Distribucions (taules) de freqüències

Recordeu que la **freqüència absoluta** d'un valor (n_i) és el número de vegades que es repeteix aquest valor. La llista de freqüencies constitueix la taula de freqüències. La suma de totes les freqüències absolutes és el número total de casos (n).

Quan una variable és numèrica (quantitativa), pot prendre un gran nombre de valors diferents. Per exemple, l'edat d'una persona, els ingressos d'una llar,... En aquest casos, una taula de freqüències on es compten les repeticions de cada valor és poc informativa. En aquests casos, l'opció de treball consisteix en agrupar la variable en intervals i comptar el nombre de casos que estan dins de cada interval.

Exemple: Les següents dades són les notes (puntació sobre 100) obtingudes en un examen,

55.8	60.9	37.0	91.3	65.8
42.3	33.8	60.6	76.0	69.0
45.9	39.1	35.5	56.0	44.6
71.7	61.2	61.5	47.2	74.5
83.2	40.0	31.7	36.7	62.3
47.3	94.6	56.3	30.0	68.2
75.3	71.4	65.2	52.6	58.2
48.0	61.8	78.8	39.8	65.0
60.7	77.1	59.1	49.5	69.3
69.8	64.9	27.1	87.1	66.3

En aquest cas, l'agrupació que té més sentit és la de considerar els intervals $[0, 50)$, $[50, 70)$, $[70, 90)$ i $[90, 100]$ (que correspon a suspens, aprovat, notable i excel.lent).

Aleshores si té més sentit considerar la taula de freqüències,

Nota	n_i	f_i	p_i	n_i	f_i	p_i
$[0, 50)$	17	0.34	34%	17	0.34	34%
$[50, 70)$	22	0.44	44%	39	0.78	78%
$[70, 90)$	9	0.18	18%	48	0.96	96%
$[90, 100]$	2	0.04	4%	50	1	100
Total	50	1	100			

Quan dividim la regió de possibles valors en intervals I_1,I_2, ..., I_m, podem parlar dels límits dels intervals.

Donat un interval $I_i = [L_i, L_{i+1})$, L_i és el **límit inferior** i L_{i+1} és el **límit superior**. La **longitud** o **amplitud** a_i de l'interval és $L_{i+1} - L_i$.

La **marca de classe** x_i de l'interval I_i és el seu punt mig $x_i = \frac{L_i + L_{i+1}}{2}$.

Per exemple, 50 és el límit inferior i 70 és el límit superior de $[50, 70)$, i la seva marca de classe és 60. La seva longitud és 20.

Les marques de classe de l'exemple anterior són 25, 60, 80 i 95.

Com agrupar les dades en intervals o classes: quants i de quina longitud.

Quan no hi ha una divisió lògica o aprapiada pel nostre estudi, el nombre d'intervals en què dividirem els valors depèn del número total de casos.

- Si el número total d'observacions n és inferior o igual a 100 ($n \leq 100$), aleshores el número d'intervals k és l'aproximació entera de $\sqrt{n}$.

- Si el número total d'observacions n és superior a 100 ($n > 100$), aleshores el número d'intervals k és l'aproximació entera de $\frac{\log n}{\log 2}$.

Aleshores cada interval ha de tenir una longitud aproximada a de $a = \frac{X_{max} - X_{min}}{k}$ (la diferència entre el valor màxim i el valor mínim dividit pel nombre d'intervals).

A l'exemple anterior teníem 50 observacions, aleshores com que és menor que 100, el nombre d'intervals és l'aproximació entera de $\sqrt{50} = 7.071$. Per tant, prenem 7 intervals.

El valor més petit és 27.1 i el més gran és 94.6. Aleshores l'amplitud $\frac{94.6-27.1}{7} = 9.64$. Aleshores podem prendre 9.7. Com que hem estat arrodonint, podem començar en 27.

Els intervals són $[27, 27+9.7 = 36.7)$, $[36.7, 36.7+9.7 = 46.4)$, $[46.4, 56.1)$, $[56.1, 65.8)$, $[65.8, 75.5)$, $[75.5, 85.2)$, i $[85.2, 94.9]$.

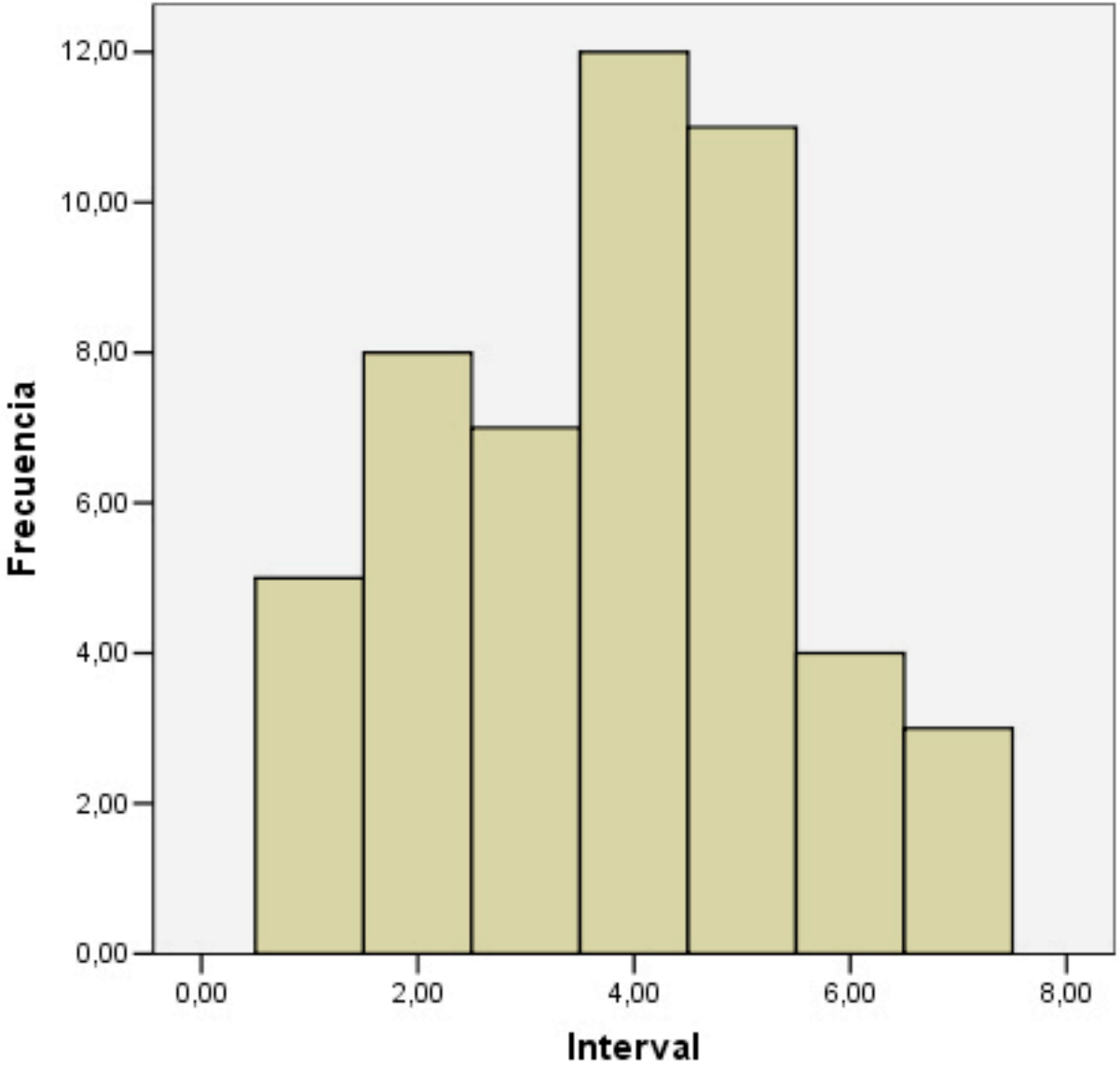
La taula de freqüència és

	n_i	f_i	p_i	N_i	F_i	P_i
[27, 36.7)	5	0.1	10%	5	0.1	10%
[36.7,46.4)	8	0.16	16%	13	0.26	26%
[46.4,56.1)	7	0.14	14%	20	0.4	40%
[56.1,65.8)	12	0.24	24%	32	0.64	64%
[65.8,75.5)	11	0.22	22%	43	0.86	86%
[75.5,85.2)	4	0.08	8%	47	0.94	94%
[85.2,94.9]	3	0.06	6%	50	1	100%
	50	1	100%			

Un 86% dels alumnes ha tret com molt un 75.5, o bé, un 14% dels alumnes ha tret més d'un 75.5.

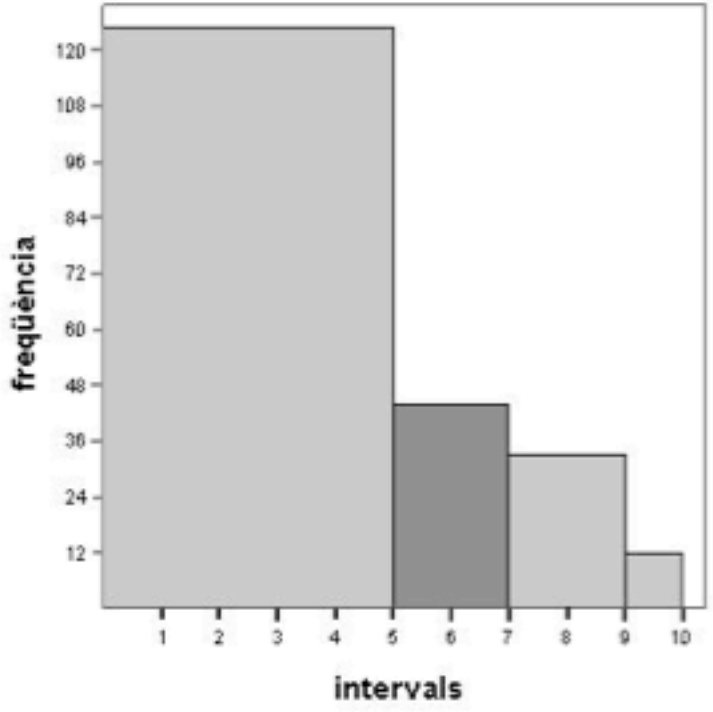
Representació gràfica d'una variable quantitativa

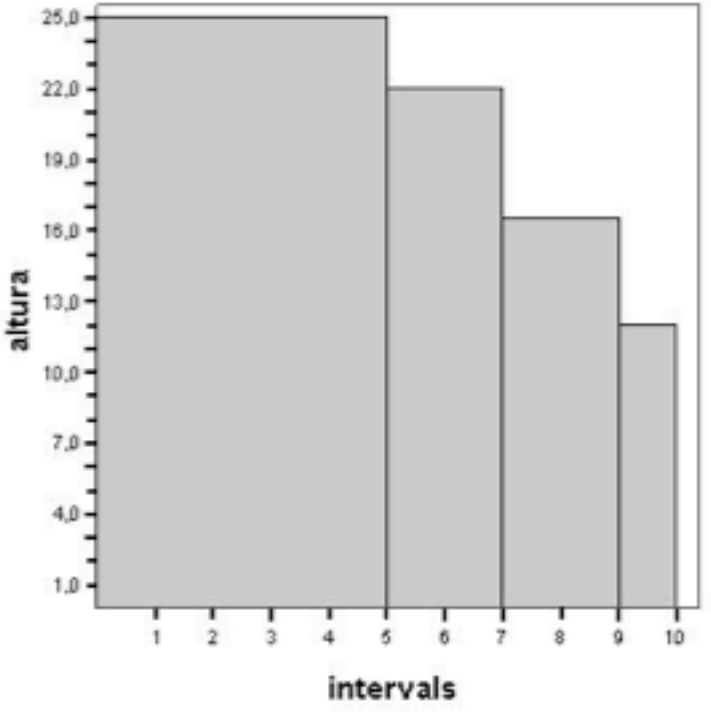
- **L'histograma** Consisteix en aixecar, per cada interval, un rectangle d'àrea proporcional a la freqüència.


12,00
10,00
8,00
6,00
4,00
2,00
0,00
Frecuencia
0,00
2,00
4,00
6,00
8,00
Interval
Media =3,80
Desviación típica =1,
66599
N =50
Casos ponderados por Freq

Quan els intervals no són tots de la mateixa amplada, cal anar amb compte al calcular l'alçada de cada barra. El motiu és que l'area ha de ser proporcional a la freqüència (sinó la impressió que donaria el gràfic seria errònia).

L'alçada h de cada rectangle dependrà de la freqüència i de la amplitud de l'interval en qüestió. L'alçada serà $h = \frac{n}{a}$, la freqüència dividida per l'amplitud i s'anomena **freqüència per unitat d'amplitud**.


freqüència
120
108
96
84
72
60
48
36
24
12
1
2
3
4
5
6
7
8
9
10
intervals


altura
25,0
22,0
19,0
16,0
13,0
10,0
7,0
4,0
1,0
1
2
3
4
5
6
7
8
9
10
intervals

Resumint, si la nostra variable és numèrica, podem representar-la dividint el rang dels seus valors en intervals i comptant la freqüencia de cada interval.

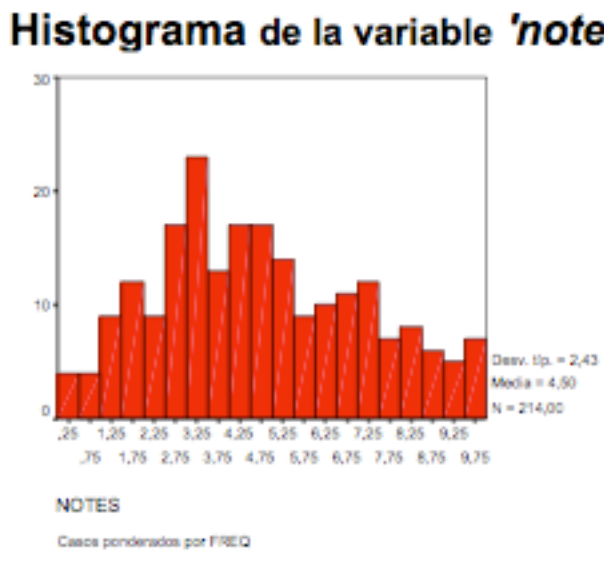
L'histograma ens serveix per representar una variable numèrica agrupada en intervals. A sobre de cada interval s'aixeca un rectangle d'àrea proporcional a la freqüència.

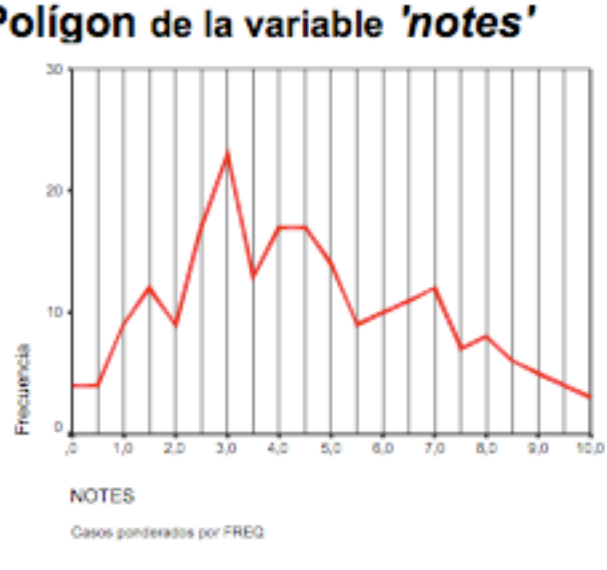
- **Tots els intervals tenen la mateixa longitud**: aleshores l'alçada de cada rectangle és la freqüència de l'interval.

- **Els intervals tenen longituds diferents**: aleshores l'alçada de cada rectangle és la **freqüència per unitat d'amplitud**, és a dir, cal dividir la freqüència de l'interval per la seva amplitud.

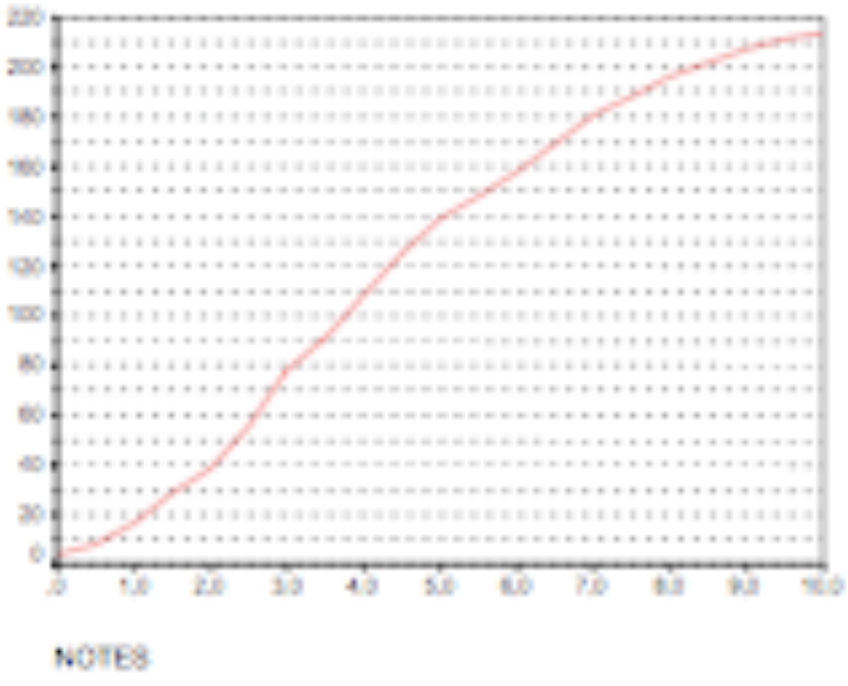
- **El polígon de freqüencies** Es construeix a partir de l'histograma, unint mitjançant una línia poligonal els punts mitjans de les bases superiors dels rectangles. També es coneix com **gràfica de línies**.

(b) Histograma i polígon de la variable *notes*:





- **L'ogiva o polígon de freqüències acumulades** És una línia poligonal però sempre creixen ja que representa les freqüències acumulades.



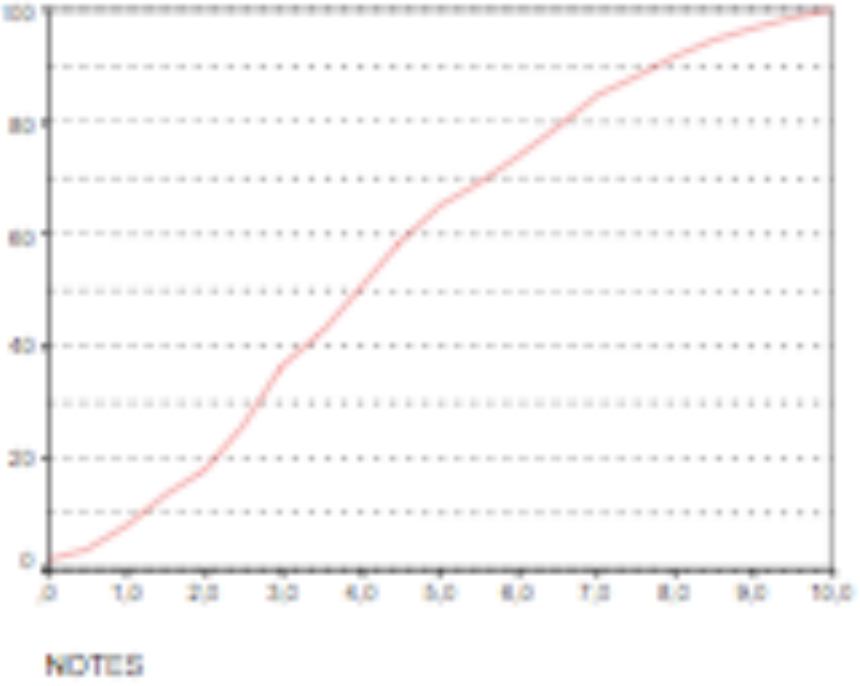


Figura 1.10: Ogives de la variable notes, que representen freqüències i percentatge acumulats.