En un estudio sobre el número de personas que formaban las bandas de "gangsters" en el Chicago de 1927, se obtuvieron los siguientes datos sobre 825 de dichas bandas:

Tamaño banda	frecuencia	Tamaño banda	frecuencia
$3-6$	37	$41-51$	51
$6-11$	198	$51-76$	26
$11-16$	191	$76-101$	25
$16-21$	149	$101-201$	25
$21-26$	79	$201-501$	11
$26-31$	46	$501-1000$	2
$31-41$	55		

(El autor de este estudio, F.M. Thrasher, consideró que las personas aisladas o los grupos de dos "gangsters" no constituían una banda.)

Analizar descriptivamente estos datos.

El enunciado nos da la distribución de frecuencias absolutas de unos datos correspondientes a un carácter cuantitativo, *número de personas que componen la banda*, clasificados por intervalos.

Siguiendo la notación habitual (CB-sección 2.2), entenderemos que los intervalos son cerrados por la izquierda y abiertos por la derecha, menos el último que es cerrado por ambos lados; así por ejemplo, si una banda está compuesta por 501 "gansters", ésta deberá ser contabilizada en el último intervalo y no en el penúltimo.

Las cuatro distribuciones de frecuencias (absolutas, relativas, absolutas acumuladas y relativas acumuladas) son, respectivamente

	I_i	n_i	f_i	N_i	F_i
I_1	$3-6$	37	$0'04134$	37	$0'04134$
I_2	$6-11$	198	$0'22123$	235	$0'26257$
I_3	$11-16$	191	$0'21341$	426	$0'47598$
I_4	$16-21$	149	$0'16648$	575	$0'64246$
I_5	$21-26$	79	$0'08827$	654	$0'73073$
I_6	$26-31$	46	$0'05140$	700	$0'78213$
I_7	$31-41$	55	$0'06145$	755	$0'84358$
I_8	$41-51$	51	$0'05698$	806	$0'90056$
I_9	$51-76$	26	$0'02905$	832	$0'92961$
I_{10}	$76-101$	25	$0'02793$	857	$0'95754$
I_{11}	$101-201$	25	$0'02793$	882	$0'98547$
I_{12}	$201-501$	11	$0'01229$	893	$0'99776$
I_{13}	$501-1000$	2	$0'00224$	895	1
		895	1		

La representación gráfica de este tipo de datos es (CB-sección 2.3.1) un *histograma* para las distribuciones absolutas y relativas sin acumular y un *polígono de frecuencias acumuladas* para las distribuciones acumuladas.

Consideraremos solamente un histograma para la distribución de frecuencias absolutas. Al ser el histograma una representación por áreas, debemos calcular la altura de cada rectángulo de forma que se cumpla la ecuación

$$\text{Frecuencia absoluta (área)} = \text{base} \cdot \text{altura}$$

por lo que tomando, para simplificar, una amplitud unidad igual a 10, la siguiente tabla nos da las longitudes de cada intervalo, el factor g_i que multiplica a la amplitud unidad 10 para conseguir la longitud del intervalo y la altura h_i determinada como cociente $h_i = n_i/g_i$, ya que, en ese caso, será

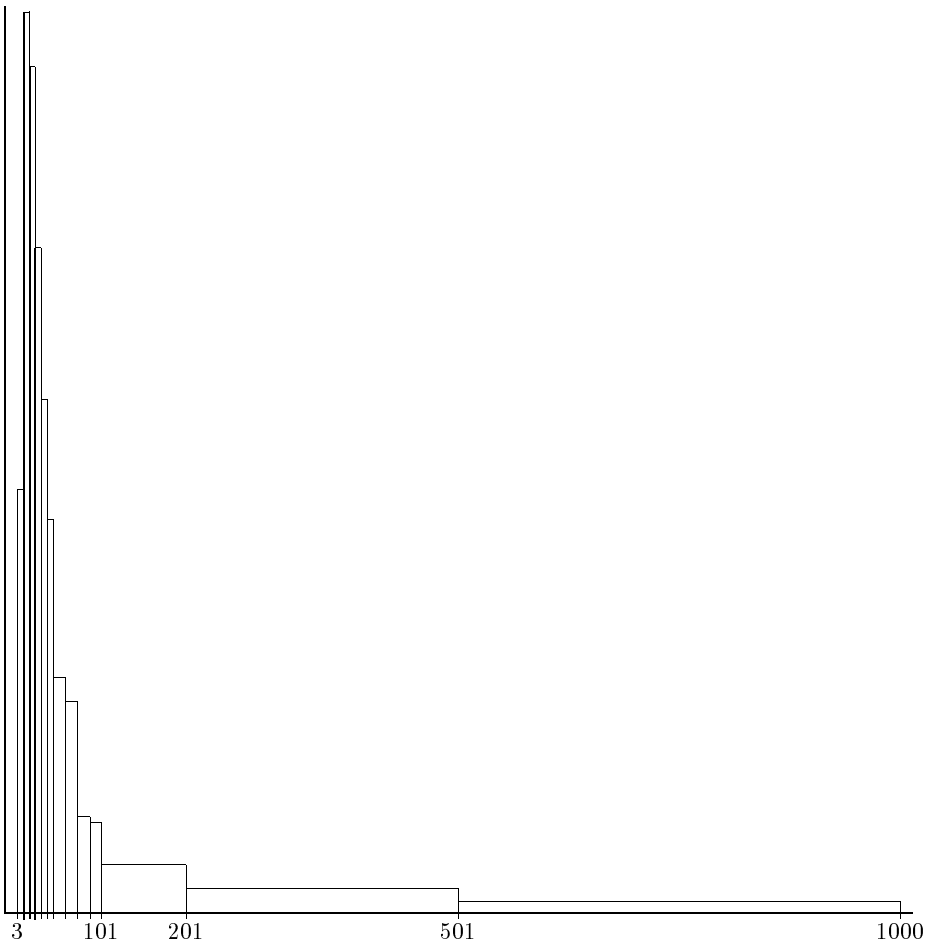
$$\text{Frecuencia absoluta} = n_i = (1 \cdot g_i) \cdot (n_i/g_i) = (\text{base}) \cdot (\text{altura})$$

	longitud	g_i	h_i(altura)
I_1	3	$0'3$	$123'3$
I_2	5	$0'5$	396
I_3	5	$0'5$	382
I_4	5	$0'5$	298
I_5	5	$0'5$	158
I_6	5	$0'5$	92
I_7	10	1	55
I_8	10	1	51
I_9	25	$2'5$	$10'4$
I_{10}	25	$2'5$	10
I_{11}	100	10	$2'5$
I_{12}	300	30	$0'37$
I_{13}	499	$49'9$	$0'04$

El histograma que se obtiene será el de la página siguiente, el cual muestra una clara asimetría a la derecha de la distribución de frecuencias, al descender más despacio las frecuencias por el lado derecho.

En la determinación de las medidas de posición, dispersión y asimetría, utilizaremos la siguiente tabla de cálculos, en la cual, al estar los datos agrupados en intervalos $[e_{i-1}\,,\,e_i)$, juegan un papel esencial las *marcas de clase* $x_i = (e_i + e_{i-1})/2$

I_i	x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
$3-6$	$4'5$	37	$166'5$	$749'25$
$6-11$	$8'5$	198	1683	$14305'5$
$11-16$	$13'5$	191	$2578'5$	$34809'75$
$16-21$	$18'5$	149	$2756'5$	$50995'25$
$21-26$	$23'5$	79	$1856'5$	$43627'75$
$26-31$	$28'5$	46	1311	$37363'5$
$31-41$	36	55	1980	71280
$41-51$	46	51	2346	107916
$51-76$	$63'5$	26	1651	$104838'5$
$76-101$	$88'5$	25	$2212'5$	$195806'25$
$101-201$	151	25	3775	570025
$201-501$	351	11	3861	1355211
$501-1000$	$750'5$	2	1501	$1126500'5$
		895	$27678'5$	$3713428'25$



De la tabla de cálculos se obtiene que la *media aritmética* es

$$a = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \frac{1}{895}\sum_{i=1}^{13} x_i \cdot n_i = \frac{27678'5}{895} = 30'93.$$

Respecto a la *mediana*, a partir de la distribución de frecuencias absolutas obtenemos que es

$$N_3 = 426 < \frac{n}{2} = \frac{895}{2} = 447'5 < 575 = N_4$$

con lo que la mediana M_e está en el intervalo $[16\,,\,21)$, siendo la mediana el valor

$$M_e = x_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{n_j} \cdot c_j = 16 + \frac{447'5 - 426}{149} \cdot 5 = 16'72.$$

Se observa que la mediana $M_e = 16'72$ se ve menos influenciada por los valores extremos —en este caso, muy altos— que la media aritmética $a = 30'93$. Este hecho hace que se recomienda utilizar la mediana en lugar de la media como medida representativa de los datos.

Respecto a la *moda*, al tener los intervalos diferente amplitud, primero debemos *normalizar* los intervalos calculando los cocientes

$$l_j = \frac{n_j}{c_j}$$

pero, como la longitud c_j de cada intervalo es $c_j = 10 \cdot g_j$, será

$$l_j = \frac{n_j}{c_j} = \frac{n_j}{10 \cdot g_j} = \frac{h_j}{10}$$

siendo el

$$\max\{l_1, ..., l_k\} = \frac{1}{10}\max\{h_1, ..., h_k\} = \frac{396}{10} = 39'6$$

y, por tanto, el intervalo modal el $I_2 = [6\,,\,11)$, con lo que la moda será

$$M_d = x_{j-1} + \frac{l_{j+1} \cdot c_j}{l_{j-1} + l_{j+1}} = x_{j-1} + \frac{c_j \cdot h_{j+1}/10}{(h_{j-1} + h_{j+1})/10} = 6 + \frac{382 \cdot 5}{123'3 + 382} = 9'78.$$

Respecto a los *cuantiles*, si consideramos el *primer cuartil*, al ser

$$37 < \frac{1}{4} \cdot n = 223'75 < 235$$

será $p_{1/4} \in [6\,,\,11)$ y, en concreto,

$$p_{1/4} = x_{j-1} + \frac{\frac{1}{4} \cdot n - N_{j-1}}{n_j} \cdot c_j = 6 + \frac{\frac{1}{4} \cdot 895 - 37}{198} \cdot 5 = 10'716$$

el cual será igual al centil 25.

Para calcular el *sexto decil* (que es igual al centil sesenta), acotamos el valor

$$\frac{6}{10}\cdot n = \frac{60}{100}\cdot n = \frac{6}{10}\cdot 895 = 537$$

por las frecuencias absolutas acumuladas

$$N_3 = 426 < 537 < 575 = N_4$$

estando dicho valor, por tanto, en el intervalo $[16\,,\,21)$, y siendo igual a

$$p_{6/10} = x_{j-1} + \frac{\frac{6}{10}\cdot n - N_{j-1}}{n_j}\cdot c_j = 16 + \frac{537-426}{149}\cdot 5 = 19'72.$$

Respecto a las *medidas de dispersión* (CB-sección 2.3.3), el *recorrido* es igual a

$$R = x_{max} - x_{min} = 750'5 - 4'5 = 746$$

al ser $750'5$ la última marca de clase y $4'5$ la primera.

Utilizando la tabla de cálculos antes determinada, obtenemos que la *varianza* es igual a

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - a)^2\, n_i = \frac{1}{n}\sum_{i=1}^{k} x_i^2\, n_i - a^2 = \frac{3713428'25}{895} - 30'93^2 = 3192'417$$

que permite calcular la *cuasivarianza*

$$S^2 = \frac{n\cdot s^2}{n-1} = \frac{895\cdot 3192'417}{894} = 3195'988.$$

Si calculamos sus raíces cuadradas, que expresan mejor la dispersión de los datos al ser de esa manera los valores obtenidos *número de individuos*, la *desviación típica* es igual a

$$s = \sqrt{s^2} = \sqrt{3192'417} = 56'50$$

y la *cuasidesviación típica*

$$S = \sqrt{S^2} = \sqrt{3195'988} = 56'53$$

valores que, como se ve, apenas si se diferencian, al ser grande el número de datos considerados.

El *coeficiente de variación de Pearson* resulta igual a

$$V_p = \frac{s}{a} \cdot 100 = \frac{56'50}{30'93} \cdot 100 = 182'67.$$

Por último, el *coeficiente de asimetría de Pearson* (CB-sección 2.3.4),

$$A_p = \frac{a - M_d}{s} = \frac{30'93 - 9'78}{56'50} = 0'374$$

confirma la asimetría a la derecha de la distribución de frecuencias.