Estadística descriptiva: problemas resueltos

BENITO J. GONZÁLEZ RODRÍGUEZ (bjglez@ull.es)
DOMINGO HERNÁNDEZ ABREU (dhabreu@ull.es)
MATEO M. JIMÉNEZ PAIZ (mjimenez@ull.es)
M. ISABEL MARRERO RODRÍGUEZ (imarrero@ull.es)
ALEJANDRO SANABRIA GARCÍA (asgarcia@ull.es)

Departamento de Análisis Matemático
Universidad de La Laguna

Índice

5. Problemas resueltos

5.1. Variables no agrupadas

Ejercicio 5.1. *En una clínica infantil se ha ido anotando, durante un mes, el número de metros que cada niño anda, seguido y sin caerse, el primer día que comienza a caminar, obteniéndose la tabla de información adjunta:*

número de metros	1	2	3	4	5	6	7	8
número de niños	2	6	10	5	10	3	2	2

Se pide:

a) *Tabla de frecuencias. Diagrama de barras para frecuencias absolutas, relativas y acumuladas.*

b) *Mediana, media aritmética, moda y cuartiles.*

c) *Varianza y desviación típica.*

d) *¿Entre qué dos valores se encuentra, como mínimo, el 75 % de las observaciones?*

RESOLUCIÓN. La variable considerada en el estudio es cuantitativa discreta.

a) Al tratarse de una variable discreta podemos confeccionar directamente la tabla de frecuencias (Cuadro 5.1).

x_i	n_i	N_i	f_i	f_i (%)	F_i	F_i (%)
1	2	2	0.050	5.0	0.050	5.0
2	6	8	0.150	15.0	0.200	20.0
3	10	18	0.250	25.0	0.450	45.0
4	5	23	0.125	12.5	0.575	57.5
5	10	33	0.250	25.0	0.825	82.5
6	3	36	0.075	7.5	0.900	90.0
7	2	38	0.050	5.0	0.950	95.0
8	2	40	0.050	5.0	1.000	100.0

Cuadro 5.1. Tabla de frecuencias para la variable del Ejercicio 5.1.

Los diagramas de barras de frecuencias se representan en las Figuras 5.1, 5.2, 5.3 y 5.4.

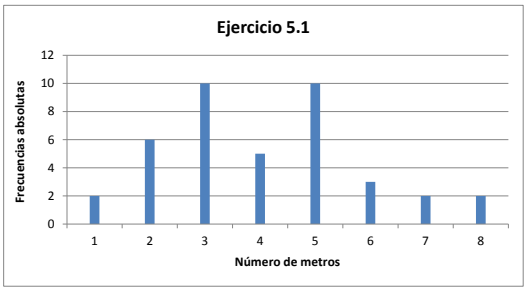

Figura 5.1. Diagrama de barras de frecuencias absolutas para la variable del Ejercicio 5.1.

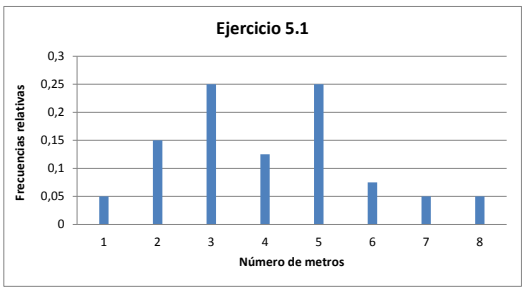

Figura 5.2. Diagrama de barras de frecuencias relativas para la variable del Ejercicio 5.1.

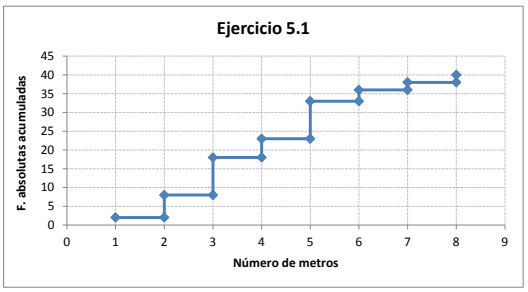

Figura 5.3. Diagrama de barras acumulativo de frecuencias absolutas para la variable del Ejercicio 5.1.

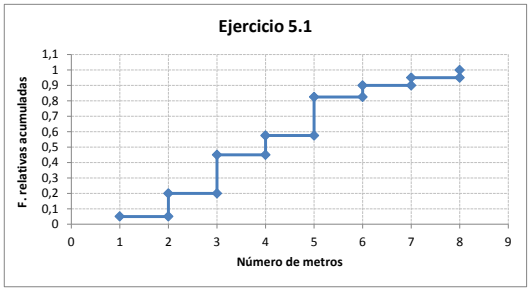

Figura 5.4. Diagrama de barras acumulativo de frecuencias relativas para la variable del Ejercicio 5.1.

b) Nos ocupamos en primer lugar de las medidas de centralización. La media $\bar{x}$ viene dada por:

$$\bar{x} = \frac{1}{40}(1\cdot 2+2\cdot 6+3\cdot 10+4\cdot 5+5\cdot 10+6\cdot 3+7\cdot 2+8\cdot 2) = 4.05 \simeq 4.$$

En la tabla de frecuencias (Cuadro 5.1) observamos que la variable es bimodal, con modas

$$M_{o_1} = 3 \quad \text{y} \quad M_{o_2} = 5,$$

pues estos dos valores de la variable son los que presentan una mayor frecuencia absoluta, a saber, 10.

La mediana divide la distribución en dos partes iguales. Como en el Cuadro 5.1 no existe ningún valor de la variable que acumule el 50 % de los datos, la mediana será el primero que supere este porcentaje:

$$M_e = 4.$$

De manera análoga se procede para calcular el primer, segundo y tercer cuartiles. Estos son los valores de la variable que acumulan, respectivamente, el 25 %, 50 % y 75 % de las observaciones. Al no comparecer en la columna de frecuencias relativas acumuladas del Cuadro 5.1 exactamente estos porcentajes tomamos los inmediatamente superiores, de modo que

$$P_{1/4} = 3, \quad P_{2/4} = M_e = 4, \quad P_{3/4} = 5.$$

c) Ahora determinaremos las medidas de dispersión. Utilizando la relación

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

se tiene que la varianza viene dada por:

$$\begin{aligned}\sigma^2 =& \frac{1}{40}(1\cdot 2+4\cdot 6+9\cdot 10+16\cdot 5+25\cdot 10+36\cdot 3+49\cdot 2+64\cdot 2) - 4.05^2\\ =& 19.5 - 16.4025\\ =& 3.0975 \simeq 3.10.\end{aligned}$$

Consecuentemente, la desviación típica es

$$\sigma = \sqrt{3.0975} \simeq 1.76.$$

d) El Teorema de Chebyshev garantiza que, como mínimo, el $(1-\frac{1}{k^2})\cdot 100\,\%$ de los datos se concentran en el intervalo $(\bar{x}-k\sigma,\bar{x}+k\sigma)$ y, por tanto, fuera de dicho intervalo se encuentra, a lo sumo, el $\frac{1}{k^2}\cdot 100\,\%$ de ellos.

Conforme a este teorema, imponemos que

$$75 = \left(1-\frac{1}{k^2}\right)\cdot 100,$$

de donde

$$100-75 = \frac{1}{k^2}\cdot 100$$

y

$$k^2 = \frac{100}{25} = 4.$$

Por lo tanto, $k=2$. Podemos así garantizar que, al menos, el 75 % de los datos se encuentran entre los valores

$$\bar{x}-k\sigma = 4.05-2\cdot 1.76 = 0.53$$

y

$$\bar{x}+k\sigma = 4.05+2\cdot 1.76 = 7.57.$$

□

Ejercicio 5.2. *Las cifras dadas en la tabla adjunta corresponden a miligramos de hidroxiprolina absorbidos por* 1 *miligramo de masa intestinal, analizados en distintos pacientes:*

hidroxiprolina (mg)	77.3	61.2	82.4	75.9	61	70.2	65	80
número de pacientes	3	10	15	13	8	5	2	0

Se pide:

a) *Confeccionar la tabla de frecuencias.*

b) *Calcular la media, mediana, moda y cuartiles.*

c) *Calcular la desviación típica de la muestra.*

d) *¿Qué porcentaje de observaciones se concentra en el intervalo* $(\bar{x}-5\sigma,\bar{x}+5\sigma)$*?*

RESOLUCIÓN. La variable considerada en el estudio es cuantitativa discreta.

a) Al tratarse de una variable discreta podemos confeccionar directamente la tabla de frecuencias (Cuadro 5.2).

x_i	n_i	N_i	$f_i\ (\%)$	$F_i\ (\%)$
61.0	8	8	14.3	14.3
61.2	10	18	17.9	32.2
65.0	2	20	3.6	35.8
70.2	5	25	8.9	44.7
75.9	13	38	23.2	67.9
77.3	3	41	5.4	73.3
80.0	0	41	0.0	73.3
82.4	15	56	26.8	100.0

Cuadro 5.2. Tabla de frecuencias para la variable del Ejercicio 5.2.

b) La media aritmética viene dada por:

$$\begin{aligned}\bar{x} &= \frac{1}{56}\left(8\cdot 61.0+10\cdot 61.2+2\cdot 65.0+5\cdot 70.2+13\cdot 75.9+3\cdot 77.3+0\cdot 80.0+15\cdot 82.4\right)\\ &= 72.06428571 \simeq 72.1.\end{aligned}$$

La moda es

$$M_o = 82.4,$$

ya que a este valor de la variable le corresponde la mayor frecuencia absoluta, a saber, 15.

La mediana viene dada por

$$M_e = 75.9,$$

pues en el Cuadro 5.2 ninguna puntuación de la variable acumula exactamente el 50 % de los datos, siendo 75.9 la primera con una frecuencia relativa acumulada superior al 50 %.

Nótese que la mediana coincide con el segundo cuartil:

$$P_{1/2} = M_e = 75.9.$$

El primer cuartil $P_{1/4}$ viene dado por

$$P_{1/4} = 61.2,$$

mientras que el tercer cuartil es

$$P_{3/4} = 82.4.$$

En efecto, en la columna de frecuencias relativas acumuladas del Cuadro 5.2 vemos que estos valores de la variable son los primeros que superan, respectivamente, al 25 % y al 75 % de las observaciones.

c) La varianza de la muestra viene dada por

$$\begin{aligned}\sigma^2 =& \frac{1}{56}(8\cdot 61^2 + 10\cdot 61.2^2 + 2\cdot 65^2 + 5\cdot 70.2^2 + 13\cdot 75.9^2 + +3\cdot 77.3^2 + 0\cdot 80^2 + 15\cdot 82.4^2) - 72.1^2 \\ =& 69.007857 \simeq 69.0.\end{aligned}$$

Luego, la desviación típica es

$$\sigma = \sqrt{69.007857} \simeq 8.3.$$

d) De acuerdo con el Teorema de Chebyshev, en el intervalo $(\bar{x}-5\sigma, \bar{x}+5\sigma)$ podemos encontrar un mínimo del

$$\left(1 - \frac{1}{5^2}\right)\cdot 100\,\% = \frac{25-1}{25}\cdot 100\,\% = \frac{24}{25}\cdot 100\,\% = 96\,\%$$

de las observaciones. □

5.2. Variables agrupadas

Ejercicio 5.3. *Los valores del pH sanguíneo de* 32 *individuos son los siguientes:*

7.33	7.31	7.26	7.33	7.37	7.27	7.30	7.33
7.33	7.32	7.35	7.39	7.33	7.38	7.33	7.31
7.37	7.35	7.34	7.32	7.29	7.35	7.38	7.32
7.32	7.33	7.32	7.40	7.33	7.32	7.34	7.33

a) *Agrupar los datos en* 5 *intervalos y confeccionar la tabla de frecuencias.*

b) *Calcular la media aritmética, la moda y la mediana.*

c) *Hallar el tercer decil.*

d) *Determinar el porcentaje de individuos que se concentra fuera del intervalo* $(\bar{x}-4\sigma,\bar{x}+4\sigma)$.

RESOLUCIÓN. En primer lugar, nótese que la variable considerada en el estudio es una variable cuantitativa continua. Por esta razón distribuimos los datos observados en varios intervalos de clase.

a) Para establecer la longitud de cada intervalo de clase hemos de determinar el rango de la variable:

$$R = x_{\max} - x_{\min} = 7.40 - 7.26 = 0.14.$$

Consecuentemente,

$$\ell = \frac{R}{5} = \frac{0.14}{5} = 0.028.$$

Redondeando por exceso podemos tomar $\ell = 0.03$.

La tabla de frecuencias para la variable en estudio queda recogida en el Cuadro 5.3.

intervalos de clase	x_i	n_i	N_i	f_i (%)	F_i (%)
$[7.26, 7.29)$	7.275	2	2	6.250	6.250
$[7.29, 7.32)$	7.305	4	6	12.500	18.750
$[7.32, 7.35)$	7.335	17	23	53.125	71.875
$[7.35, 7.38)$	7.365	5	28	15.625	87.500
$[7.38, 7.41)$	7.395	4	32	12.500	100.000

Cuadro 5.3. Tabla de frecuencias para la variable del Ejercicio 5.3.

b) Calculemos la media aritmética:

$$\bar{x} = \frac{1}{32}(2\cdot 7.275 + 4\cdot 7.305 + 17\cdot 7.335 + 5\cdot 7.365 + 4\cdot 7.395) = \frac{234.87}{32} \simeq 7.34.$$

La mayor frecuencia absoluta registrada en la tabla de frecuencias es 17, que corresponde al intervalo $[7.32, 7.35)$. Dicho intervalo es, por tanto, el intervalo modal, o intervalo donde se encuentra la moda M_o.

Finalmente, para calcular la mediana trazamos el polígono de frecuencias absolutas acumuladas (Figura 5.5).

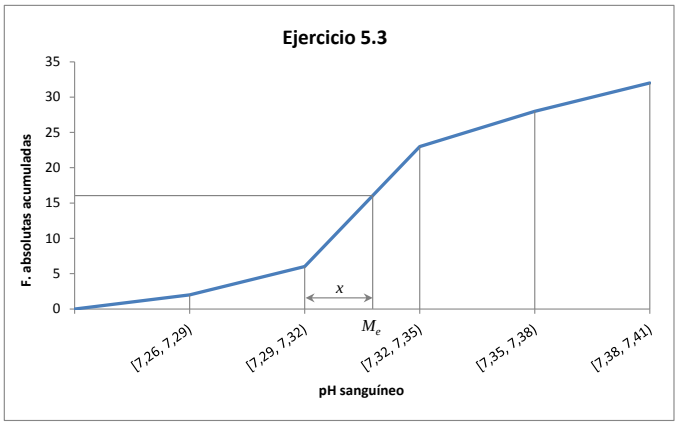

Figura 5.5. Diagrama de frecuencias absolutas acumuladas y cálculo de la mediana para la variable del Ejercicio 5.3.

La mediana divide el número total de observaciones en dos partes iguales, esto es, en 16 observaciones. Atendiendo a la gráfica de la Figura 5.5 y al Cuadro 5.3, se verifica

$$M_e = 7.32 + x,$$

donde x satisface la relación:

$$\frac{16-6}{23-6} = \frac{x}{7.35-7.32}.$$

Entonces

$$x = \frac{0.3}{17} \simeq 0.02,$$

y se concluye que

$$M_e = 7.32 + 0.02 = 7.34.$$

c) Los deciles dividen la distribución en diez partes iguales. Por tanto, el tercer decil se corresponde con el valor de la variable que acumula una frecuencia de

$$3 \cdot \frac{N}{10} = \frac{3 \cdot 32}{10} = 9.6.$$

Para calcularlo procedemos de manera similar que con la mediana:

$$D_3 = 7.32 + x,$$

donde ahora x satisface la relación

$$\frac{0.03}{17} = \frac{x}{9.6-6}.$$

Se infiere que

$$x = \frac{0.03 \cdot 3.6}{17} \simeq 0.006,$$

y concluimos:

$$D_3 = 7.32 + 0.006 = 7.326 \simeq 7.33.$$

d) El Teorema de Chebyshev garantiza que, como mínimo, el $(1-\frac{1}{k^2}) \cdot 100\,\%$ de las observaciones se encuentra en el intervalo $(\bar{x}-k\sigma, \bar{x}+k\sigma)$, mientras que fuera de dicho intervalo se encuentra a lo sumo el $\frac{1}{k^2} \cdot 100\,\%$ de ellas. Consiguientemente, un máximo del

$$\frac{1}{4^2} \cdot 100\,\% = \frac{1}{16} \cdot 100\,\% = 6.25\,\%$$

de los datos cae fuera del intervalo $(\bar{x}-4\sigma, \bar{x}+4\sigma)$. □

Ejercicio 5.4. *En pacientes con distrofia muscular progresiva (*enfermedad de Duchenne*), la actividad de creatinquinasa sérica se eleva llamativamente sobre el valor normal de* 50 *unidades por litro. Los siguientes datos son niveles séricos de creatinquinasa (en unidades por litro) medidos en* 24 *jóvenes pacientes con la enfermedad confirmada:*

3720	3795	3340	5600	3800	3580
5500	2000	1570	2360	1500	1840
3725	3790	3345	3805	5595	3575
1995	5505	2055	1575	1835	1505

Se pide:

a) *Agrupar los datos en* 5 *intervalos de clase.*

b) *Determinar la media y la desviación típica. Calcular la moda y la mediana.*

c) *Determinar el tercer cuartil, el séptimo decil y el centil* 25.

RESOLUCIÓN. Nótese que la variable considerada en el estudio es cuantitativa continua.

a) El rango de la variable es:

$$R = 5600 - 1500 = 4100.$$

Luego,

$$\ell = \frac{R}{5} = \frac{4100}{5} = 820.$$

Redondeando por exceso, tomamos $\ell = 821$.

La tabla de frecuencias para la variable en estudio queda recogida en el Cuadro 5.4.

intervalos de clase	x_i	n_i	N_i
$[1500, 2321)$	1910.5	9	9
$[2321, 3142)$	2731.5	1	10
$[3142, 3963)$	3552.5	10	20
$[3963, 4784)$	4373.5	0	20
$[4784, 5605)$	5194.5	4	24

Cuadro 5.4. Tabla de frecuencias para la variable del Ejercicio 5.4.

b) La media aritmética viene dada por:

$$\bar{x} = \frac{1}{24}(9 \cdot 1910.5 + 1 \cdot 2731.5 + 10 \cdot 3552.5 + 0 \cdot 4373.5 + 4 \cdot 5194.5) = 3176.208333 \simeq 3176.21.$$

Por otro lado, la varianza viene dada por la expresión

$$\begin{aligned}\sigma^2 =& \frac{1}{24}\left[(1910.5 - 3176.21)^2 \cdot 9 + (2731.5 - 3176.21)^2 \cdot 1 + (3552.5 - 3176.21)^2 \cdot 10 + \right.\\ &\left. + (4373.5 - 3176.21)^2 \cdot 0 + (5194.5 - 3176.21)^2 \cdot 4)\right]\\ =& 1346911.79,\end{aligned}$$

de donde la desviación típica es

$$\sigma = \sqrt{1346911.79} \simeq 1160.6.$$

El intervalo modal es $[3142, 3963)$, pues en él se agrupa el mayor número de observaciones (es decir, 10).

Finalmente, para calcular la mediana hacemos una representación gráfica de las frecuencias absolutas acumuladas (Figura 5.6).

Como el número total de observaciones es 24, la mediana será aquel valor que divide a la muestra en dos

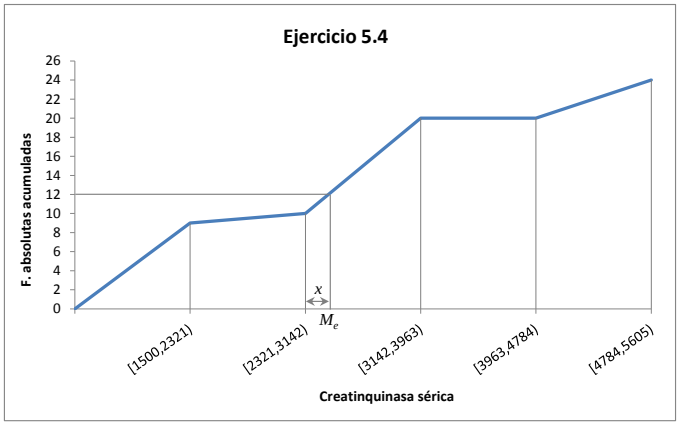

Figura 5.6. Diagrama de frecuencias absolutas acumuladas y cálculo de la mediana para la variable del Ejercicio 5.4.

partes iguales, esto es, en 12 observaciones. Teniendo en cuenta la Figura 5.6 y el Cuadro 5.4, se verifica

$$M_e = 3142 + x,$$

donde x satisface la relación:

$$\frac{12-10}{20-10} = \frac{x}{821}.$$

Luego

$$x = \frac{2 \cdot 821}{10} = 164.2,$$

y, por tanto,

$$M_e = 3142 + 164.2 = 3306.2.$$

c) Los cuartiles $P_{i/4}$, donde i toma los valores 1, 2 ó 3, dividen el número de observaciones en cuatro partes iguales. Así, el tercer cuartil $P_{3/4}$ será aquel valor de la variable que acumula una frecuencia de

$$3 \cdot \frac{N}{4} = \frac{3 \cdot 24}{4} = 18.$$

Siguiendo un procedimiento similar al empleado al calcular la mediana encontramos que

$$P_{3/4} = 3142 + x,$$

donde x satisface ahora la relación

$$\frac{18-10}{20-10} = \frac{x}{821}.$$

Consecuentemente, $x = 656.8$ y

$$P_{3/4} = 3142 + 656.8 = 3798.8.$$

Los deciles dividen la distribución en diez partes iguales. Así pues, el séptimo decil D_7 se corresponde con una frecuencia acumulada de

$$7 \cdot \frac{24}{10} = 16.8.$$

De manera análoga

$$D_7 = 3142 + x,$$

siendo

$$x = \frac{6.8 \cdot 821}{10} = 558.28 \simeq 558.3.$$

Luego,

$$D_7 = 3142 + 558.3 = 3700.3.$$

Finalmente, los centiles P_i, donde i toma valores desde 1 hasta 99, dividen la distribución en cien partes iguales, de forma que P_{25} se corresponderá con aquel valor de la variable que acumula una frecuencia de

$$25 \cdot \frac{24}{100} = 6.$$

Así pues, escribimos

$$P_{25} = 1500 + x$$

donde

$$x = \frac{6 \cdot 821}{9} \simeq 547.3$$

para concluir:

$$P_{25} = 1500 + 547.3 = 2047.3.$$

Nótese que, al ser

$$25 \cdot \frac{1}{100} = \frac{1}{4},$$

el percentil P_{25} es, en realidad, el primer cuartil. □