tienen un hijo, mientras que son 5 las que no tienen, 3 las que tienen 2 y 1 la que tiene 4, podemos concluir que el valor que más se repite es el 1, y por lo tanto la moda es un hijo; esto es, $M_o=1$.

Más rápidamente se puede especificar la moda, si utilizamos la distribución de frecuencias de la variable X, donde X denota el número de hijos de cada una de las 15 personas, pues al ser:

NÚMERO DE HIJOS	
x_i	n_i
0	5
1	6
2	3
4	1

basta observar cual es la mayor frecuencia, en este caso $n_2=6$, para poder afirmar que la moda es x_2, ésto es, que el valor que más se repite es 1 hijo.

Obviamente pueden darse tanto el supuesto de que todos los valores de la variable tengan frecuencia unitaria, en donde no tendría sentido la moda, como que haya más de un valor de la variable que tenga la mayor frecuencia, tal y como se propone en el siguiente supuesto.

Ejemplo 4.4.2. Supongamos que a las 15 personas del ejemplo anterior le añadimos otra que no tiene hijos. Entonces, la nueva distribución de frecuencias de la variable se puede especificar:

NÚMERO DE HIJOS	
x_i	n_i
0	6
1	6
2	3
4	1

por lo que en este caso existen dos valores de la variable, x_1 y x_2 que se presentan el mayor número de veces, 6, y así, podemos afirmar que en esta distribución existen dos modas, $M_o=0$ y $M_o=1$.

Las distribuciones que presentan una única moda se denominan ***unimodales***, las que tienen dos, ***bimodales***, las que tienen tres ***trimodales***, etc.

Ahora bien, cuando se considera una variable agrupada, al igual que con los promedios y la mediana, será imprescindible efectuar alguna *hipótesis adicional* para poder especificar la moda de la distribución, tal y como se va a contemplar a continuación.

Ejemplo 4.4.3. Consideremos la altura, en centímetros, de las 15 personas a las que se refiere el ejemplo 3.1.1. Puesto que los datos, tal y como se recogen en el mencionado ejemplo son:

168, 185, 160, 178, 193, 187, 172, 164, 175, 173, 195, 192, 166, 171 y 176,

podemos concluir que no hay moda, pues no hay ningún valor que aparezca más veces que otro.

Supongamos a continuación que únicamente se dispone de la distribución de frecuencias que se propone en el ejemplo 3.2.9, que venía dada por:

ALTURA			
$L_{i-1} - L_i$	a_i	x_i	n_i
160 - 165	5	162.5	2
165 - 170	5	167.5	2
170 - 175	5	172.5	4
175 - 180	5	177.5	2
180 - 185	5	182.5	1
185 - 190	5	187.5	1
190 - 195	5	192.5	3

Vamos a admitir que aquel intervalo que tenga una mayor frecuencia, esto es, aquel con un mayor número de observaciones, es el que contiene a la moda. Entonces, dado que en este caso, el intervalo en el que existe un mayor número de valores de la variable es (170-175], pues es el que tiene mayor frecuencia, suponemos que en dicho intervalo se encuentra la moda de la variable. Ahora bien, para fijar un valor concreto de la moda, siempre comprendida entre 170 y 175, debemos realizar alguna hipótesis adicional, al igual que cuando se han obtenido las expresiones correspondientes a los promedios y a la mediana. Dicha hipótesis podría ser tomar como moda el extremo inferior del intervalo, o el extremo superior, o su punto medio. Sin embargo, en este caso, vamos a establecer como criterio que el valor modal se encuentre más cerca de aquel intervalo, de los dos contiguos, que tenga mayor frecuencia, pues entendemos que dicho intervalo

ejerce una mayor "atracción" sobre el valor modal, pudiendo entonces ofrecer un valor aproximado de la moda de la variable cuando se encuentra agrupada.

En el supuesto que nos ocupa, la moda corresponde a la marca de clase del intervalo (170-175], pues las frecuencias de los dos intervalos contiguos a éste son iguales; esto es, $n_2=n_4=2$, y así, la atracción que ejercen dichos intervalos sobre el valor modal es idéntica, por lo que, $M_o=172.5$.

Ahora bien, si en lugar de la distribución de frecuencias que se acaba de ofrecer, se considera la que se propone en el ejemplo 3.2.4, que venía dada por:

ALTURA			
$L_{i-1} - L_i$	a_i	x_i	n_i
160 - 170	10	165	4
170 - 180	10	175	6
180 - 190	10	185	2
190 - 195	5	192.5	3

podríamos afirmar, en principio, que la moda se encuentra en el intervalo (170-180], dado que es el que presenta una mayor frecuencia. Ahora bien, dado que la amplitud de los intervalos no es constante, no debemos efectuar la comparación directa de sus frecuencias, pues si bien hay 6 observaciones comprendidas en el intervalo (170-180], también es cierto que hay 3 en el intervalo (190-195]. Evidentemente cabe esperar que un intervalo de mayor amplitud tenga mayor frecuencia, esto es, más valores de la variable, que uno que la tiene menor, sin que ello tenga que implicar que el valor que más se repite se encuentra en el primero de estos intervalos.

Para resolver este problema, esto es, para determinar cual es el intervalo en el que se encuentra la moda cuando la variable agrupada presenta intervalos con amplitud variable, vamos a recurrir a la densidad de frecuencia, y así, el intervalo modal será aquel, o aquellos, que tengan mayor densidad de frecuencia, esto es, aquel o aquellos, cuya altura en su correspondiente histograma sea mayor. De esta forma, para el ejemplo considerado, al existir intervalos con amplitudes distintas, deben calcularse las densidades de frecuencia, $d_i=n_i/a_i$, i=1,2,...,k, y seleccionar como intervalo donde se encuentra la moda, aquel cuya densidad de frecuencia sea mayor. Entonces, al ser la distribución de frecuencias:

ALTURA				
$L_{i-1} - L_i$	a_i	x_i	n_i	d_i
160 - 170	10	165	4	0.4
170 - 180	10	175	6	0.6
180 - 190	10	185	2	0.2
190 - 195	5	192.5	3	0.6

Podemos concluir que los intervalos con mayor densidad de frecuencia son dos, (170-180] y (190-195], por lo que esta distribución de frecuencias es bimodal.

El histograma correspondiente a esta distribución pone de manifiesto rápidamente su carácter bimodal, pues los rectángulos de mayor altura corresponden, obviamente, a los de mayor densidad de frecuencia, tal y como muestra la figura 4.4.1.

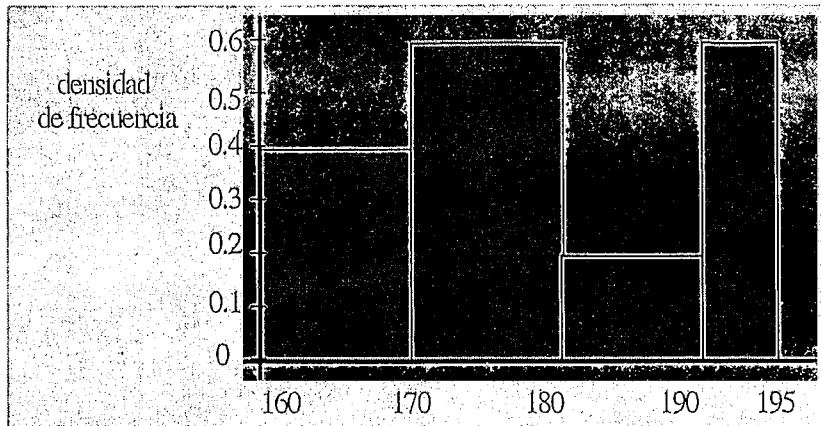

Figura 4.4.1: Histograma correspondiente al ejemplo 3.2.4.

El problema estriba ahora en calcular las dos modas; esto es, especificar cual es el valor del intervalo (170-180] y del intervalo (190-195] que se van a considerar como modas. En este sentido, por lo que respecta al primero de estos intervalos, la moda debe estar más cerca de 170 que de 180, pues la densidad de frecuencia del intervalo (160-170] es superior a la del intervalo (180-190], y así, el primero de estos intervalos ejerce una mayor atracción sobre el valor modal que el segundo. Lógicamente, sustituimos las densidades de frecuencias por las frecuencias a la hora de evaluar la atracción que ejercen los intervalos contiguos para la determinación de la moda. Por lo que respecta al otro valor modal, al ser el último intervalo, y no existir frecuencias en un intervalo superior, éste no ejerce ninguna influencia sobre el valor de la moda, por lo que dicho valor modal debe situarse en el extremo inferior del intervalo, siendo entonces M_o=190.

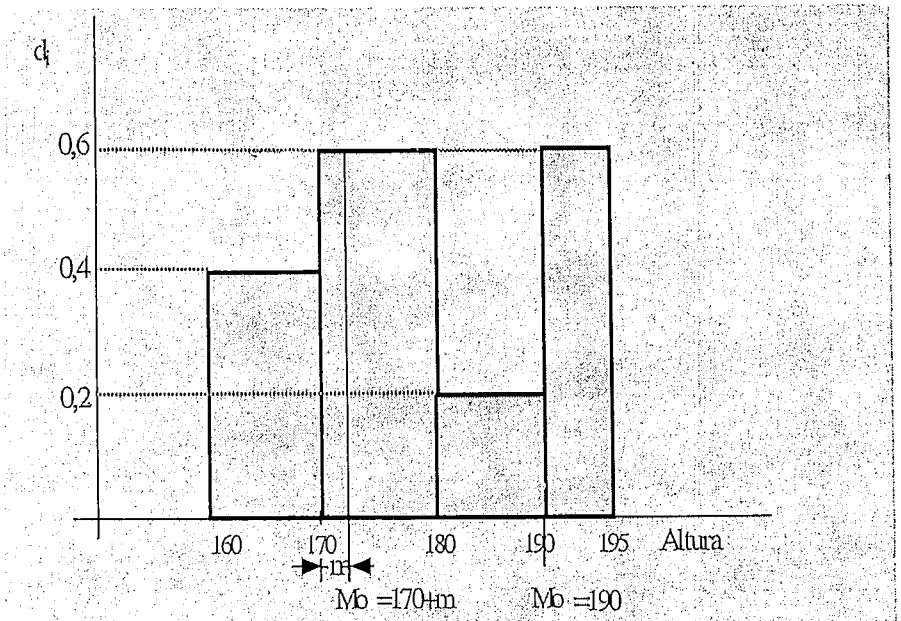

Figura 4.4.2: Determinación de las modas correspondientes al ejemplo 3.2.4

Evidentemente, el primer valor modal es M_o=170+m, tal y como muestra la figura 4.5.2, y puesto que la condición de que la moda debe estar más cerca de aquel intervalo contiguo con mayor densidad de frecuencia, se puede traducir en que la moda será aquel punto para el cual las distancias a los extremos inferior y superior del intervalo donde se encuentra la moda sean inversamente proporcionales a las densidades de frecuencia de dichos intervalos contiguos, podemos expresar:

$$\frac{m}{\frac{1}{0.4}} = \frac{10-m}{\frac{1}{0.2}}$$

y teniendo en cuenta las propiedades de sumas de antecedentes y consecuentes de las proporciones, se tiene

$$\frac{m}{\frac{1}{0.4}} = \frac{10-m}{\frac{1}{0.2}} = \frac{m+10-m}{\frac{1}{0.4}+\frac{1}{0.2}} = \frac{10}{\frac{1}{0.4}+\frac{1}{0.2}}$$

pudiendo expresar entonces,

$$\frac{m}{\frac{1}{0.4}} = \frac{10}{\frac{1}{0.4}+\frac{1}{0.2}}$$

por lo que

$$m = \frac{\frac{1}{0.4}}{\frac{1}{0.4}+\frac{1}{0.2}}10 = \frac{\frac{1}{0.4}}{\frac{0.2+0.4}{0.2\cdot 0.4}}10 = \frac{0.2}{0.2+0.4}10$$

y así, la moda que venimos queriendo determinar es:

$$M_o = 170 + m = 170 + \frac{0.2}{0.2+0.4}10 = 173.33$$

pudiendo concluir entonces que la distribución de frecuencias considerada tiene dos modas, que son precisamente 173.33 y 190 cm.

Obsérvese entonces, que en la determinación del valor de la moda de una variable agrupada, al igual que en los de los promedios y que en él de la mediana, interviene la forma en la que se agrupan los valores de la variable en los distintos intervalos.

Consideremos ahora una distribución genérica de la variable agrupada X, donde sólo existe un intervalo con mayor densidad de frecuencia, el intervalo i-ésimo, y obtengamos una expresión general

para la determinación de la moda. De esta forma, sea la siguiente distribución de frecuencias:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE AGRUPADA X				
L_0-L_1	a_1	x_1	n_1	d_1
L_1-L_2	a_2	x_2	n_2	d_2
...	...	...	...	...
L_{i-2}-L_{i-1}	a_{i-1}	x_{i-1}	n_{i-1}	d_{i-1}
L_{i-1}-L_i	a_i	x_i	n_i	d_i
L_i-L_{i+1}	a_{i+1}	x_{i+1}	n_{i+1}	d_{i+1}
...	...	...	...	...
L_{k-1}-L_k	a_k	x_k	n_k	d_k

donde $d_i>d_j$, j=1,2,..,i-1,i+1,...k. Bajo estos supuestos, y *considerando la* ***hipótesis*** *de que la moda se encuentra en el intervalo con mayor densidad de frecuencia*, la moda es un valor que se encuentra comprendido entre L_{i-1} y L_i, pues el intervalo i-ésimo, i=1,2,...,k, es el que presenta mayor densidad de frecuencia. Supongamos también, sin pérdida de generalidad, que $d_{i+1}<d_{i-1}$. De esta forma si representamos la parte del histograma de esta distribución correspondiente al intervalo modal, y a los intervalos anterior y posterior a éste, se obtiene la figura 4.4.3:

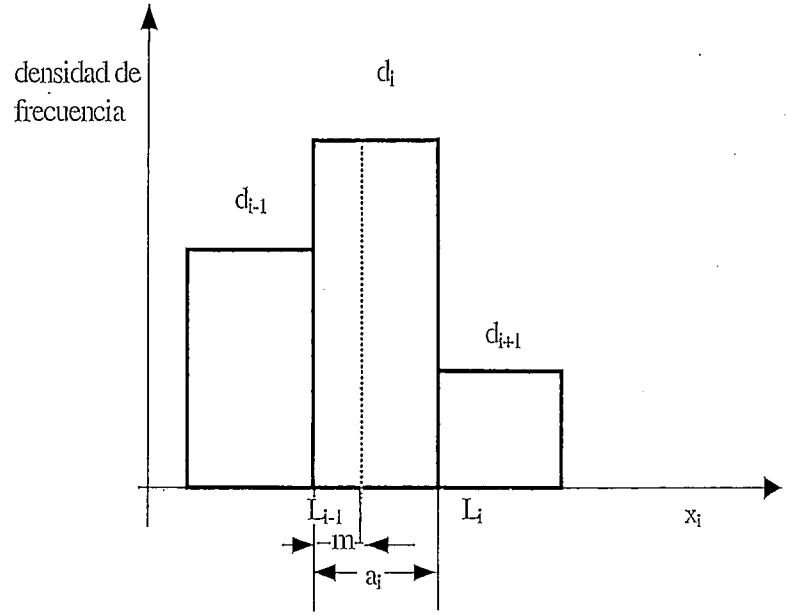

Figura 4.4.3: Determinación de la moda en distribuciones agrupadas en intervalos de distinta amplitud.

La moda será $M_o=L_{i-1}+m$, tal y como muestra la figura 4.4.3, y puesto que la condición de que *la moda debe estar más cerca de aquel intervalo contiguo con mayor densidad de frecuencia*, recordemos que se puede traducir en que la moda será aquel punto para el cual las distancias a los extremos inferior y superior del intervalo donde se encuentra la moda sean inversamente proporcionales a las densidades de frecuencia de dichos intervalos contiguos, podemos expresar:

$$\frac{m}{\frac{1}{d_{i-1}}}=\frac{a_i-m}{\frac{1}{d_{i+1}}}$$

y teniendo en cuenta las propiedades de sumas de antecedentes y consecuentes de las proporciones, se tiene

$$\frac{m}{\frac{1}{d_{i-1}}}=\frac{a_i-m}{\frac{1}{d_{i+1}}}=\frac{m+a_i-m}{\frac{1}{d_{i-1}}+\frac{1}{d_{i+1}}}=\frac{a_i}{\frac{1}{d_{i-1}}+\frac{1}{d_{i+1}}}$$

pudiendo expresar entonces,

$$\frac{m}{\frac{1}{d_{i-1}}}=\frac{a_i}{\frac{1}{d_{i-1}}+\frac{1}{d_{i+1}}}$$

por lo que

$$m=\frac{\frac{1}{d_{i-1}}}{\frac{1}{d_{i-1}}+\frac{1}{d_{i+1}}}a_i=\frac{\frac{1}{d_{i-1}}}{\frac{d_{i+1}+d_{i-1}}{d_{i+1}\cdot d_{i-1}}}a_i=\frac{d_{i+1}}{d_{i+1}+d_{i-1}}a_i$$

y así, la moda es:

$$M_o=L_{i-1}+m=L_{i-1}+\frac{d_{i+1}}{d_{i+1}+d_{i-1}}a_i$$

Ejemplo 4.5.4. Vamos a aplicar la expresión anterior a la determinación de la moda de los dos supuestos realizados en el ejemplo 4.5.3. Para el primero de ellos, dado que la distribución de frecuencias se puede especificar:

ALTURA				
$L_{i-1}-L_i$	a_i	x_i	n_i	d_i
160 - 165	5	162.5	2	0.4
165 - 170	5	167.5	2	0.4
170 - 175	5	172.5	4	0.8
175 - 180	5	177.5	2	0.4
180 - 185	5	182.5	1	0.2
185 - 190	5	187.5	1	0.2
190 - 195	5	192.5	3	0.6

se tiene que el intervalo con mayor densidad de frecuencia es el tercero, por lo que en el intervalo (170-175] se encuentra la moda de la distribución. Entonces,

$$M_o=L_{i-1}+\frac{d_{i+1}}{d_{i+1}+d_{i-1}}a_i=170+\frac{0.4}{0.4+0.4}5=170+2.5=172.5$$

y por lo tanto el valor modal es 172.5 cm.

Obsérvese que en este caso no sería necesario calcular las densidades de frecuencia, pues al tener todos los intervalos la misma amplitud, la mayor densidad de frecuencia corresponderá siempre a la mayor frecuencia, por lo que el intervalo en el que se encuentra la moda será el mismo independientemente de que se determine a través de la frecuencia o de la densidad de frecuencia; y además, dado que

$$M_o = L_{i-1} + \frac{\frac{n_{i+1}}{a_{i+1}}}{\frac{n_{i+1}}{a_{i+1}} + \frac{n_{i-1}}{a_{i-1}}} a_i = 170 + \frac{\frac{2}{5}}{\frac{2}{5} + \frac{2}{5}} 5$$

al ser $a_{i+1}=a_{i-1}=5$, podemos expresar:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i+1} + n_{i-1}} a_i = 170 + \frac{2}{2+2} 5 = 170 + 2.5 = 172.5$$

por lo que cuando los intervalos tienen la misma amplitud, puede reemplazarse, en la expresión de la moda de la variable X, las densidades de frecuencia de los intervalos contiguos al de la moda, por sus respectivas frecuencias.

Para la otra distribución en intervalos propuesta en el mencionado ejemplo, que venía dada por:

ALTURA				
$L_{i-1} - L_i$	a_i	x_i	n_i	d_i
160 - 170	10	165	4	0.4
170 - 180	10	175	6	0.6
180 - 190	10	185	2	0.2
190 - 195	5	192,5	3	0.6

se tiene que al ser $d_2=d_4=0.6$, la distribución es bimodal, encontrándose las modas en el segundo y en el último de los intervalos propuestos. De esta forma, para el primero de estos intervalos,

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i+1} + d_{i-1}} a_i = 170 + \frac{0.2}{0.2 + 0.4} 10 = 170 + 3.33 = 173.33$$

mientras que al considerar que $d_5=0$, pues sólo hay 4 intervalos,

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i+1} + d_{i-1}} a_i = 190 + \frac{0}{0 + 0.2} 10 = 190 + 0 = 190$$

por lo que las modas de esta distribución de frecuencias son 173.33 y 190 cm.

Obsérvese que en el supuesto de que el intervalo modal fuese el primero en lugar del último, al considerar que $d_0=0$, la moda se encontraría en el extremo superior del mencionado intervalo.

Obviamente, la moda, al igual que la mediana, al no obtenerse mediante un promedio de todos los valores de la variable, *no será* ***sensible*** *a la situación de que en su distribución de frecuencias se presenten valores anormalmente pequeños o grandes*. Sin embargo, cuando la variable se encuentra agrupada en intervalos, únicamente es posible especificar el valor o valores modales de forma aproximada, pudiendo tener una notable influencia la forma en la que se agrupen los valores de la variable en los distintos intervalos, tal y como se pone de manifiesto en el ejemplo 4.4.4.

También, debemos resaltar que la hipótesis que se ha utilizado en este epígrafe para determinar, cuando la variable ha sido agrupada, cual es el valor modal, no es tan comúnmente aceptada, como lo son las que se han hecho para el cálculo de los promedios, marca de clase, o de la mediana, esto es el reparto uniforme. En este sentido, se han propuesto *otras alternativas*. Así, por ejemplo, se puede entender que

una vez identificado el intervalo o intervalos en los que se encuentra la moda, tomemos como valor modal aquel tal que sus distancias a los intervalos contiguos sean directamente proporcionales a las diferencias de altura entre el modal y dichos contiguos, que conduce a la siguiente expresión:

$$M_o = L_{i-1} + \frac{d_i - d_{i-1}}{(d_i - d_{i-1}) + (d_i - d_{i+1})} a_i$$

De cualquier forma, como todas ellas son aproximaciones, y puesto que no existe un criterio unificado para su cálculo, salvo que se disponga de los valores originales o de la distribución de frecuencias de una variable no agrupada, no creemos conveniente detenernos más en su estudio.

4.5 Medidas de posición no central: los cuantiles.

En los epígrafes anteriores, se ha intentado de alguna forma buscar el centro de la distribución de frecuencias de la variable de interés, y más concretamente, la mediana era aquel valor de la variable tal que, una vez ordenados todos sus valores, ocupaba el lugar central. Puede resultar interesante, en lugar de buscar el valor de la variable que divide la distribución en dos partes de forma que en cada una de ellas se encuentre el 50% de las observaciones, dividirla en un conjunto de partes, de forma que en cada una de ellas haya el mismo número de observaciones, como por ejemplo el 10%, o el 20%, etc.

En este sentido, *una vez ordenados los valores de la variable en sentido creciente, a los valores que dividen en partes en las que haya el mismo número de observaciones, se les denomina* ***cuantiles*** *o* ***cuantilas***, recibiendo nombres específicos según el número de partes en que lo hacen; cuartiles, si son 4, deciles, si son 10 y percentiles o centiles, si son 100.

Obsérvese que los requisitos que se imponen para su determinación son los mismos que los que se especificaron para calcular la mediana, salvo que en ésta se consideraban dos partes con el mismo número de observaciones, y por lo tanto, la determinación de cualquier cuantil debe encontrarse íntimamente relacionada con la de la mediana. A continuación vamos a considerar los cuantiles más significativos.

A) **Los *cuartiles* son aquellos valores de la variable, tal que una vez ordenados todos sus valores en sentido creciente, los dividen en cuatro partes en los que haya el mismo número de observaciones.** Así, el primer cuartil Q_1, es aquel valor de la variable tal que el 25% de las observaciones son inferiores a él, y por tanto, el 75% restante son superiores. El segundo cuartil, Q_2, es aquel valor de la variable tal que el 50% de las observaciones son inferiores a él, siendo entonces el 50% restante superior; mientras que el tercer cuartil, Q_3, es aquel valor de la variable tal que el 75% de las observaciones son inferiores a él, y por tanto, el 25% restante son superiores.

Para su determinación, debemos distinguir si la variable de interés se encuentra agrupada o no. En el caso de que no lo esté, puede utilizarse la regla de actuación ofrecida para el cálculo de la mediana, con la modificación pertinente según el cuartil que se pretenda obtener. Así, *para el primer cuartil*:

1. Ordenar los valores de la variable.
2. Calcular las frecuencias acumuladas N_i, i=1,2,...,k.
3. Obtener el valor de N/4.
4. Si no hay ningún valor de N_i que coincida con N/4, determinar el primero tal que N_i>N/4, y si éste es N_h, entonces $Q_1=x_h$.
5. Si hay un valor de N_i que coincida con N/4, si éste es N_h, entonces la hipótesis considerada es que $Q_1=(x_h+x_{h+1})/2$.

En el caso de que se pretenda determinar el segundo cuartil, en lugar de obtener el valor de N/4 habrá que calcular 2N/4, esto es, N/2, y por lo tanto dicho cuartil coincide con la mediana. Para el tercer cuartil, debemos obtener el valor de 3N/4, y a continuación considerar los pasos 4 y 5 de la regla de actuación anterior.

Ejemplo 4.5.1. Consideremos el número de hijos de cada una de las 15 personas a las que se refiere el ejemplo 3.1.2, cuya distribución de frecuencias se puede especificar:

NÚMERO DE HIJOS		
x_i	n_i	N_i
0	5	5
1	6	11
2	3	14
4	1	15

Para el cálculo del primer cuartil, dado que N/4=15/4=3.75, es imposible al no ser N/4 entero que exista algún N_i, i=1,2,3,4, que coincida con él, y por lo tanto únicamente debemos buscar cual es la primera frecuencia acumulada que sobrepasa a 3.75; como quiera que ésta es N_1=5, se tiene que $Q_1=x_1$, y por lo tanto el primer cuartil es 0 hijos.

Por lo que respecta al segundo cuartil, al ser $Q_2=M_e$, se tiene que Q_2=1, tal y como se obtuvo en el ejemplo 4.4.5, y así el segundo cuartil es 1 hijo.

Para determinar el valor de Q_3, dado que 3N/4=45/4=11.25, es imposible nuevamente al no ser 3N/4 entero que exista algún N_i, i=1,2,3,4, que coincida con, él, y por lo tanto, dado que la primera frecuencia acumulada que sobrepasa a este valor es N_3=14, se tiene que $Q_3=x_3$, y por lo tanto el tercer cuartil es 2 hijos.

Así pues, los tres cuartiles de esta distribución de frecuencias son 0, 1 y 2 hijos.

Obsérvese que en ese caso, pudiera darse la circunstancia de que dos, o incluso los tres cuartiles alcanzaran el mismo valor, pues si para este supuesto sustituimos una persona que tiene 2 hijos por otra que sólo tiene 1, se puede comprobar que entonces $Q_2=Q_3$=1.

De esta forma, podemos ofrecer la siguiente regla de actuación general para calcular el cuartil r-ésimo, r=1,2,3, en aquellas distribuciones de frecuencias cuya variable no se encuentra agrupada:

1. Ordenar los valores de la variable.
2. Calcular las frecuencias acumuladas N_i, i=1,2,...,k.
3. Obtener el valor de rN/4, r=1,2,3.
4. Si no hay ningún valor de N_i que coincida con rN/4, determinar el primero tal que N_i>rN/4, y si éste es N_h, entonces $Q_r=x_h$.
5. Si hay un valor de N_i que coincida con rN/4, si éste es N_h, entonces la ***hipótesis*** considerada es que $Q_r=(x_h+x_{h+1})/2$.

En el supuesto de que se traten de determinar los cuartiles de una distribución de frecuencias cuya variable se encuentra agrupada, hemos de realizar la misma ***hipótesis*** que la que se consideró al calcular la mediana, y recordemos que ésta es el *reparto uniforme de las frecuencias dentro del intervalo en el que se encuentra el cuartil considerado.* Supuesto que éste sea el intervalo i-ésimo, i=1,2,...,k, tal hipótesis conduciría a la siguiente expresión:

$$Q_r = L_{i-1} + \frac{\frac{rN}{4} - N_{i-1}}{N_i - N_{i-1}} a_i = L_{i-1} + \frac{\frac{rN}{4} - N_{i-1}}{n_i} a_i \quad r = 1,2,3$$

donde es necesario, obviamente, que el intervalo en él que se encuentra el cuartil esté bien definido. Además, en todo caso $Q_2=M_e$.

Ejemplo 4.5.2. Consideremos la altura, en centímetros, de cada una de las 15 personas a las que se refiere el ejemplo 3.1.1, cuya distribución de frecuencias se puede especificar, considerando la agrupación realizada en el ejemplo 3.2.4:

ALTURA				
$L_{i-1} - L_i$	a_i	x_i	n_i	N_i
160 - 170	10	165	4	4
170 - 180	10	175	6	10
180 - 190	10	185	2	12
190 - 195	5	192.5	3	15

La aplicación de la expresión anterior a la determinación de los tres cuartiles de esta distribución de frecuencias conduce a los siguientes resultados:

$$Q_1 = L_{i-1} + \frac{\frac{1\cdot 15}{4} - N_{i-1}}{n_i} a_i = 160 + \frac{3.75 - 0}{4} 10 = 169.375$$

pues $N_0=0$,

$$Q_2 = M_e = L_{i-1} + \frac{\frac{2\cdot 15}{4} - N_{i-1}}{n_i} a_i = 175.83$$

tal y como se obtuvo en el ejemplo 4.3.6, y

$$Q_3 = L_{i-1} + \frac{\frac{3\cdot 15}{4} - N_{i-1}}{n_i} a_i = 180 + \frac{11.25 - 10}{2} 10 = 186.25$$

por lo que podemos concluir que los tres cuartiles son 169.375, 175.83 y 186.25 cm. Nótese que para calcular un determinado cuartil, es imprescindible que el intervalo en el que se encuentra esté completamente especificado, pudiendo no estarlo el resto de los mismos.

Téngase en cuenta que cuando se considera una distribución de frecuencias de una variable agrupada, *es imposible que dos cuartiles puedan tomar el mismo valor*, a diferencia de cuando la variable no se encuentra agrupada. Ahora bien, la agrupación en intervalos, puede conducir, cuando se varía, a que se puedan obtener otros valores, y así, la forma en la que se agrupe una variable influye en los valores de los cuartiles.

B) **Los *deciles* son aquellos valores de la variable, tal que una vez ordenados todos sus valores en sentido creciente, los dividen en diez partes en los que haya el mismo número de observaciones.** Así, el primer decil D_1, es aquel valor de la variable tal que el 10% de las observaciones son inferiores a él, y por tanto, el 90% restante son superiores. El segundo decil, D_2, es aquel valor de la variable tal que el 20% de las observaciones son inferiores a él, siendo entonces el 80% restante superior, ..., y el noveno decil, D_9, es aquel valor de la variable tal que el 90% de las observaciones son inferiores a él, y por tanto, el 10% restante son superiores.

En el caso de que la variable no se encuentre agrupada, puede utilizarse la siguiente regla de actuación para calcular el decil r-ésimo, r=1,2,...,9:

1. Ordenar los valores de la variable.
2. Calcular las frecuencias acumuladas N_i, i=1,2,...,k.
3. Obtener el valor de rN/10, r=1,2,...,9.
4. Si no hay ningún valor de N_i que coincida con rN/10, determinar el primero tal que N_i>rN/10, y si éste es N_h, entonces $D_r=x_h$.
5. Si hay un valor de N_i que coincida con rN/10, si éste es N_h, entonces la ***hipótesis*** considerada es que $D_r=(x_h+x_{h+1})/2$.

En ese caso, es posible que dos o más deciles *alcancen el mismo valor*. Sin embargo, esta situación es imposible cuando la variable se encuentra agrupada, en cuyo caso, dichos valores pueden depender de la forma en la que se hayan agrupado, pudiéndose especificar, bajo la ***hipótesis*** del reparto uniforme de las frecuencias dentro del intervalo considerado, según la siguiente expresión:

$$D_r = L_{i-1} + \frac{\frac{rN}{10} - N_{i-1}}{N_i - N_{i-1}} a_i = L_{i-1} + \frac{\frac{rN}{10} - N_{i-1}}{n_i} a_i \quad r = 1,2,...,9$$

donde es necesario, obviamente, que el intervalo en el que se encuentra el decil se encuentre bien definido, coincidiendo siempre el quinto decil con la mediana.

C) Los ***percentiles*** o ***centiles*** **son aquellos valores de la variable, tal que una vez ordenados todos sus valores en sentido creciente, los dividen en cien partes en los que haya el mismo número de observaciones.** Así, el primer percentil P_1, es aquel valor

de la variable tal que el 1% de las observaciones son inferiores a él, y por tanto, el 99% restante son superiores. El segundo percentil, P_2, es aquel valor de la variable tal que el 2% de las observaciones son inferiores a él, siendo entonces el 98% restante superior, ..., y el percentil noventa y nueve, P_{99}, es aquel valor de la variable tal que el 99% de las observaciones son inferiores a él, y por tanto, el 1% restante son superiores.

En el caso de que la variable no se encuentre agrupada, puede utilizarse la siguiente regla de actuación para calcular el percentil r-ésimo, r=1,2,...,99:

1. Ordenar los valores de la variable.
2. Calcular las frecuencias acumuladas N_i, i=1,2,...,k.
3. Obtener el valor de rN/100, r=1,2,...,99.
4. Si no hay ningún valor de N_i que coincida con rN/100, determinar el primero tal que N_i>rN/100, y si éste es N_h, entonces $P_r=x_h$.
5. Si hay un valor de N_i que coincida con rN/100, si éste es N_h, entonces la ***hipótesis*** considerada es que $P_r=(x_h+x_{h+1})/2$.

En ese caso, es posible que dos o incluso un número bastante elevado de percentiles *alcancen el mismo valor*. Sin embargo, esta situación es imposible cuando la variable se encuentra agrupada, en cuyo caso, dichos valores pueden depender de la forma en la que se hayan agrupado, pudiéndose especificar, bajo la ***hipótesis*** del reparto uniforme de las frecuencias dentro del intervalo considerado, según la siguiente expresión:

$$P_r = L_{i-1} + \frac{\frac{rN}{100} - N_{i-1}}{N_i - N_{i-1}} a_i = L_{i-1} + \frac{\frac{rN}{100} - N_{i-1}}{n_i} a_i \quad r = 1,2,...,99$$

donde es necesario, obviamente, que el intervalo en el que se encuentra el percentil se encuentre bien definido, siendo en todo caso el percentil de orden 50 igual a la mediana. Evidentemente, existe un numeroso conjunto de relaciones entre los distintos cuantiles que se han considerado. Así, a título de ejemplo podemos citar que $P_{10}=D_1$, $P_{25}=Q_1$, ..., y naturalmente, $P_{50}=D_5=Q_2=M_e$.

Ejemplo 4.5.3. Consideremos la altura, en centímetros, de cada una de las 15 personas a las que se refiere el ejemplo 4.5.2, cuya distribución de frecuencias recordemos que viene dada por:

ALTURA				
L_{i-1} - L_i	a_i	x_i	n_i	N_i
160 - 170	10	165	4	4
170 - 180	10	175	6	10
180 - 190	10	185	2	12
190 - 195	5	192.5	3	15

Vamos a calcular cuál es la altura más pequeña del 15% de las personas más altas, y cuál es la altura más grande del 30% de las personas más bajas. En ambos casos, se nos está solicitando un determinado cuantil, que será el percentil 85, para el primero de los supuestos; y el percentil treinta o el tercer decil, para el segundo. Así, utilizando la hipótesis de que las frecuencias se reparten de forma uniforme en aquellos intervalos en los que sea necesario, los cuantiles solicitados serían:

$$P_{85}=L_{i-1}+\frac{\frac{85\cdot 15}{100}-N_{i-1}}{n_i}a_i=190+\frac{12.75-12}{3}5=190+1.25=191.25$$

$$P_{30}=L_{i-1}+\frac{\frac{30\cdot 15}{100}-N_{i-1}}{n_i}a_i=170+\frac{4.5-4}{6}10=170+0.83=170.83$$

y así, la altura más pequeña correspondiente al 15% de las personas más altas es 191.25 cm., mientras que la altura más grande del 30% de las personas más bajas es precisamente 170.83 cm.

4.6 Atributos y medidas de posición.

Que la característica de interés no sea cuantitativa limita de forma notable la posibilidad de utilizar las medidas de posición. Así, los **promedios**, no pueden obtenerse cuando consideramos un atributo, pues las modalidades no son valores, y por lo tanto no se pueden sumar. Sin embargo, sí que parece posible emplear la mediana y la moda, así como los cuantiles.

En este sentido, un aspecto que determina la medida de tendencia central que se puede utilizar es la escala de medida en la que venga dada la característica cualitativa. De esta forma, si ésta es **ordinal**, puede utilizarse tanto la mediana como la moda, pues si existe una ordenación entre las distintas modalidades, puede buscarse cuál de ellas es la que forma dos grupos en los que haya el mismo número de observaciones. Además, la posibilidad de asignación numérica de las modalidades, llevaría aparejada el posible cálculo de los promedios, aunque éste es un aspecto en el que debe obrarse con mucha cautela.

Sin embargo, si la escala de medida es **nominal**, únicamente tiene sentido determinar cual es la modalidad que más veces se presenta, y por lo tanto sólo puede especificarse la moda del atributo.

Obviamente, cuando se pueda utilizar la mediana, pueden considerarse los distintos **cuantiles**.

Ejemplo 4.6.1. Consideremos los supuestos especificados en el ejemplo 3.1.7, que hacía referencia a la religión que les hubiera gustado tener a un conjunto de 15 personas, cuya distribución de frecuencias era:

RELIGION	
m_i	n_i
Católico	12
Testigo de Jehová	2
Protestante	1

En este caso, puesto que este atributo no admite la ordenación de sus modalidades, únicamente es posible determinar cuál es la que tiene mayor frecuencia, y así, la moda de dicho atributo es "Católico".

Ejemplo 4.6.2. Supongamos que el nivel socio-económico de un conjunto de 500 familias es el que se ofrece en la siguiente distribución de frecuencias:

NIVEL SOCIO ECONÓMICO		
m_i	n_i	N_i
Bajo	45	45
Medio-bajo	200	245
Medio-medio	125	370
Medio-alto	95	465
Alto	35	500

En este caso, además de poder ofrecer la modalidad que con mayor frecuencia se presenta, siendo ésta la "Medio-bajo", al admitir este atributo la ordenación de sus modalidades, puede determinarse

cuál es la que forma dos grupos en los que hay el mismo número de familias, siendo entonces la modalidad "Medio-medio", la mediana de esta distribución de frecuencias.

Se propone al lector como ejercicio que compruebe que en la modalidad "Medio-alto" constituye el percentil 80 de dicha distribución de frecuencias.

RESUMEN

Las medidas de posición se clasifican en centrales y no centrales. Las primeras, tratan de poner de manifiesto, mediante algún criterio, cuál es el centro de una distribución de frecuencias, con objeto, básicamente, de representarla. En el supuesto de que la característica sea cuantitativa, el promedio por excelencia es la media aritmética, y así, para la variable X, ésta es:

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

que viene expresada en las mismas unidades en que lo haga la variable X. Si trabajamos con distribuciones agrupadas, caso en el que x_i son las marcas de clase, $\overline{x}$ es un valor aproximado, pues supone que las frecuencias de cada intervalo se concentran en su marca de clase. Del conjunto de propiedades de este promedio destaca que la suma de diferencias de cada valor de la variable a la media aritmética es nula, y que los cambios de origen y/o escala en la variable afectan al promedio. También, presenta notable interés la media aritmética ponderada, destacando en este sentido el caso particular de la composición de poblaciones.

Otras medidas de posición central son la mediana, M_e, y la moda, M_o. La primera representa el valor de la variable que ocupa el lugar central de la distribución de frecuencias, una vez han sido ordenados los mismos, mientras que la segunda indica el valor de la variable que más se repite. En el supuesto de que la variable se presente agrupada en intervalos, debe realizarse alguna hipótesis sobre "el reparto" de las frecuencias en todos o en algún intervalo, para poder determinar, de forma aproximada el valor de la mediana y de la moda o modas. Así, bajo el supuesto de que la frecuencias se reparten de forma uniforme en el intervalo (i-ésimo) en el que se encuentra la mediana, ésta se puede aproximar mediante la expresión:

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

De forma análoga, si se considera que en el intervalo (i-ésimo) donde hay mayor densidad de frecuencia, es aquel en el que se encuentra la moda, entonces, entre otras fórmulas, puede emplearse como aproximación a su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i+1} + d_{i-1}} a_i$$

De entre las medidas de posición no central o cuantiles, valores de la distribución que dividen la misma en diversas partes iguales, cabe destacar los cuartiles, los deciles y los percentiles, Q_r r=1,2,3; D_r r=1,...,9; y P_r r=1,...,99, respectivamente, cuyos valores o aproximaciones se obtienen de forma similar a la mediana, salvo lógicamente, por el porcentaje de observaciones de deja a cada lado.

Por último, debemos señalar que si la característica objeto de estudio es un atributo, según la escala de medida utilizada podrá determinarse la modalidad mediana, y por lo tanto los cuantiles, y la modal, si dicha escala es ordinal, o únicamente la modalidad modal, en el caso de que la escala sea nominal.

CAPÍTULO 5

MEDIDAS DE DISPERSIÓN Y TRANSFORMACIONES

5.1 La dispersión y su medida.

Además de las medidas de posición existen otras, denominadas de ***dispersión,*** que intentan mostrar la variabilidad existente en torno a los distintos valores de una determinada distribución de frecuencias. Una primera aproximación a esta idea podría venir derivada de la diferencia entre el valor más pequeño y el más grande que toma la variable. Así, buscando una correspondencia entre variabilidad y distancia, dicha diferencia trata de determinar el grado de variación de la distribución. Pues bien, a esta diferencia la vamos a denominar ***recorrido*** o ***rango***, R, y así:

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

donde x_{max} denota el valor ***mayor de la variable*** X, y x_{min} el ***menor de la variable.***

Ejemplo 5.1.1. Consideremos el número de hijos de cada una de las 15 personas a las que se refiere el ejemplo 3.1.2, donde si X representa el número de hijos de cada una de ellas, su distribución de frecuencias se puede expresar;

NÚMERO DE HIJOS	
x_i	n_i
0	5
1	6
2	3
4	1

En este caso, el recorrido de la variable es 4 hijos, pues

$$R = x_{max} - x_{mín} = 4 - 0 = 4$$

Dada la expresión del recorrido, es obvio que dicha medida es enormemente ***sensible*** *ante la presencia de valores anormalmente grandes o pequeños en la distribución de frecuencias*, pues son precisamente los valores extremos los únicos que se consideran para su determinación.

Ejemplo 5.1.2. Consideremos ahora la altura, en centímetros, de cada una de las 15 personas a que se refiere el ejemplo 3.1.1, donde si X representa la altura, en centímetros, de cada una de ellas, su distribución de frecuencias agrupada, tal y como se propone en el ejemplo 3.2.9, es:

ALTURA			
$L_{i-1} - L_i$	a_i	x_i	n_i
160 - 165	5	162.5	2
165 - 170	5	167.5	2
170 - 175	5	172.5	4
175 - 180	5	177.5	2
180 - 185	5	182.5	1
185 - 190	5	187.5	1
190 - 195	5	192.5	3

Para este supuesto, el recorrido de la variable es 35 cm., pues

$$R = x_{max} - x_{mín} = 195 - 160 = 35$$

En el caso de que la distribución de frecuencias de una variable agrupada no tenga bien definido el primero y/o el último de los intervalos, tal y como se plantea en el ejemplo 3.2.3, cuya distribución de frecuencias se especificaba:

ALTURA	
Menos de 170	4 personas
Más de 170 pero menos de 180	6 personas
Más de 180 pero menos de 190	2 personas
Más de 190	3 personas

no es posible calcular el recorrido. Para resolver esta situación, puede considerarse el empleo de medidas alternativas, como el ***recorrido intercuartílico*** $R_Q = Q_3 - Q_1$, que se podrá especificar siempre y cuando el primer y el tercer cuartil se encuentren en intervalos completamente especificados; la diferencia entre el percentil 90 y el percentil 10, siempre y cuando dichos percentiles se encuentren en intervalos completamente especificados, etc.

Tal y como hemos podido comprobar en los ejemplos anteriores, el recorrido es una medida de dispersión que viene expresada en las mismas unidades que presenta la variable objeto de estudio. Pues bien, en estos casos, decimos que dicha medida de dispersión es absoluta, y así, una ***medida de dispersión absoluta*** viene expresada en la misma unidad de medida que la variable.

Ahora bien, si quisiéramos saber cual de las dos distribuciones propuestas en los ejemplos anteriores presenta un mayor grado de variabilidad, determinada por el recorrido, deberíamos buscar una medida que fuese adimensional, esto es, que no viniera expresada en ninguna unidad de medida. Así, a las medidas de dispersión que no tienen unidad de medida se denominan ***medidas de dispersión relativas***. En este sentido, podríamos emplear el ***coeficiente de apertura*** o de disparidad, A, cuya expresión es:

$$A = \frac{x_{max}}{x_{min}}$$

que al definirse a través de un cociente de valores de la variable, no presenta unidades de medida, y por lo tanto es una medida de dispersión relativa.

Así, para los dos ejemplos propuestos anteriormente, se tiene que dicho coeficiente quedaría indeterminado para el primero de ellos, pues

$$A = \frac{4}{0}$$

mientras que sería 1.21875 para el segundo, dado que

$$A = \frac{195}{160} = 1.21875$$

que viene a indicar que para esta distribución de frecuencias, el valor más grande que toma la variable es 1.21875 veces mayor que el valor más pequeño que toma la misma. Así pues, cuando el valor más pequeño de la variable es 0, no puede calcularse el coeficiente de apertura, situación que puede presentarse con cierta asiduidad.

Además de los inconvenientes reseñados, tanto el recorrido como el coeficiente de apertura parecen indicadores muy burdos de la variabilidad de la distribución de frecuencias, tal y como se pone de manifiesto en el siguiente supuesto.

Ejemplo 5.1.3. Supongamos la siguiente distribución de frecuencias correspondiente a la altura de 15 personas:

ALTURA			
$L_{i-1} - L_i$	a_i	x_i	n_i
160 - 165	5	162.5	1
165 - 170	5	167.5	1
170 - 175	5	172.5	1
175 - 180	5	177.5	9
180 - 185	5	182.5	1
185 - 190	5	187.5	1
190 - 195	5	192.5	1

Podemos observar que en este caso, R=195-160=35 y A=195/160=1.21875, valores idénticos a los obtenidos al considerar la distribución de frecuencias planteada en el ejemplo 5.1.2. La diferencia entre ambas distribuciones estriba en el número de observaciones de los intervalos, esto es, en las frecuencias.

Fácilmente podemos concluir que la distribución de frecuencias propuesta en el presente ejemplo presenta un menor grado de variabilidad, pues la mayoría de los valores de la variable se encuentran incluidos en el intervalo (175-180], mientras que en la propuesta en el ejemplo 5.1.2, que recordemos era:

ALTURA			
$L_{i-1} - L_i$	a_i	x_i	n_i
160 - 165	5	162.5	2
165 - 170	5	167.5	2
170 - 175	5	172.5	4
175 - 180	5	177.5	2
180 - 185	5	182.5	1
185 - 190	5	187.5	1
190 - 195	5	192.5	3

los valores se encuentran más repartidos entre los distintos intervalos, por lo que existe más dispersión, aunque ni el recorrido ni el coeficiente de apertura son capaces de ponerlo de manifiesto.

Por todo ello, parece necesario buscar medidas de dispersión a través de otros caminos. En este sentido, debemos recordar que habitualmente se utiliza una medida de tendencia central para resumir en un único valor todos los que conforman la distribución de frecuencias. Así, supongamos que tres personas miden 169, 170 y 171 cm., mientras que otras tres 160, 170 y 180 cm. La altura media en ambos casos es 170 cm., pero la representatividad de esta media aritmética no es la misma en ambos casos, pues los valores 169 y 171 se encuentran próximos a la media aritmética, mientras que 160 y 180 están mucho más alejados, apareciendo entonces la idea de "*discrepancia, diferencia, distancia*" entre los valores de la variable y su media aritmética.

Así, supongamos que una determinada medida de tendencia central de una distribución de frecuencias es c. Entonces, podríamos

considerar alguna medida de la variabilidad que presenta c, mostrando si los valores que toma la variable están "alejados" o si, por el contrario, están "próximos" a c, evaluando de esta forma la discrepancia, diferencia o distancia existente entre c y el conjunto de valores que toma la variable. De esta forma, si dicha distancia es pequeña, es porque todos los valores de la variable están próximos a c, por lo que esta medida de tendencia central representa adecuadamente a la distribución de frecuencias, mientras que si dicha distancia es grande, será porque los valores de la variable se separan notablemente de c, por lo que ésta no resume de forma adecuada al conjunto de valores de la variable.

Por lo tanto, las medidas de dispersión que vamos a considerar a partir de este momento, van a tratar de mostrar la representatividad de una medida de posición central, mediante la separación o distancia entre dicha medida y el conjunto de valores que toma la variable. *A mayor distancia o separación, mayor dispersión presenta la distribución de frecuencias, y por lo tanto, menor es la representatividad de la medida de tendencia central sobre la que se calcula dicha distancia.*

Recordemos que las medidas de dispersión, tal y como se ha contemplado anteriormente, pueden ser absolutas, si vienen expresadas en una determinada unidad de medida, y relativas, en el caso de que no vengan expresadas en unidad de medida alguna.

5.2 Medidas de dispersión absolutas.

Consideremos una medida de tendencia central cualquiera c. Para valorar, en término medio, la dispersión (distancia o separación) que existe entre dicha medida de tendencia central y el conjunto de valores que toma la variable de interés, podrían utilizarse las siguientes medidas de dispersión:

$$D_1 = \frac{\sum_{i=1}^{k} |x_i - c| n_i}{N}$$

o bien

$$D_2 = \frac{\sum_{i=1}^{k} (x_i - c)^2 n_i}{N}$$

pues debemos buscar una medida de dispersión que evite que diferencias de signo positivo se compensen con diferencias de signo negativo.

En este sentido, y puesto que en el Apéndice demostramos que el mínimo de la medida de dispersión

$$D_2 = \frac{\sum_{i=1}^{k} |x_i - c| n_i}{N}$$

se alcanza cuando $c=M_e$, se tiene que **la mediana de una distribución de frecuencias tiene como medida de dispersión**

$$D_{M_e} = \frac{\sum_{i=1}^{k} |x_i - M_e| n_i}{N}$$

denominada ***desviación media respecto de la mediana***, o desviación absoluta media respecto de la mediana.

Ejemplo 5.2.1. Consideremos la distribución de frecuencias del número de hijos de las 15 personas a que se refiere el ejemplo 5.1.1, que recordemos venía dada por la siguiente tabla:

NÚMERO DE HIJOS	
x_i	n_i
0	5
1	6
2	3
4	1

La media aritmética de esta distribución de frecuencias es 16/15 hijos, mientras que el valor mediano es 1 hijo, tal y como se obtuvo a través de los ejemplos 4.2.1 y 4.4.1.

Pues bien, si procedemos a calcular las diferencias, en valor absoluto, de cada uno de los valores de la variable a cada una de estas medidas de tendencia central, se tiene,

NÚMERO DE HIJOS			
x_i	n_i	$\lvert x_i - \bar{x} \rvert n_i$	$\lvert x_i - M_e \rvert n_i$
0	5	80/15	5
1	6	6/15	0
2	3	42/15	3
4	1	44/15	3
TOTALES	15	172/15	11

De esta forma,

$$\frac{\sum_{i=1}^{4}\left|x_i-\frac{16}{15}\right|n_i}{15}=\frac{\frac{172}{15}}{15}=0.764$$

mientras que

$$D_{M_e}=\frac{\sum_{i=1}^{4}|x_i-1|n_i}{15}=\frac{11}{15}=0.733$$

por lo que para el ejemplo planteado

$$0.764=\frac{\sum_{i=1}^{4}|x_i-\bar{x}|n_i}{N}>\frac{\sum_{i=1}^{4}|x_i-M_e|n_i}{N}=0.733$$

Si ahora consideramos la segunda de las medidas de dispersión propuestas, y puesto que en el Apéndice también demostramos que el mínimo de la medida de dispersión

$$D_2=\frac{\sum_{i=1}^{k}(x_i-c)^2 n_i}{N}$$

se produce cuando $c=\bar{x}$, resultado conocido como ***Teorema de Köning***, se tiene que la **media aritmética de una distribución de frecuencias tiene como medida de dispersión**

$$s^2=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{N}$$

denominada ***varianza*** o variancia.

Ejemplo 5.2.4. Consideremos de nuevo el número de hijos de cada una de las 15 personas a que se refiere el ejemplo 5.1.1, cuya distribución de frecuencias venía dada por:

NÚMERO DE HIJOS	
x_i	n_i
0	5
1	6
2	3
4	1

Dado que la media aritmética de esta distribución de frecuencias es 16/15 hijos, mientras que el valor mediano es 1 hijo, tal y como se ha especificado en el ejemplo 5.2.1, se tiene que

NÚMERO DE HIJOS			
x_i	n_i	$(x_i-\bar{x})^2 n_i$	$(x_i-M_e)^2 n_i$
0	5	$1280/15^2$	5
1	6	$6/15^2$	0
2	3	$588/15^2$	3
4	1	$1936/15^2$	9
TOTALES	15	$3810/15^2$	17

De esta forma,

$$s^2=\frac{\sum_{i=1}^{4}\left(x_i-\frac{16}{15}\right)^2 n_i}{15}=\frac{\frac{3810}{15^2}}{15}=1.129$$

mientras que

$$\frac{\sum_{i=1}^{4}(x_i - 1)^2 n_i}{15} = \frac{17}{15} = 1.133$$

por lo que para el ejemplo planteado

$$1.129 = \frac{\sum_{i=1}^{4}(x_i - \bar{x})^2 n_i}{N} < \frac{\sum_{i=1}^{4}(x_i - M_e)^2 n_i}{N} = 1.133$$

pues al ser la varianza una medida de dispersión óptima para la media aritmética, sería imposible que

$$\frac{\sum_{i=1}^{4}(x_i - \bar{x})^2 n_i}{N} > \frac{\sum_{i=1}^{4}(x_i - M_e)^2 n_i}{N}$$

Obsérvese nuevamente que si la variable se encuentra agrupada, entonces el valor de la varianza, al igual que el de la media aritmética, es aproximado, pues al utilizar las marcas de clase en lugar de los auténticos valores de la variable, se está suponiendo que todos los valores de un intervalo se agrupan en su marca de clase.

Ejemplo 5.2.5. Consideremos la altura, en centímetros, de cada una de las 15 personas que se especifica en el ejemplo 5.1.2, cuya distribución de frecuencias agrupada venía dada por:

ALTURA			
$L_{i-1} - L_i$	a_i	x_i	n_i
160 - 170	10	165	4
170 - 180	10	175	6
180 - 190	10	185	2
190 - 195	5	192,5	3

Puesto que según esta agrupación de los valores de la variable, la media aritmética es 177.167 cm., tal y como se obtuvo al desarrollar el ejemplo 4.2.2, se tiene que:

ALTURA			
$L_{i-1} - L_i$	x_i	n_i	$(x_i - \bar{x})^2 n_i$
160 - 170	165	4	592.144
170 - 180	175	6	28.175
180 - 190	185	2	122.712
190 - 195	192.5	3	705.303
TOTALES	- -	15	1448.334

y por lo tanto,

$$s^2 = \frac{\sum_{i=1}^{4}(x_i - 177.167)^2 n_i}{N} = \frac{1448.334}{15} = 96.556$$

y así, la varianza, aproximada, de la altura de estas 15 personas es 96.556 cm^2.

Así pues, y a modo de resumen, podemos afirmar que si el promedio seleccionado para representar a la distribución de frecuencias es la media aritmética, podemos considerar como ***medida de dispersión óptima*** a la *varianza*,

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N}$$

mientras que si la medida de tendencia central elegida es la mediana, la ***medida de dispersión óptima*** es la *desviación media respecto de la mediana*,

$$D_{M_e}=\frac{\sum_{i=1}^{k}|x_i-M_e|n_i}{N}$$

Evidentemente, *pueden utilizarse las* ***medidas de dispersión que no imponen la condición de que ser óptimas***. Así, por ejemplo, puede calcularse la desviación media respecto de la media aritmética,

$$D_{\bar{x}}=\frac{\sum_{i=1}^{k}|x_i-\bar{x}|n_i}{N}$$

pero entonces,

$$D_{\bar{x}}\geq D_{M_e}$$

Dado que el promedio por excelencia es la media aritmética, vamos a ofrecer en el siguiente epígrafe las propiedades más relevantes de su medida de dispersión absoluta, la varianza, dejando para un epígrafe posterior el estudio de las medidas de dispersión relativas tanto de la media como de la mediana.

5.3 Propiedades de la varianza. La desviación típica.

Tal y como se ha demostrado en el epígrafe anterior, la varianza es la medida de dispersión óptima para la media aritmética. Si los cálculos han de hacerse manualmente, suele utilizarse la expresión,

$$s^2=\frac{\sum_{i=1}^{k}x_i^2n_i}{N}-\bar{x}^2$$

que se obtiene al desarrollar

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

En efecto, como

$$(x_i-\bar{x})^2=x_i^2-2x_i\bar{x}+\bar{x}^2$$

podemos expresar

$$s^2=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2n_i}{N}=\frac{\sum_{i=1}^{k}x_i^2n_i}{N}+\frac{\bar{x}^2\sum_{i=1}^{k}n_i}{N}-\frac{2\bar{x}\sum_{i=1}^{k}x_in_i}{N}=$$
$$=\frac{\sum_{i=1}^{k}x_i^2n_i}{N}+\frac{\bar{x}^2N}{N}-2\bar{x}\bar{x}=\frac{\sum_{i=1}^{k}x_i^2n_i}{N}-\bar{x}^2$$

Como ***propiedades*** más importantes de esta medida de dispersión podemos destacar las siguientes:

1) La varianza de una variable es siempre *no negativa*. La demostración es inmediata a partir de su expresión, pues como

$$(x_i-\bar{x})^2\geq 0;\ n_i\geq 0;\ N\geq 0;\ i=1,2,..,k$$

se tiene que

$$s^2=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2n_i}{N}\geq 0$$

Obsérvese que únicamente en el caso de que $x_i - \overline{x} = 0 \; \forall i$, i=1,2,...,k, será $s^2=0$, y así, sólo cuando la variable tome un único valor, que lógicamente coincidirá con su media aritmética, podrá ser nula la varianza de dicha variable.

2) Si a todos los valores de la variable les sumamos (o restamos) una constante, la varianza de la nueva variable coincidirá con la varianza de la variable original. Es decir, si Y=X+a, donde a es una constante, entonces $s_Y^2 = s_X^2$.

En efecto, dado que

$$\overline{y} = \overline{x} + a$$

tal y como se ha demostrado en el segundo epígrafe del capítulo anterior, al ser

$$s_Y^2 = \frac{\sum_{i=1}^{k} (y_i - \overline{y})^2 n_i}{N}$$

podemos expresar:

$$s_Y^2 = \frac{\sum_{i=1}^{k} (y_i - \overline{y})^2 n_i}{N} = \frac{\sum_{i=1}^{k} (x_i + a - [\overline{x} + a])^2 n_i}{N} = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{N} = s_X^2$$

Así, puesto que la varianza de la nueva variable Y es la misma que la de X, para cualquier constante, podemos afirmar que la varianza es invariante frente a ***cambios de origen*** en la variable.

3) Si todos los valores de la variable se multiplican, (o dividen) por una constante, la varianza de la nueva variable es igual a la varianza de la variable original multiplicada por el cuadrado de la constante. Es decir, si Y=bX, donde b es una constante, entonces $s_Y^2 = b^2 s_X^2$.

En efecto, dado que

$$\overline{y} = b\,\overline{x}$$

tal y como se ha demostrado en el segundo epígrafe del capítulo anterior, al ser

$$s_Y^2 = \frac{\sum_{i=1}^{k} (y_i - \overline{y})^2 n_i}{N}$$

podemos expresar:

$$s_Y^2 = \frac{\sum_{i=1}^{k} (y_i - \overline{y})^2 n_i}{N} = \frac{\sum_{i=1}^{k} (b\,x_i - b\,\overline{x})^2 n_i}{N} = \frac{b^2 \sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{N} = b^2 s_X^2$$

Así, puesto que la varianza de la nueva variable Y no es la misma que la de X, salvo que la constante sea la unidad, podemos afirmar que la varianza no es invariante frente a ***cambios de escala*** en la variable.

Conjugando esta propiedad con la anterior, se tiene que si se realiza simultáneamente una cambio de origen y de escala en la variable X, tal que Y=a+bX, entonces,

$$\boxed{s_Y^2 = b^2 s_X^2}$$

Nuevamente debemos resaltar que la utilidad práctica de los cambios de origen y escala estriba en la posibilidad de facilitar los cálculos, aspecto que podemos poner de manifiesto en el siguiente supuesto.

Ejemplo 5.3.1. Consideremos el salario anual de cada una de las 15 personas que vienen conformando nuestro colectivo objeto de estudio,

que se especifica en el ejemplo 4.2.4, cuya distribución de frecuencias es la que se ofrece a continuación:

SALARIO	
x_i	n_i
2000000	3
2500000	4
3000000	2
3450000	1
3900000	3
4300000	2

El cálculo de la varianza a partir de los valores que toma la variable X puede hacerse de varias formas. En primer lugar, como se ha de obtener la media aritmética, podemos añadir a la tabla anterior una columna que recoja el producto de cada valor de la variable por su frecuencia, aunque ya conocemos que dicha media es 3050000 ptas., tal y como se muestra en el ejemplo 4.2.4. A ello añadiremos otra que indique la diferencia entre cada valor de la variable y su media aritmética, cuyo valor ascendería a 3050000 ptas. tal y como se muestra en el mencionado ejemplo; y otra que recoja cada una de estas diferencias al cuadrado multiplicadas por sus respectivas frecuencias, así como una última fila en la que se recoja la suma de las columnas de interés, tal y como se muestra a continuación:

SALARIO					
	x_i	n_i	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 n_i$
	2000000	3	6000000	-1050000	$3.3075\ 10^{12}$
	2500000	4	10000000	-550000	$1.21\ 10^{12}$
	3000000	2	6000000	-50000	$5\ 10^{9}$
	3450000	1	3450000	400000	$1.6\ 10^{11}$
	3900000	3	11700000	850000	$2.1675\ 10^{12}$
	4300000	2	8600000	1250000	$3.125\ 10^{12}$
TOTAL	- -	15	45750000	- -	$9.975\ 10^{12}$

Entonces, se tiene que

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N} = \frac{9.975 \cdot 10^{12}}{15} = 6.65 \cdot 10^{11}$$

por lo que la varianza del salario para estas 15 personas es 665000000000 (seiscientos sesenta y cinco mil millones) ptas^2., tal y como se ha obtenido anteriormente.

También, cabe la posibilidad de añadir a la tabla primitiva, además de la columna que ofrece el producto de cada valor de la variable por su frecuencia, otra que muestre el producto del cuadrado de los valores de la variable por su frecuencia, así como una última fila en la que se recoja la suma de las tres últimas columnas, tal y como se muestra a continuación:

SALARIO				
	x_i	n_i	$x_i n_i$	$x_i^2 \cdot n_i$
	2000000	3	6000000	$1,2\ 10^{13}$
	2500000	4	10000000	$2.5\ 10^{13}$
	3000000	2	6000000	$1.8\ 10^{13}$
	3450000	1	3450000	$1.19025\ 10^{13}$
	3900000	3	11700000	$4.563\ 10^{13}$
	4300000	2	8600000	$3.698\ 10^{13}$
TOTAL	- -	15	45750000	$14.95125\ 10^{13}$

y entonces, se tiene que

$$s^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \overline{x}^2 = \frac{14,95125 \cdot 10^{13}}{15} - \left(\frac{45750000}{15}\right)^2 = 6.65 \cdot 10^{11}$$

por lo que la varianza del salario para estas 15 personas es 665000000000 ptas^2., tal y como se ha obtenido anteriormente.

Ahora bien, si construimos una nueva variable Y, tal que

$$Y = \frac{X - 3000000}{1000000}$$

podemos expresar entonces;

Y				
	y_i	n_i	$y_i n_i$	$y_i^2 n_i$
	-1	3	-3	3
	-0,5	4	-2	1
	0	2	0	0
	0.45	1	0.45	0.2025
	0,9	3	2.7	2.43
	1.3	2	2.6	3.38
TOTAL	- -	15	0.75	10.0125

De esta forma, la varianza de la variable Y es

$$s_Y^2 = \frac{\sum_{i=1}^{6} y_i^2 n_i}{N} - \overline{y}^2 = \frac{10.0125}{15} - \left(\frac{0.75}{15}\right)^2 = 0.665$$

y entonces, al ser

$$Y = \frac{X - 3000000}{1000000}$$

se tiene que

$$s_Y^2 = \left(\frac{1}{1000000}\right)^2 s_X^2$$

por lo que

$$s_X^2 = s_Y^2 \cdot 1000000^2 = 0.665 \cdot 1000000^2 = 6.65 \cdot 10^{11}$$

y así se puede concluir nuevamente, que la varianza del salario para estas 15 personas es 665000000000 ptas^2.

Obsérvese la gran diferencia que existe entre los valores de las tablas anteriores cuando se emplea la variable original X, en cualquiera de sus posibilidades, o cuando en su lugar se utiliza la variable Y.

También, al igual que al analizar la media aritmética, debemos abordar la determinación de la varianza de una variable, cuando ésta se considera a través de una ***composición de poblaciones***. En este sentido, sea una población P de tamaño N, formada por varias subpoblaciones P_1, P_2, ..., P_m, de tamaños N_1, N_2, ..., N_m, siendo entonces

$$N = \sum_{j=1}^{m} N_j$$

y sea n_{ij}, i=1,2,...,k, j=1,2,...,m, el número de elementos que en la subpoblación P_j alcanzan el valor x_i de la variable X. Así, para la población P, el número de elementos que toman el valor x_i, n_i, es precisamente

$$n_i = \sum_{j=1}^{m} n_{ij}$$

Bajo el supuesto de que la variable no se agrupa, simplificando entonces la presentación, podemos especificar su distribución de frecuencias mediante la siguiente tabla:

DISTRIBUCIÓN DE FRECUENCIAS DE LA VARIABLE X					
	P_1	P_2	... P_j ...	P_m	P (TOTALES)
x_1	n_{11}	n_{12}	... n_{1j} ...	n_{1m}	n_1
x_2	n_{21}	n_{22}	... n_{2j} ...	n_{2m}	n_2
...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	... n_{ij} ...	n_{im}	n_i
...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	... n_{kj} ...	n_{km}	n_k
TOTAL	N_1	N_2	... N_j ...	N_m	- -

La media y la varianza de la variable X en la subpoblación j, $\overline{x}_j$, j=1,2,..,m, se pueden expresar

$$\overline{x}_j = \frac{\sum_{i=1}^{k} x_i n_{ij}}{N_j}$$

$$s_j^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x}_j)^2 n_{ij}}{N_j}$$

mientras que la media y la varianza de dicha variable para la población P vienen dadas por

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

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N}$$

Ahora bien, dado que

$$\sum_{i=1}^{k}(x_i - \overline{x}_j)^2 n_{ij} = \sum_{i=1}^{k}(x_i - \overline{x} + \overline{x} - \overline{x}_j)^2 n_{ij} =$$

$$= \sum_{i=1}^{k}(x_i - \overline{x})^2 n_{ij} + (\overline{x}_j - \overline{x})^2 \sum_{i=1}^{k} n_{ij} - 2(\overline{x}_j - \overline{x}) \sum_{i=1}^{k}(x_i - \overline{x}) n_{ij} =$$

$$= \sum_{i=1}^{k}(x_i - \overline{x})^2 n_{ij} - (\overline{x}_j - \overline{x})^2 N_j$$

al ser:

$$\sum_{i=1}^{k} n_{ij} = N_j$$

$$\sum_{i=1}^{k}(x_i - \overline{x}) n_{ij} = \sum_{i=1}^{k} x_i n_{ij} - \overline{x} \sum_{i=1}^{k} n_{ij} = \overline{x}_j N_j - \overline{x} N_j = N_j(\overline{x}_j - \overline{x})$$

y puesto que

$$\sum_{i=1}^{k}(x_i - \overline{x}_j)^2 n_{ij} = s_j^2 N_j$$

a partir de

$$\sum_{i=1}^{k}(x_i - \overline{x})^2 n_{ij} = \sum_{i=1}^{k}(x_i - \overline{x}_j)^2 n_{ij} + (\overline{x}_j - \overline{x})^2 N_j$$

podemos expresar

$$\sum_{i=1}^{k}(x_i - \overline{x})^2 n_{ij} = s_j^2 N_j + (\overline{x}_j - \overline{x})^2 N_j$$

y entonces:

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N} = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 \sum_{j=1}^{m} n_{ij}}{N} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_{ij}}{N} =$$

$$= \frac{\sum_{j=1}^{m} N_j s_j^2}{N} + \frac{\sum_{j=1}^{m} N_j(\overline{x}_j - \overline{x})^2}{N}$$

De esta forma, se tiene que la varianza de la variable X en la población P, puede expresarse:

$$s^2 = \frac{\sum_{j=1}^{m} N_j s_j^2}{N} + \frac{\sum_{j=1}^{m} N_j(\overline{x}_j - \overline{x})^2}{N}$$

esto es, la varianza de la variable X en la población es igual a la media aritmética ponderada de las varianzas de cada una de las subpoblaciones, más la varianza de las medias de cada una de dichas subpoblaciones, donde tanto la ponderación en el primer caso, como la frecuencia en el segundo, es precisamente el número de elementos de la subpoblación.

Ejemplo 5.3.2. Consideremos nuevamente los supuestos especificados en el ejemplo 4.2.7. En el mismo se nos ofrecía la información relativa al número de hijos y comunidad de nacimiento de un conjunto de 15 personas, la cual se puede plasmar en la siguiente tabla:

NÚMERO DE HIJOS					
x_i	Nacidos en Andalucía	Nacidos en Extremadura	Nacidos en la Comunidad de Madrid	Nacidos en la Comunidad Valenciana	Total de personas con x_i hijos (n_i)
0	1	1	3	0	5
1	1	1	2	2	6
2	2	1	0	0	3
4	1	0	0	0	1
Total de personas por comunidad de nacimiento	5	3	5	2	--

A partir de la información relativa al total de personas que no tienen hijos, que tienen 1, que tienen 2 y que tienen 4, se puede calcular fácilmente la varianza de la variable número de hijos, que tal y como se ha ofrecido en el ejemplo 5.2.4 es de 1.129 hijos^2. Ahora bien, supongamos que en lugar de ofrecer la tabla anterior, se nos indica que las cinco personas que nacieron en Andalucía tienen una media de 1.8 hijos y una varianza de 1.76 hijos^2; que las 3 que lo hicieron en Extremadura, tienen 1 hijo por término medio, con una varianza de 0.666 hijos^2; que las cinco que nacieron en la Comunidad de Madrid, tienen una media de 0.4 hijos y una varianza de 0.24 hijos^2; y que los dos que lo hicieron en la Comunidad Valenciana, tienen también 1 hijo por término medio, con una varianza nula. Bajo este nuevo supuesto, también es posible especificar la varianza de la variable número de hijos, pues dado que la media de dicha variable es 16/15, tal y como se obtuvo al desarrollar el ejemplo 4.2.8, podemos ofrecer la siguiente tabla:

Subpoblación	$\bar{x}_j$	s_j^2	N_j	$s_j^2 N_j$	$(\bar{x}_j - \bar{x})^2 N_j$
1	1.8	1.76	5	8.8	2.689
2	1	0.666	3	2	0.013
3	0.4	0.24	5	1.2	2.222
4	1	0	2	0	0.009
TOTAL	--	--	15	12	4.933

donde las subpoblaciones hacen referencia a cada una de las cuatro comunidades autónomas.

Así, se tiene que

$$s^2 = \frac{\sum_{j=1}^{m} N_j s_j^2}{N} + \frac{\sum_{j=1}^{m} N_j (\bar{x}_j - \bar{x})^2}{N} = \frac{12}{15} + \frac{4.933}{15} = 0.8 + 0.329 = 1.129$$

por lo que podemos concluir que la varianza de la variable número de hijos es 1.129 hijos^2.

Parece pues que la varianza, que es una medida de dispersión absoluta óptima para la media aritmética, reúne además un conjunto de propiedades que la pueden hacer muy útil. Sin embargo, tiene un **gran inconveniente**, y es que no viene expresada en las mismas unidades en las que lo hace la variable, sino en dichas unidades al cuadrado. La solución que proponemos ante esta situación es tomar la raíz cuadrada positiva de la varianza; esto es,

$$+\sqrt{\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{N}}$$

A esta medida la vamos a conocer como ***desviación típica*** *o* ***desviación estándar***, s, y entonces, la desviación típica de la variable X se puede especificar:

$$s=+\sqrt{\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{N}}=+\sqrt{s^2}$$

Ejemplo 5.3.3. Si consideramos nuevamente el número de hijos de cada una de las 15 personas a que se refiere el ejemplo 5.1.1, cuya distribución de frecuencias venía dada por:

NÚMERO DE HIJOS	
x_i	n_i
0	5
1	6
2	3
4	1

se tiene que la varianza de esta variable es 1.129 hijos^2, según se obtuvo en el ejemplo 5.2.4. Entonces, puesto que

$$s=+\sqrt{1.129}=1.0625$$

se tiene que la desviación típica de la variable número de hijos de las 15 personas es 1.0625 hijos.

Las ***propiedades*** de la desviación típica, que se desprenden directamente de las de la varianza, son las siguientes:

1) $s \geq 0$, pues se ha despreciado la solución negativa de la raíz.

2) Si Y=X+a, donde a es una constante, entonces $s_Y=s_X$.

3) Si Y=bX, donde b es una constante, entonces $s_Y=|b|s_X$, pues si b fuese negativo, si no se considera el valor absoluto, la desviación típica de Y sería negativa. Al conjugar esta propiedad con la anterior, se tiene que si Y=a+bX, entonces $s_Y=|b|s_X$.

Además, la desviación típica **viene expresada en las mismas unidades** en las que lo haga la variable objeto de estudio.

Sin embargo, aún siendo este último aspecto una notable ventaja sobre la varianza, se sigue presentando el problema de *interpretar adecuadamente su valor*. Así, una desviación típica de 1.06 hijos, ¿se puede considerar grande? o, por el contrario, ¿es pequeña?. Obviamente, habrá que compararla con el valor de la media aritmética, pues si dicho valor medio es 20 hijos, 1.06 es pequeño, pero si éste es 2 hijos, entonces es grande.

Y no sólo eso, sino que además, si se pretende comparar la dispersión que con respecto a sus medias presentan dos o más distribuciones de frecuencias, si la desviación típica en una de ellas es 1.06 hijos, en otra es 9.83 cm. ($\sqrt{96.556}$), y en otra es 815475.32 ptas. ($\sqrt{6.65\cdot 10^{11}}$), es imposible comparar estas desviaciones típicas, pues vienen expresadas en distintas unidades.

Para resolver ambos problemas, vamos a considerar una *medida de dispersión relativa, el coeficiente de variación*, que va a ser especificada en el siguiente epígrafe.