

Rappresentazioni grafiche

Su una popolazione di $n = 20$ unità sono stati rilevati i seguenti fenomeni:

- stato civile (X)
- livello di scolarità (Y)
- numero di figli a carico (Z)
- reddito in migliaia di € (W)

$$X = \begin{cases} N & = \text{Nubile} \\ C & = \text{Coniugato} \\ V & = \text{Vedovo} \\ S & = \text{Separato, divorziato} \end{cases}$$

$$Y = \begin{cases} A & = \text{Analfabeta, alfabeto} \\ O & = \text{Scuola dell'obbligo} \\ S & = \text{Diploma di scuola superiore} \\ L & = \text{Laurea e superiore} \end{cases}$$

u	X	Y	Z	W
unità stat.	stato civile	grado di scolarità	numero di figli	reddito in €
1	N	L	0	72.50
2	S	O	1	54.28
3	V	A	3	50.02
4	V	O	4	88.88
5	C	L	1	62.30
6	N	S	1	45.21
7	C	S	0	57.50
8	C	O	2	78.40
9	V	L	3	75.13
10	N	O	0	58.00
11	N	S	1	53.70
12	N	A	0	91.29
13	S	S	1	74.70
14	C	S	4	41.22
15	N	S	3	65.20
16	C	L	0	63.58
17	V	O	2	48.27
18	S	O	2	52.52
19	C	S	4	69.50
20	C	S	4	85.98

Problemi

1. Come possono essere classificati i fenomeni X , Y , Z e W ?
 2. Costruire le tabelle di frequenza: relative, percentuali e cumulate;
 3. Rappresentare in modo opportuno i dati;
- stato civile (X) qualitativo nominale
si presenta con $k = 4$ modalità

$$x_1 = N \quad x_2 = C \quad x_3 = V \quad x_4 = S$$

La tabella delle frequenze è

x_i	n_i	$f_i = n_i/n$	$p_i = f_i \cdot 100\%$
N	6	0.30	30
C	7	0.35	35
V	4	0.20	20
S	3	0.15	15
	$n = 20$	1.00	100

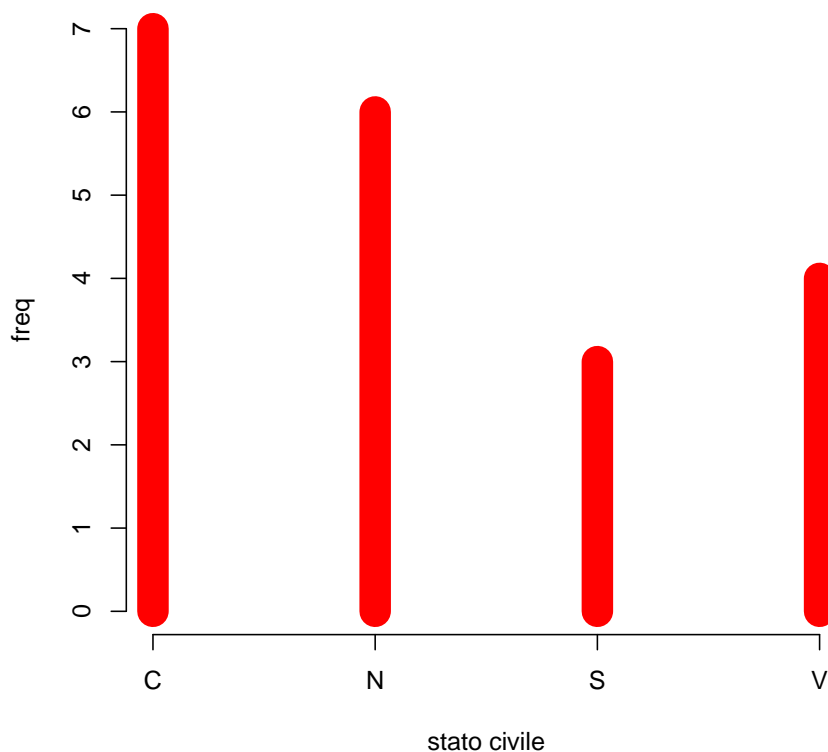
Un carattere *qualitativo sconnesso* può essere rappresentato graficamente in diversi modi:

- *tramite rettangoli*
- *grafici a torta*
- *rappr. tramite figure*

Rappresentazione tramite rettangoli

Le modalità x_1, x_2, \dots, x_k del carattere si sistemano su un segmento orizzontale in *qualsiasi ordine* e in *modo equispaziato*.

In corrispondenza di ciascuna modalità si disegnano rettangoli di *stessa base* e altezza *proporzionale* alle frequenze n_i, f_i o p_i .

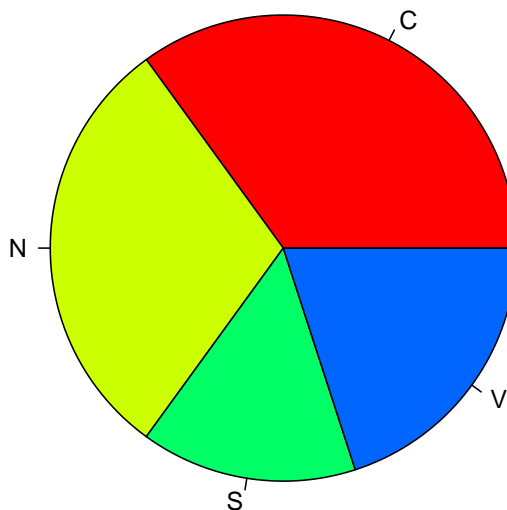


Grafici o diagrammi a torta

Si costruisce un cerchio e si identificano dei settori circolari la cui ampiezza (o la cui area) è *proporzionale* alle frequenze n_i , f_i o $f_i \cdot 100\%$.

Anche in questo caso i settori vengono disegnati in un ordine qualsiasi.

stato civile



Rappresentazione tramite figure

Si sceglie una *figura* per rappresentare l'unità di misura:

$$\begin{array}{c} \circ \quad \circ \\ | \\ \smile \end{array} = 1$$

Si rappresentano le modalità del carattere riportando un numero di figure *proporzionale* alle frequenze n_i , f_i o $f_i \cdot 100\%$.

Anche in questo caso le modalità vengono sistemate in un ordine qualsiasi.

N		30%
C		35%
S		15%
V		20%

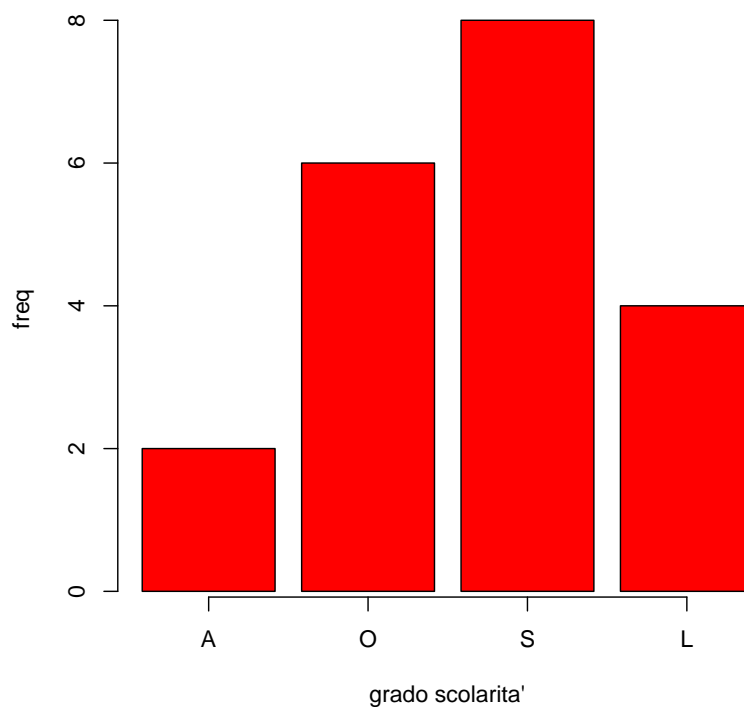
- livello di scolarità (Y) qualitativo ordinale

Le modalità con cui si presenta il fenomeno sono $k = 4$

$$x_1 = A \quad x_2 = O \quad x_3 = S \quad x_4 = L$$

La tabella delle frequenze è:

x_i	n_i	f_i	p_i	N_i	F_i
A	2	0.1	10	2	0.1
O	6	0.3	30	8	0.4
S	8	0.4	40	16	0.8
L	4	0.2	20	20	1.0
	20	1.0	100	—	—

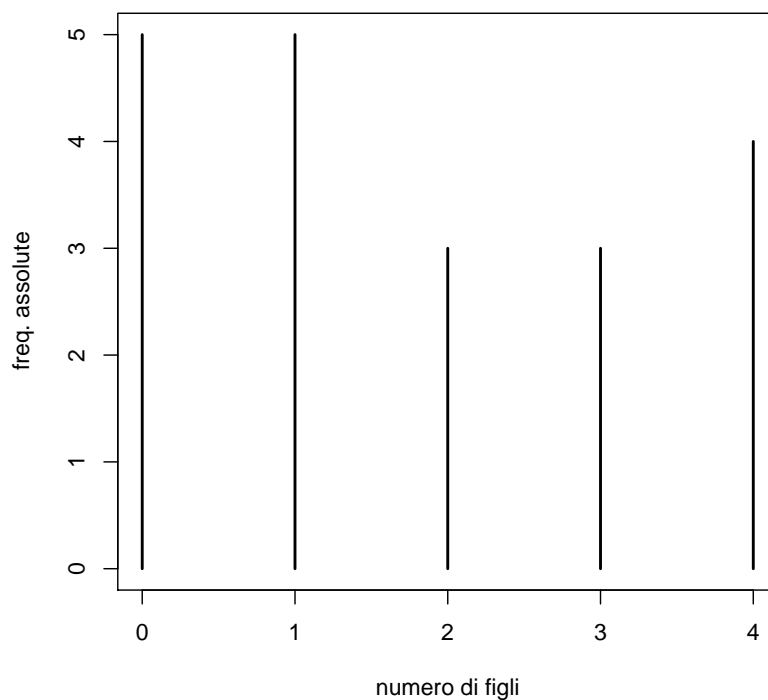


- numero di figli a carico (Z) quantitativo discreto (misurabile su scala di rapporto, lo zero naturale esiste).
Le intensità con cui si presenta il fenomeno sono $k = 5$

$$x_1 = 0 \quad x_2 = 1 \quad x_3 = 2 \quad x_4 = 3 \quad x_5 = 4 \text{ (o più)}$$

La tabella delle frequenze è:

x_i	n_i	f_i	p_i	N_i	F_i
0	5	0.25	25	5	0.25
1	5	0.25	25	10	0.50
2	3	0.15	15	13	0.65
3	3	0.15	15	16	0.80
4	4	0.20	20	20	1.00
	20	1.00	100		



Attenzione : *per i fenomeni quantitativi, l'asse su cui rappresentano i dati è di tipo numerico, per cui si deve prestare attenzione a come si rappresentano i dati rispettando l'unità di misura dell'asse.*

- reddito in $\text{€}(W)$ quantitativo continuo

Le intensità con cui si presenta il fenomeno sono tutte distinte ($k = 20$)

Ricorriamo allora ad un raggruppamento dei dati in classi. Introduciamo:

a_i : l' ampiezza di ciascuna classe

$l_i = n_i/a_i$: la *densità di frequenza*

La tabella delle frequenze è:

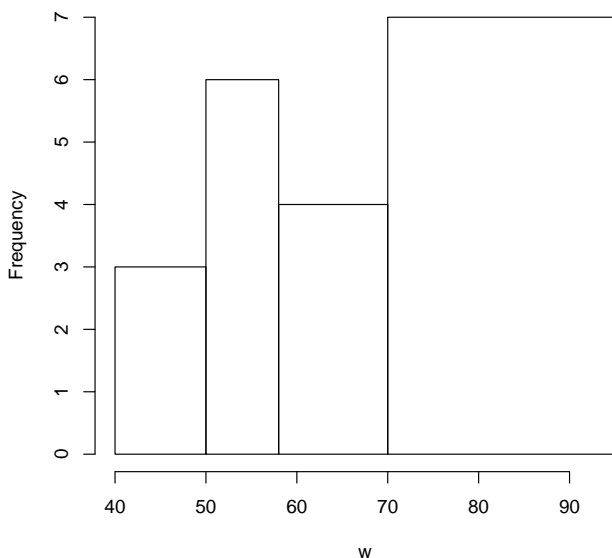
x_i	n_i	f_i	N_i	a_i	l_i
40 – 50	3	0.15	3	10	0.30
50 – 58	6	0.30	9	8	0.75
58 – 70	4	0.20	13	12	0.33
70 – 95	7	0.35	20	25	0.28
	20	1.00			

Rappresentazione tramite istogrammi

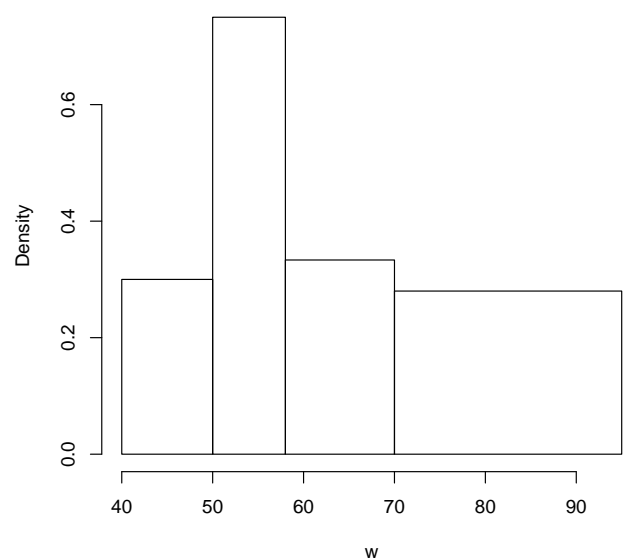
Quando abbiamo un fenomeno quantitativo continuo con dati raggruppati in classi si costruisce un istogramma procedendo come segue:

1. Si dispongono i valori degli estremi degli intervalli delle classi sull'asse delle ascisse ripetendo l'unità di misura dell'asse
2. si tracciano dei rettangoli avendo come base gli estremi dell'intervallo e come altezza la densità di frequenza l_i . **Attenzione** : utilizzare le frequenze n_i , f_i o p_i può portare a grafici completamente sballati.

Istogramma sbagliato!



Istogramma corretto



Indici di posizione

Vogliamo descrivere alcune caratteristiche delle distribuzioni o farne confronti attraverso degli indici di sintesi

- esistono un valore o dei valori attorno ai quali si aggregano i dati?
- quando questo avviene, quanto i dati sono sparpagliati attorno a tali valori?

La moda

Può essere calcolata per qualunque fenomeno

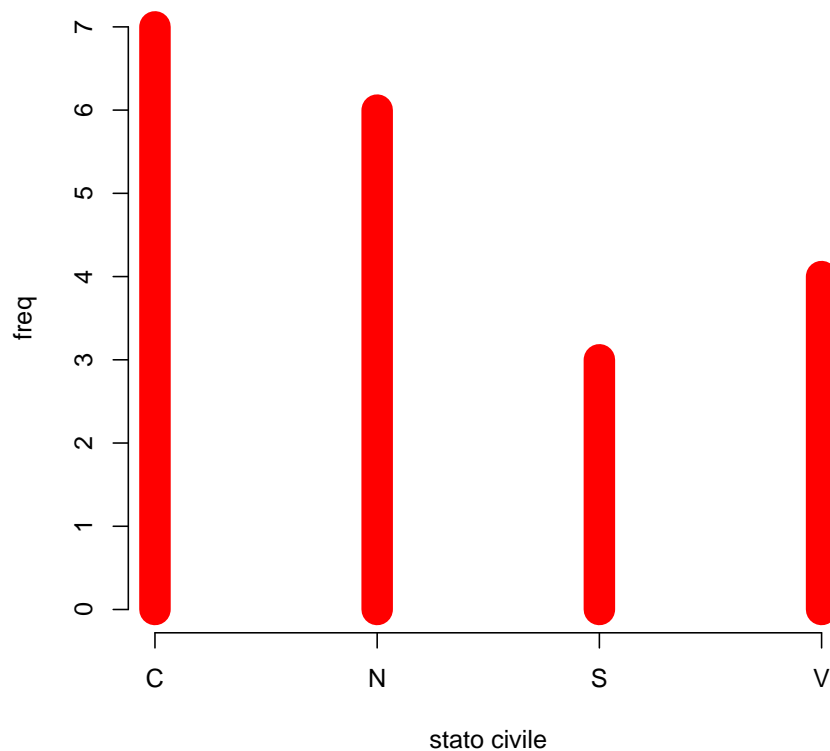
Moda

valore x_i di una distribuzione con frequenza n_i (f_i o p_i) massima o, se il fenomeno è raggruppato in classi, il punto medio dell'intervallo con densità di frequenza l_i più elevata.

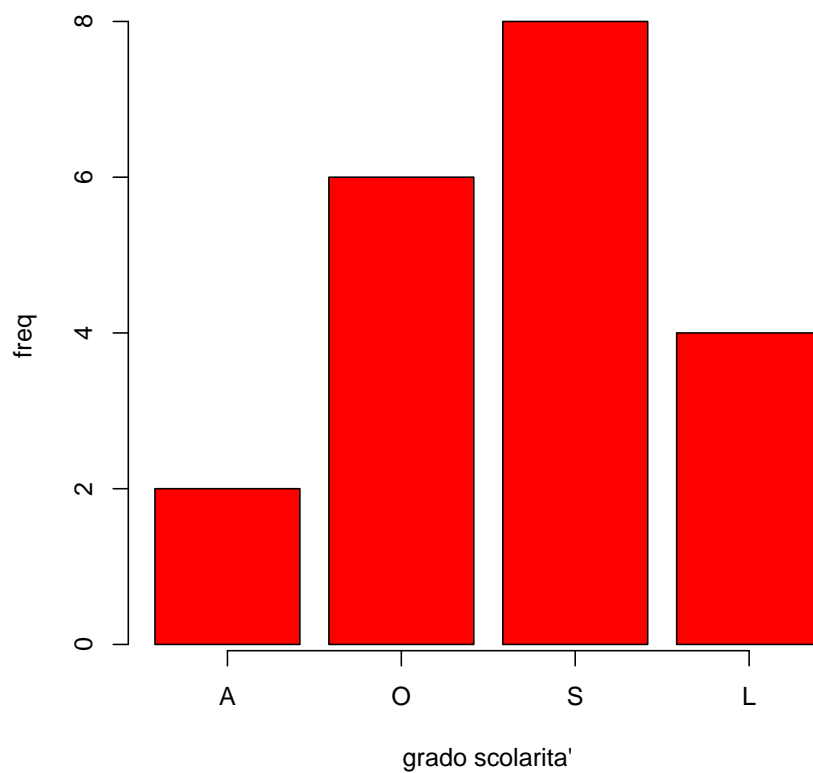
Se esistono più mode si parla di distribuzione *plurimodale*

Calcoliamo la moda per i caratteri X , Y , Z e W

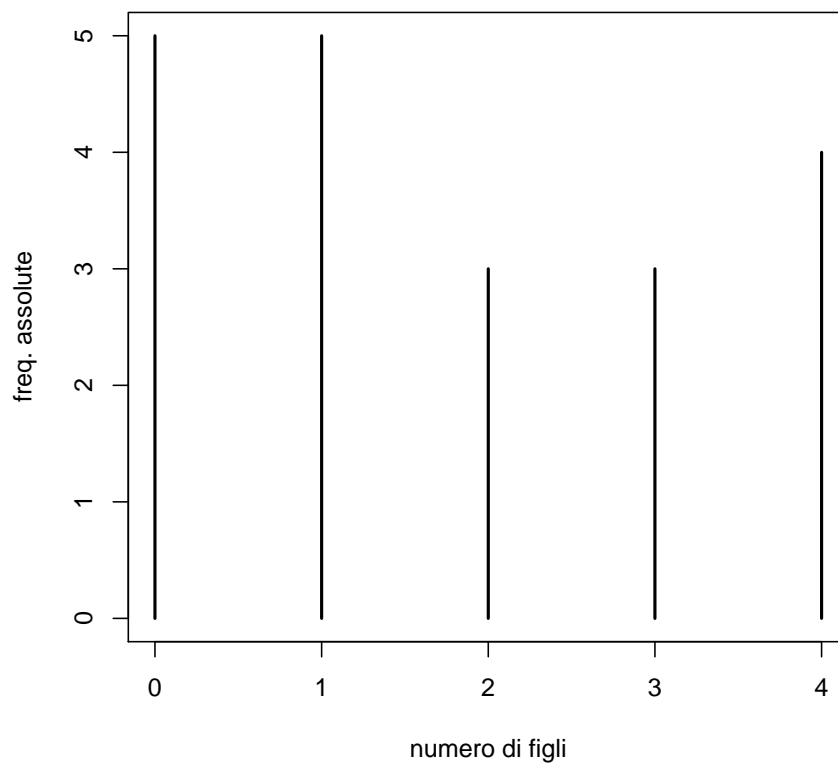
Carattere X



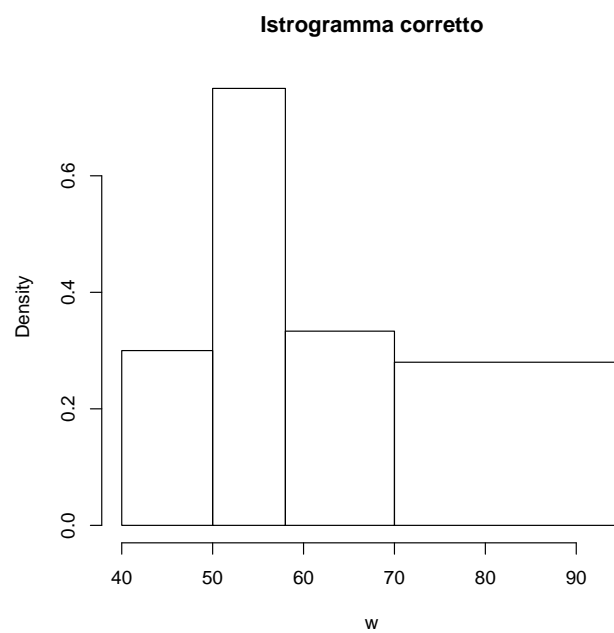
La modalità $x_i = C$ ha frequenza più elevata e pari a 7. Quindi la moda è C .

Carattere Y 

La modalità che si presenta più frequentemente è $x_i = S$, frequenza 8.

Carattere Z 

I valori $x_i = 0$ e $x_i = 1$ hanno entrambi frequenza massima pari a 5. Z ha quindi una distribuzione *bimodale*.

Carattere W 

La modalità cui corrisponde l_i più alta è la classe $50 - 58$.
Quindi il valore corretto per la moda è $(50+58)/2 = 54$

La Mediana

La mediana è quel valore che, una volta **ordinati** i dati del campione, lascia alla sua sinistra e alla sua destra la metà del campione

La mediana divide a metà la distribuzione dei dati

Può essere calcolata solo per fenomeni ordinabili

Mediana (per dati singoli)

Se n è l'ampiezza del campione, si procede così:

- 1) si ordinano i dati in ordine crescente
- 2) si calcola il valore $(n + 1)/2$
- 3a) se esiste $(n + 1)/2$ (caso n dispari) la mediana è quel valore.
- 3b) se $(n + 1)/2$ non è un numero intero (caso n pari)
 - fenomeno quantitativo: si fa la media tra il valore precedente e quello successivo alla posizione $(n + 1)/2$
 - fenomeno qualitativo: si confrontano le modalità di posto precedente e successivo alla posizione $(n + 1)/2$ e se coincidono quella è la mediana, altrimenti la mediana è indeterminata.

Calcoliamo la mediana per il carattere Y

x_i	n_i	N_i	F_i
A	2	2	0.1
O	6	8	0.4
S	8	16	0.8
L	4	20	1.0
	20		

Ordiniamo i dati della distribuzione

AAOOOOOOO |S| |S| SSSSSLLLL

$$\frac{n+1}{2} = 10.5$$

$n/2 = 10$ $n/2 + 1 = 11$ si trova la modalità S

Le frequenze cumulate ci aiutano a trovare il valore della mediana

È sufficiente conoscere in corrispondenza di quale valore si ha $F_i = 0.5$, quel valore è la mediana

In questo caso $F_i = 0.5$ casca nella modalità S

Calcoliamo la mediana per la seguente distribuzione

x_i	n_i	N_i	F_i
A	4	4	0.2
O	6	10	0.5
S	8	18	0.9
L	2	20	1.0
	20		

Ordiniamo i dati della distribuzione

AAAAOOOOOO |O| |S| SSSSSSLL

$n/2 = 10$ troviamo O ma $n/2 + 1 = 11$ c'è S

Poiché $O \neq S$ concludiamo che *la mediana è indeterminata*

$F_i = 0.5$ si verifica al limite tra una modalità e la successiva: la mediana risulta indeterminata

Calcoliamo la mediana per il carattere Z

x_i	n_i	f_i	N_i	F_i	
0	5	0.25	5	0.25	
1	5	0.25	10	0.50	
2	3	0.15	13	0.65	
3	3	0.15	16	0.80	
4	4	0.20	20	1.00	
	20	1.00	100		

$n/2 = 10$ corrisponde a 1 $n/2 + 1 = 11$ corrisponde a 2

In questo caso $F_i = 0.5$ si verifica al limite tra una modalità e la successiva: la mediana risulta la media aritmetica di questo due valori

La mediana è

$$Me = \frac{1 + 2}{2} = 1.5$$

Calcoliamo la mediana per la seguente distribuzione

x_i	n_i	f_i	N_i	F_i	
0	5	0.25	5	0.25	
1	6	0.30	11	0.55	
2	2	0.10	13	0.65	
3	3	0.15	16	0.80	
4	4	0.20	20	1.00	
	20	1.00	100		

$n/2 = 10$ corrisponde a 1 $n/2 + 1 = 11$ corrisponde a 1

In questo caso $F_i = 0.5$ casca nella modalità 1: la mediana risulta 1

La mediana per i dati raccolti in classi

Si cerca l'intervallo con frequenza cumulata

$$N_i = (n + 1)/2 \text{ o } F_i = 0.5$$

Se questo valore cade a cavallo di due classi contigue, si sceglie come mediana il valore che separa le due classi

Se non accade la mediana è data da

$$Me = x_i + \frac{\frac{n}{2} - N_{i-1}}{l_i}$$

x_i è l'estremo inferiore della classe in cui *casca* $F_i = 0.5$

N_{i-1} è la frequenza cumulata della classe precedente quella in cui *casca* $F_i = 0.5$

Calcoliamo la mediana per il carattere W

x_i	n_i	f_i	F_i	N_i	a_i	l_i
40 – 50	3	0.15	0.15	3	10	0.30
50 – 58	6	0.30	0.45	9	8	0.75
58 – 70	4	0.20	0.65	13	12	0.33
70 – 95	7	0.35	1.00	20	25	0.28
	20	1.00				

$$(n + 1)/2 = 10.5$$

La classe che contiene la mediana è la numero 3, 58 – 70

$$Me = x_3 + \frac{\frac{n}{2} - N_2}{l_3} = 58 + \frac{10 - 9}{0.33} = 58 + 3 = 61.$$

Quartili

- $Q1$ è il valore tale che a sinistra, c'è il 25% dei dati (e il rimanente 75% a destra)
- $Q2 = Me$ è il valore che divide a metà la distribuzione
- $Q3$ è il valore tale che a sinistra, c'è il 75% dei dati (e il rimanente 25% a destra)

Quindi tra $Q1$ e $Q3$ vi si trova il 50% dei dati: è il 50% *centrale* dell'intera distribuzione dei dati.

Primo quartile $Q1$

Se n è l'ampiezza del campione, si procede come segue:

- 1) Si calcola il valore $\frac{1}{4} \cdot (n + 1)$.
- 2) si procede come per la mediana tenendo come riferimento sempre la posizione $\frac{1}{4} \cdot (n + 1)$.

Per i dati in classe la formula è

$$Q1 = x_i + \frac{\frac{n}{4} - N_{i-1}}{l_i}$$

Terzo quartile Q_3

Se n è l'ampiezza del campione, si procede come segue:

- 1) Si calcola il valore $\frac{3}{4} \cdot (n + 1)$.
- 2) si procede come per la mediana tenendo come riferimento sempre la posizione $\frac{3}{4} \cdot (n + 1)$.

Per i dati in classe la formula è

$$Q_3 = x_i + \frac{\frac{3}{4} \cdot n - N_{i-1}}{l_i}$$

Esempio Calcoliamo Q_1 e Q_3 per il carattere Z

$$\frac{1}{4} \cdot (n + 1) = 5.25$$

Le posizione 5 cade in 0, la posizione 6 cade in 1

$$Q_1 = \frac{0 + 1}{2} = 0.5$$

$$\frac{3}{4} \cdot (n + 1) = 15.75$$

Le posizione 15 e la posizione 16 cadono in 3

$$Q_3 = 3$$

Percentili

- C_p con $p = 1, \dots, 100$ è il valore tale che a sinistra, c'è il $p\%$ dei dati (e il rimanente $(1 - p)\%$ a destra)
- C_5 è il valore tale che a sinistra, c'è il 5% dei dati (e il rimanente 95% a destra)

Quindi tra C_5 e C_{95} vi si trova il 90% dei dati: è il 90% *centrale* dell'intera distribuzione dei dati.

p-percentile C_p

Se n è l'ampiezza del campione, si procede come segue:

- 1) Si calcola il valore $\frac{p}{100} \cdot (n + 1)$.
- 2) si procede come per i quartili tenendo come riferimento sempre la posizione $\frac{p}{100} \cdot (n + 1)$.

Per i dati in classe la formula è

$$C_p = x_i + \frac{\frac{p}{100} \cdot n - N_{i-1}}{l_i}$$

La media aritmetica

La *media aritmetica* o semplicemente la *media* è uno degli indici maggiormente impiegati dagli statistici.

$$\bar{x}_n = \frac{1}{n} (x_1 + x_2 + \cdots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

Si calcola solo per dati numerici cioè per fenomeni quantitativi

Media aritmetica

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^k (x_i \cdot n_i) = \sum_{i=1}^k (x_i \cdot f_i)$$

dove k è il numero di valori diversi assunti dal fenomeno statistico.

Esempio Calcoliamo la media di Z

$$\bar{z}_n = \frac{1}{n} \sum_{i=1}^k (z_i \cdot n_i)$$

$$\begin{aligned} \bar{z}_n &= \frac{0 \cdot 5 + 1 \cdot 5 + 2 \cdot 3 + 3 \cdot 3 + 4 \cdot 4}{20} \\ &= \frac{5 + 5 + 6 + 9 + 16}{20} = \frac{36}{20} \\ &= 1.8 \end{aligned}$$

Esempio Calcoliamo la media di W

$$\bar{w}_n = \sum_{i=1}^k (\tilde{w}_i \cdot f_i), \quad \tilde{w}_i = \frac{w_i + w_{i+1}}{2}$$

$$\begin{aligned} \bar{w}_n &= \frac{40 + 50}{2} 0.15 + \frac{50 + 58}{2} 0.30 + \frac{58 + 70}{2} 0.20 + \\ &\quad + \frac{70 + 95}{2} 0.35 \\ &= 45 \cdot 0.15 + 54 \cdot 0.30 + 64 \cdot 0.20 + 82.5 \cdot 0.35 \\ &= 63.875 \end{aligned}$$

Media o Mediana?

Supponiamo di avere 3 valori

10 20 30

la media è $\frac{(10+20+30)}{3} = 20$ la mediana è 20

Cambiamo ora un solo valore

10 20 300

la media diventa $\frac{10+20+300}{3} = 110$ la mediana resta sempre 20

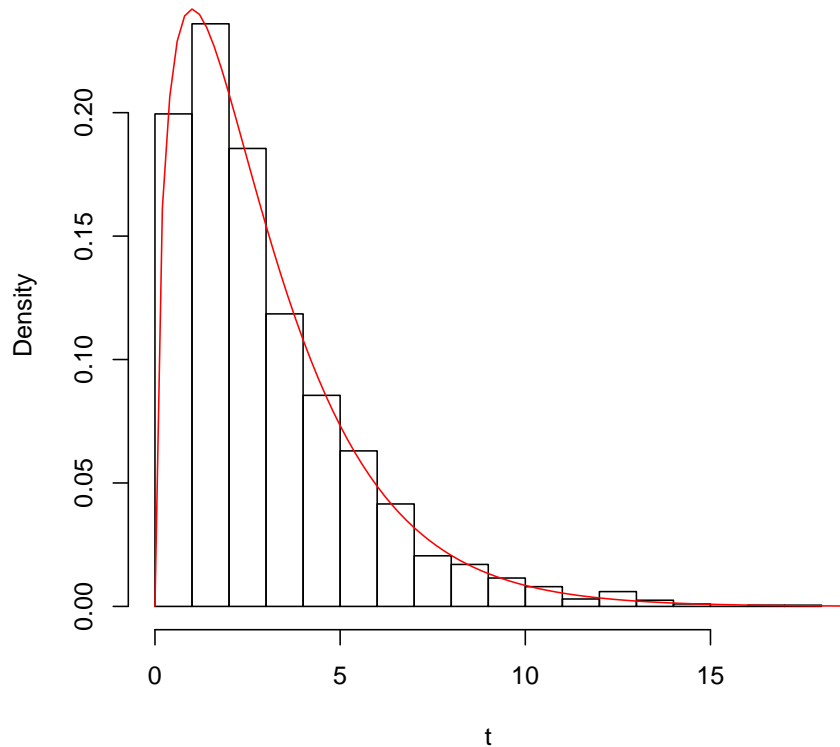
Lo stesso accade se cambiamo il valore di sinistra

0 20 30

la media diventa $\frac{0+20+30}{3} = 16.\overline{6}$ la mediana resta sempre 20

Morale: quando variamo i valori *estremi* di una distribuzione, la media aritmetica ne risente mentre la mediana no. Si dice allora che *la mediana è un indice del centro di una distribuzione più robusto della media aritmetica.*

Esempio



Distribuzione del tempo di vita dei pazienti che hanno subito un trapianto di organo vitale. In ordinata sono riportate le densità di frequenza l_i e in ascissa i tempi di vita t .

La media aritmetica dei tempi è 10 anni

La moda è 2 anni

La mediana è 2.3 anni (2 anni e poco più di 3 mesi)

$Q1 = 1.21$, $Q3 = 4.12$