OCW-UPM Estadística para Ingeniería Civil y Medioambiental

Autores: E. M. García del Toro, C. Hermoso, E. J. Huertas

PROBLEMAS Y EJERCICIOS RESUELTOS

TEMA 1 - ESTADÍSTICA DESCRIPTIVA DE UNA VARIABLE

Ejercicio 1. Ejercicio completo con datos no agrupados (dispuestos en tabla de frecuencias)

Se desea estudiar la población de bacterias coliformes que contaminan el agua de cierta balsa de riego de grandes dimensiones. Para ello, se compone una muestra con 30 extracciones de agua de 1 μm^3 de volumen, realizadas en diferentes sitios de la balsa escogidos al azar, y donde se cuenta el número de bacterias encontradas en cada extracción:

7	10	11	10	9	9	11	11	10	11
12	10	11	7	12	5	11	8	12	10
12	8	12	9	9	11	8	9	10	12

a) Disponer los datos en una tabla de frecuencias de datos sin agrupar.

b) Hallar el número medio de bacterias en la balsa de riego y el error estimado en dicha medida.

c) Hallar los coeficientes de asimetría de Fisher y de apuntamiento. Comente a partir de ellos la forma de la distribución del número de bacterias obtenidas en la muestra, y dibuje el correspondiente diagrama de barras.

d) ¿Qué número máximo de bacterias/μm^3 es superado por el 15 % de las bacterias/μm^3? ¿Qué número mínimo no es superado por el 40 % de las bacterias/μm^3?

e) Dibuje el correspondiente diagrama de caja y bigotes. ¿Hay algún conteo de los 30 realizados que pueda considerarse atípico?

Solución ejercicio 1.

Apartado a) En la tabla pedida hay $k = 7$ datos diferentes, y la última fila es de totales. La variable estadística es $X =$"*número de bacterias coliformes encontradas por* μm^3 *de agua*".

i	x_i	n_i	f_i	N_i
1	5	1	0.0333	1
2	7	2	0.0666	3
3	8	3	0.1	6
4	9	5	0.1667	11
5	10	6	0.2	17
6	11	7	0.2333	24
$k = 7$	12	6	0.2	30
$-$	$-$	$N = 30$	1	$-$

Apartado b) Primero se pide la media muestral o promedio muestral $\bar{x}$. Después, como estimación del error cometido al medir de forma repetida una misma magnitud X, se suele dar la desviación típica muestral, ya que determina la dispersión respecto de la media muestral. Así, para obtener la desviación típica, primero calculamos la varianza

$$S_X^2 = \frac{1}{N}\sum_{i=1}^{k=7}(x_i - \bar{x})^2 n_i = \tag{1}$$

$$= \frac{(5-9.9)^2+(7-9.9)^2\cdot 2+(8-9.9)^2\cdot 3+(9-9.9)^2\cdot 5+(10-9.9)^2\cdot 6+(11-9.9)^2\cdot 7+(12-9.9)^2\cdot 6}{30} = \frac{90.7}{30} = 3.0233.$$

La desviación típica es simplemente la raíz cuadrada de la varianza, y además sus unidades son las mismas que las de la media muestral

$$S_X = \sqrt{3.0233} = 1.7388 \text{ bacterias}/\mu m^3.$$

Otra forma alternativa de calcular la varianza, si conocemos previamente la media $\bar{x}$, es mediante la expresión alternativa

$$\begin{aligned} S_X^2 &= \left(\frac{1}{N}\sum_{i=1}^{k=7} x_i^2 n_i\right) - \bar{x}^2 \\ &= \frac{5^2\cdot 1 + 7^2\cdot 2 + 8^2\cdot 3 + 9^2\cdot 5 + 10^2\cdot 6 + 11^2\cdot 7 + 12^2\cdot 6}{30} - 9.9^2 \\ &= 3.0233 \end{aligned}$$

cuyo resultado es, obviamente igual que en (1).

Apartado c) Ahora nos piden los coeficientes de asimetría y apuntamiento. El cálculo de estos coeficientes implica un gran número de cuentas, por lo que suele realizarse con ordenador. Al ser este un ejemplo ilustrativo, vamos a especificar su cálculo completo con calculadora. Una forma de realizar menos cuentas es fijarnos que las cantidades $(x_i - \bar{x})$ **son las mismas que para el apartado anterior del cálculo de la varianza en (1)**, luego conociendo estos valores solo hay que realizar distintas potencias de los mismos

$$m_3 = \frac{1}{N}\sum_{i=1}^{k=7}(x_i-\bar{x})^3 n_i = \frac{(5-9.9)^3+(7-9.9)^3\cdot 2+(8-9.9)^3\cdot 3+(9-9.9)^3\cdot 5+(10-9.9)^3\cdot 6+(11-9.9)^3\cdot 7+(12-9.9)^3\cdot 6}{30} = -4.192,$$

$$m_4 = \frac{1}{N}\sum_{i=1}^{k=7}(x_i-\bar{x})^4 n_i = \frac{(5-9.9)^4+(7-9.9)^4\cdot 2+(8-9.9)^4\cdot 3+(9-9.9)^4\cdot 5+(10-9.9)^4\cdot 6+(11-9.9)^4\cdot 7+(12-9.9)^4\cdot 6}{30} = 29.575.$$

así

$$A_F = \frac{m_3}{S_X^3} = \frac{-4.192}{(1.7388)^3} = -0.79739, \qquad g_2 = \frac{m_4}{S_X^4} - 3 = \frac{29.575}{(1.7388)^4} - 3 = 0.23539.$$

La distribución presenta una leve asimetría a la izquierda ($A_F < 0$) y es ligeramente leptocúrtica ($g_2 > 0$) (observar que son también valores próximos a cero).

El diagrama de barras de la distribución, que corrobora lo indicado por A_F y g_2, es el siguiente:

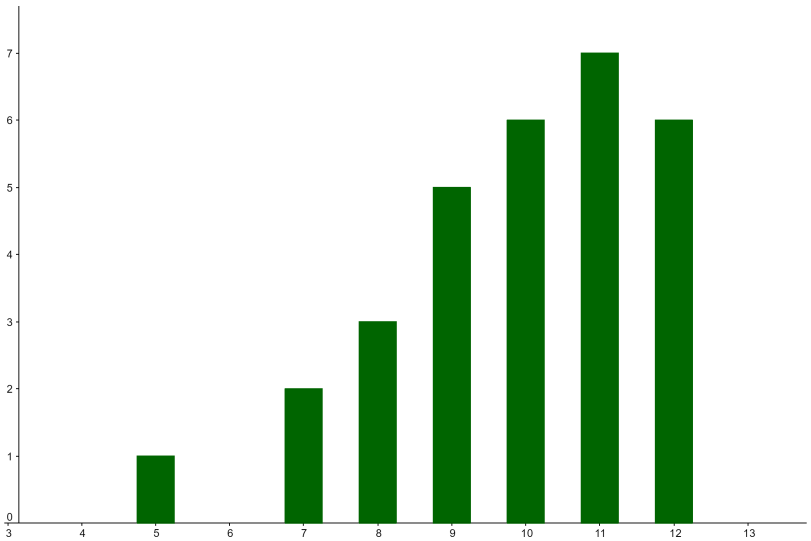

Figure 1: Diagrama de barras

Apartado d) La primera pregunta se refiere a aquél número de bacterias/μm^3 que deja por encima de él tan solo al 15 % de los números de bacterias contadas por μm^3, luego nos están pidiendo el p_{85}. Utilizando la forma de cálculo de percentiles para *datos no agrupados dispuestos en tablas de frecuencias*, que vimos en las clases de teoría, tenemos:

1º) Los datos ya están ordenados de menor a mayor valor

2º) Calculamos la posición de referencia:

$$\frac{r \cdot N}{100} = \frac{85 \cdot 30}{100} = \frac{51}{2} = 25.5.$$

Dado que no es un entero, buscamos directamente el número que lo supera en la columna de las frecuencias absolutas acumuladas. Este número es el $N_7 = 30$, luego el $p_{85} = x_7 = 11$ bacterias/μm^3.

La segunda pregunta se refiere a aquél número de bacterias/μm^3 que no es superado por el 40 % de los números de bacterias contadas por μm^3. Están pidiendo entonces el p_{40}. De igual forma, vemos que

$$\frac{r \cdot N}{100} = \frac{40 \cdot 30}{100} = 12.$$

Buscamos en la columna de frecuencias acumuladas el primer valor que supera 12 y vemos que es el $N_5 = 17$, luego $p_{40} = x_5 = 10$ bacterias/μm^3.

Apartado e) Tenemos que calcular el primer y tercer cuartiles y la mediana. Así para $q_1 = p_{25}$ tenemos $\frac{25 \cdot 30}{100} = \frac{15}{2} = 7.5$. El primer valor que supera a 7.5 en la columna de frecuencias acumuladas es $N_4 = 11$, luego $q_1 = p_{25} = x_4 = 9$ bacterias/μm^3.

Para $q_3 = p_{75}$ tenemos $\frac{75 \cdot 30}{100} = \frac{45}{2} = 22.5$. El primer valor que supera a 22.5 en la columna de frecuencias acumuladas es $N_6 = 24$, luego $q_3 = p_{75} = x_6 = 11$ bacterias/μm^3.

Para la Mediana $Me = q_2 = p_{50}$ tenemos $\frac{50 \cdot 30}{100} = 15$. El primer valor que supera a 15 en la columna de frecuencias acumuladas es $N_5 = 17$, luego $Me = q_2 = p_{50} = x_5 = 10$ bacterias/μm^3.

Obtenemos el recorrido intercuartílico: $iqr = q_3 - q_1 = 11 - 9 = 2$ bacterias/μm^3, y los valores necesarios para la posición de los bigotes en el diagrama

$$\begin{aligned} a_1 &= q_1 - 1.5 \cdot iqr = 9 - 1.5 \cdot 2 = 6, \\ a_3 &= q_3 + 1.5 \cdot iqr = 11 + 1.5 \cdot 2 = 13 \to 12 \text{ (máximo valor de la muestra)} \\ a_1 &= 6, \qquad a_3 = 12, \qquad \textit{atípico} = 5. \end{aligned}$$

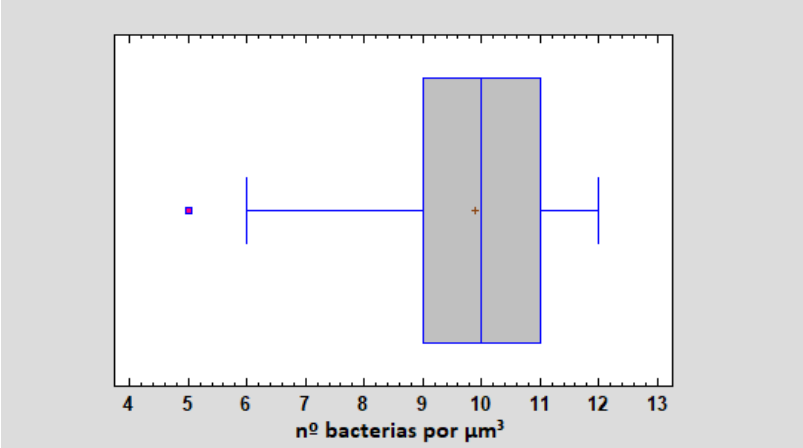

Figure 2: Diagrama de caja y bigotes

Ejercicio 2. Ejercicio completo con datos no agrupados, no dispuestos en tabla de frecuencias.

Se desea estudiar la población de truchas arcoíris que hay en el tanque de una piscifactoría. Para ello se extrae una muesta de 20 truchas y se mide su longitud en centímetros, que figura en la tabla siguiente:

12 13 16 16 16 19 19 20 24 24 24 24 26 26 28 28 28 36 36 44

Calcular:

a) El tamaño medio de las truchas de la muestra y la dispersión de tamaños respecto de la media.

b) Coeficiente de asimetría de Fisher y coeficiente de apuntamiento. Comente a partir de ellos la forma de la distribución de tamaños de las truchas de la muestra.

c) ¿Qué tamaño máximo de trucha es superado por el 15 % de los tamaños? ¿Entre qué tamaños se encuentra el 50 % de las muestras obtenidas en forma central? ¿Qué tamaño mínimo no es superado por el 33 % de los tamaños?

d) Dibuje el correspondiente diagrama de caja y bigotes. ¿Hay algún tamaño de trucha que pueda considerarse atípico en la muestra obtenida?

Solución ejercicio 2.

Apartado a) Nos piden la media muestral $\bar{x}$ y alguna medida de dispersión respecto de la media muestral. Podemos dar, por ejemplo, la varianza o la desviación típica (cualquiera de las dos valdría para contestar a la pregunta)

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N=20} x_i \quad = \frac{12+13+3\cdot 16+2\cdot 19+20+4\cdot 24+2\cdot 26+3\cdot 28+2\cdot 36+44}{20} = \frac{479}{20} = 23.95\ cm,$$

Una primera forma de calcular la varianza es

$$S_X^2 = \frac{1}{N}\sum_{i=1}^{N=20}(x_i-\bar{x})^2 =$$

$$\frac{(-11.95)^2+(-10.95)^2+3\cdot(-7.95)^2+2\cdot(-4.95)^2+(-3.95)^2+4\cdot 0.05^2+2\cdot 2.05^2+3\cdot 4.05^2+2\cdot 12.05^2+20.05^2}{20} \quad = 63.348\ (cm^2). \qquad (2)$$

Vemos que la forma anterior conlleva un buen número de cálculos en los cuales podemos equivocarnos. Habiendo calculado ya la media muestral, existe una forma alternativa del cálculo de la varianza que acarrea un menor número de cuentas. Obteniendo en primer lugar el cuadrado de todos los x_i

$$\sum_{i=1}^{N=20} n_i x_i^2 = 12^2+13^2+3\cdot 16^2+2\cdot 19^2+20^2+4\cdot 24^2+2\cdot 26^2+3\cdot 28^2+2\cdot 36^2+44^2 = 12739$$

tenemos también que

$$S_X^2 = \left(\frac{1}{N}\sum_{i=1}^{N=20} n_i x_i^2\right) - \bar{x}^2 = \frac{12739}{20} - \left(23.95^2\right) = 63.348\ (cm^2),$$

cuyo resultado es, obviamente igual que en (2). La desviación típica muestral es simplemente la raíz cuadrada de la varianza, y además sus unidades son las mismas que las de la media muestral

$$S_X = \sqrt{S_X^2} = \sqrt{63.348} = 7.9591\ cm.$$

Apartado b) Ahora nos piden los coeficientes de asimetría y apuntamiento. El cálculo de estos coeficientes implica un gran número de cuentas, por lo que suele realizarse con ordenador. Al ser este un ejemplo ilustrativo, vamos a especificar su cálculo completo con calculadora. Una forma de realizar menos cuentas es fijarnos que las cantidades $(x_i-\bar{x})$ **son las mismas que para el apartado anterior del cálculo de la varianza de la primera de las formas**, luego conociendo estos valores solo hay que realizar distintas potencias de los mismos

$$m_3 = \frac{1}{N}\sum_{i=1}^{N=20}(x_i-\bar{x})^3 \quad = \frac{(-11.95)^3+(-10.95)^3+3\cdot(-7.95)^3+2\cdot(-4.95)^3+(-3.95)^3+4\cdot 0.05^3+2\cdot 2.05^3+3\cdot 4.05^3+2\cdot 12.05^3+20.05^3}{20} = 347.25,$$

$$m_4 = \frac{1}{N}\sum_{i=1}^{N=20}(x_i-\bar{x})^4 \quad = \frac{(-11.95)^4+(-10.95)^4+3\cdot(-7.95)^4+2\cdot(-4.95)^4+(-3.95)^4+4\cdot 0.05^4+2\cdot 2.05^4+3\cdot 4.05^4+2\cdot 12.05^4+20.05^4}{20} = 12641.$$

$$A_F = \frac{m_3}{S_X^3} = \frac{347.25}{(7.9591)^3} = 0.68873, \qquad g_2 = \frac{m_4}{S_X^4} - 3 = \frac{12641}{(7.9591)^4} - 3 = 0.15011.$$

La distribución presenta una leve asimetría a derechas ($A_F > 0$) y es ligeramente leptocúrtica ($g_2 > 0$) (observar que son también valores próximos a cero)

Apartado c) La primera pregunta se refiere a aquél tamaño de trucha que deja por encima de él tan solo al 15 % de los tamaños, luego nos estan pidiendo el p_{85}. Utilizando la forma de cálculo de percentiles para datos no

agrupados que vimos en las clases de teoría, primero constatamos que los datos ya están ordenados de menor a mayor valor, con posiciones

$i:$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$x_i:$	12	13	16	16	16	19	19	20	24	24	24	24	26	26	28	28	28	36	36	44

A continuación, calculamos el valor de referencia

$$\frac{r \cdot N}{100} = \frac{85 \cdot 20}{100} = 17$$

y dado que el resultado es un número entero, sumamos 0.5 a esa posición $\rightarrow 17.5$ lo que indica que el p_{85} será la media de los datos a ambos lados de esa posición, es decir, la media entre los datos de las posiciones nº 17 y la nº 18

$$p_{85} = \frac{x_{17} + x_{18}}{2} = \frac{28 + 36}{2} = 32 \ cm.$$

La segunda pregunta se refiere a qué intervalo encierra al conjunto ordenado de los datos en una proporción exacta del 50 % en torno a la mediana M_e, repartidos exáctamente en un 25 % a cada lado de la misma (esto es lo que significa "**de forma central**") luego nos estan pidiendo el intervalo entre el primer y el tercer cuartil. Calculando ambos de la misma forma que la pregunta anterior, sabiendo que $q_1 = p_{25}$ y que $q_3 = p_{75}$, tenemos:

$$\frac{r \cdot N}{100} = \frac{25 \cdot 20}{100} = 5 \rightarrow 5.5 \rightarrow \text{media entre los datos de las posiciones nº 5 y nº 6,}$$

$$p_{25} = \frac{x_5 + x_6}{2} = \frac{16 + 19}{2} = 17.5 \ cm,$$

$$\frac{r \cdot N}{100} = \frac{75 \cdot 20}{100} = 15 \rightarrow 15.5 \rightarrow \text{media entre los datos de las posiciones nº 15 y nº 16,}$$

$$p_{75} = \frac{x_{15} + x_{16}}{2} = \frac{28 + 28}{2} = 28 \ cm.$$

Luego el intervalo pedido es $[p_{25}\,;\, p_{75}] = [17.5\,;\, 28]\ cm$.

La tercera pregunta se refiere a aquél tamaño de trucha que no es superado por el 33 % de los tamaños, luego nos estan pidiendo el p_{33}. De igual forma, vemos que

$$\frac{r \cdot N}{100} = \frac{33 \cdot 20}{100} = 6.6 \rightarrow \text{la posición del } p_{33} \text{ es el siguiente entero más grande, posición nº 7}$$

Luego $p_{33} = 19\ cm$.

Apartado d) Tenemos calculado del apartado anterior el primer y tercer cuartiles $q_1 = p_{25} = 17.5\ cm$, y $q_3 = p_{75} = 28\ cm$. La mediana es

$$\frac{r \cdot N}{100} = \frac{50 \cdot 20}{100} = 10 \rightarrow 10.5 \rightarrow \text{media entre los datos de las posiciones nº 10 y nº 11,}$$

$$Me = p_{50} = \frac{x_{10} + x_{11}}{2} = \frac{24 + 24}{2} = 24 \ cm.$$

Obtenemos el recorrido intercuartílico: $iqr = q_3 - q_1 = 28 - 17.5 = 10.5\ cm$ y los valores necesarios para la posición de los bigotes en el diagrama

$$\begin{aligned} f_1 &= q_1 - 1.5 \cdot iqr = 17.5 - 1.5 \cdot 10.5 = 1.75, \\ f_3 &= q_3 + 1.5 \cdot iqr = 28 + 1.5 \cdot 10.5 = 43.75, \\ a_1 &= 12, \qquad a_3 = 36, \qquad atípico = 44. \end{aligned}$$

Ejercicio 3 - Ejercicio completo con datos agrupados en clases.

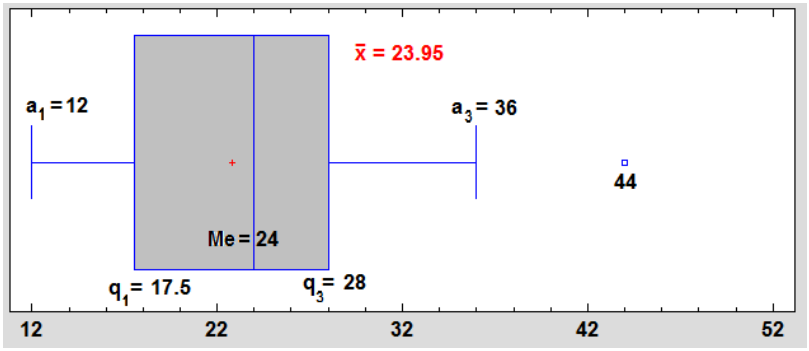

Figure 3: Diagrama de caja y bigotes

Se desea estudiar el tamaño de los tallos de ciertas plantas tropicales. Para ello se extrae una muestra de 20 plantas y se mide su longitud en centímetros. Los datos obtenidos se agrupan en la tabla siguiente:

$Intervalos:$	$[55,62)$	$[62,69)$	$[69,76)$	$[76,83)$	$[83,90]$
$Frecuencias\,(n_i):$	3	0	5	7	5

Calcular:

a) El tamaño medio de los tallos del estudio y la dispersión de tamaños respecto de la media.

b) Dibuje el histograma correspondiente y comente, a partir de él, la forma de la distribución de datos.

c) ¿Qué tamaño máximo de tallo es superado por el 15 % de los tamaños? ¿Qué tamaño mínimo no es superado por el 33 % de los tamaños?

d) ¿Entre que tamaños se encuentra el 50 % de los tallos en forma central? Obtenga también el recorrido intercuartílico, el intervalo modal, la moda y el coeficiente de variación de Pearson.

Solución ejercicio 3:

Apartado a) Ampliamos la tabla del enunciado para que incluya marcas de clase (x_i) y frecuencias acumuladas (N_i)

i	$[L_{i-1}, L_i)$	x_i	n_i	N_i
1	$[55,62)$	58.5	3	3
2	$[62,69)$	65.5	0	3
3	$[69,76)$	72.5	5	8
4	$[76,83)$	79.5	7	15
5	$[83,90]$	86.5	5	20
			$N=20$	

(3)

ahora usamos las marcas de clase y las frecuencias absolutas para calcular

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k=5} x_i n_i \quad = \frac{58.5\cdot 3+65.5\cdot 0+72.5\cdot 5+79.5\cdot 7+86.5\cdot 5}{20} = \frac{1527}{20} = 76.35\ cm,$$

$$S_X^2 = \frac{1}{N}\sum_{i=1}^{k=5} (x_i-\bar{x})^2\, n_i \quad = \frac{(-17.5)^2\cdot 3+(-10.85)^2\cdot 0+(-3.85)^2\cdot 5+(3.15)^2\cdot 7+(10.15)^2\cdot 5}{20} = 78.872.$$

Luego $S_X = \sqrt{S_X^2} = \sqrt{78.872} = 8.881\,cm$.

Apartado b) El histograma correspondiente es

En vista del histograma, puede decirse que su forma relativamente acampanada sugiere que los datos se distribuyen de forma aproximadamente normal. Tomando como eje de simetría una recta vertical que pasa por la media $\bar{x}$ de la distribución, la cola de la izquierda es ligeramente más larga que la de la derecha, por lo que hay una ligera asimetría a izquierdas o negativa (A_F negativo). En cuanto al apuntamiento, la forma acampanada del histograma sugiere que la forma distribución es aproximadamente mesocúrtica (semejante al de una distribución normal).

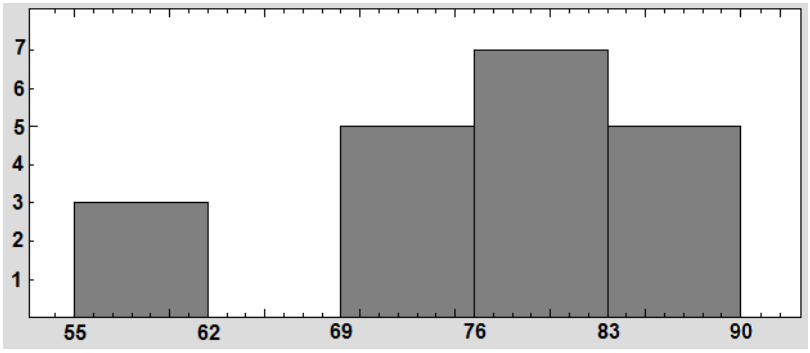

Figure 4: Histograma

Apartado c) La primera pregunta se refiere a aquél tamaño de tallo que deja por encima de él tan solo al 15 % de los tamaños, luego nos estan pidiendo el p_{85}. Utilizando la fórmula

$$p_r = L_{i-1} + \frac{r\frac{N}{100} - N_{i-1}}{n_i} \cdot c$$

debemos buscar el p_{85} en el primer intervalo cuya frecuencia acumulada sobrepase el valor de referencia

$$\frac{r \cdot N}{100} = \frac{85 \cdot 20}{100} = 17$$

y consultando en la tabla (3) ocurre para el intervalo nº 5, el $[83, 90]$, luego

$$p_{85} = L_{i-1} + \frac{r\frac{N}{100} - N_{i-1}}{n_i} \cdot c = 83 + \frac{\left(85 \cdot \frac{20}{100}\right) - 15}{5} \cdot 7 = 85.8 \ cm.$$

La segunda pregunta de este apartado se refiere a aquél tamaño de tallo que no es superado por el 33 % de los tamaños, luego nos estan pidiendo el p_{33}. De igual forma, vemos que tenemos que buscar este percentil en el primer intervalo cuya frecuencia acumulada sobrepase el valor de referencia

$$\frac{r \cdot N}{100} = \frac{33 \cdot 20}{100} = 6.6$$

y consultando de nuevo en la tabla (3), vemos que ocurre para el intervalo nº 3, el $[69, 76)$, luego

$$p_{33} = L_{i-1} + \frac{r\frac{N}{100} - N_{i-1}}{n_i} \cdot c = 69 + \frac{\left(33 \cdot \frac{20}{100}\right) - 3}{5} \cdot 7 = 74.04 \ cm.$$

Apartado d) Para calcular el intervalo en el que se encuentra el 50% de los datos **de forma central** tomamos $[p_{25}\,, p_{75}]$. Utilizando el mismo método que en el apartado anterior tenemos

$$\begin{aligned} q_1 &= p_{25} = 69 + \frac{\left(25 \cdot \frac{20}{100}\right) - 3}{5} \cdot 7 = 71.8 \ cm, \\ q_3 &= 83 + \frac{\left(75 \cdot \frac{20}{100}\right) - 15}{5} \cdot 7 = 83 \ cm. \end{aligned}$$

Por tanto, el interfalo pedido es $[p_{25}\,;\, p_{75}] = [71.8\,;\, 83]$. Los límites de este intervalo encierran los datos en una proporción exacta del 50 % en torno a la mediana

$$Me = p_{50} = 76 + \frac{\left(50 \cdot \frac{20}{100}\right) - 8}{7} \cdot 7 = 78 \ cm,$$

repartidos exáctamente en un 25 % a cada lado de la misma. El recorrido intercuartílico es $iqr = q_3 - q_1 = 83 - 71.8 = 11.2 \ cm$. El intervalo modal es el que tiene la mayor frecuencia absoluta $n_4 = 7$, es decir, es el intervalo nº 4, cuyos límites son $[76, 83)$. La moda viene dada por la fórmula

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c = 76 + \frac{5}{5+5} \cdot 7 = 79.5 \ cm.$$

El coeficiente de variación de Pearson es $CV_X = S_X/|\bar{x}| = (8.881/76.35) \cdot 100 = 11.632\,\%$.