CONCETTI di BASE

Carattere ***X*** [o *A*]	caratteristica ***quantitativa*** [o *qualitativa*] rappresentativa di un "fenomeno" sottoposto ad indagine
Popolazione	insieme dei soggetti portatori del "carattere" in esame (ossia TUTTI gli *elementi della popolazione).*
Campione	parte degli "elementi della popolazione"
Unità Statistica di Rilevazione	è un "elemento" della popolazione, ed è il *soggetto su cui viene rilevata la* ***caratteristica*** *di interesse*
Modalità $\mathbf{x_i}$ [o a_i] con i = 1, 2, 3, …n	determinazione di un "carattere", ossia il modo in cui si manifesta il carattere stesso al momento della rilevazione
Frequenza y_i	*[frequenza assoluta*] = numero di casi in cui si presenta una data *modalità* di un ***carattere:*** *frequenza della modalità* $\boldsymbol{x_i}$ ***di X*** [o della modalità a_i *di A*]
Totale frequenze	$\sum_{i=1}^{n} \mathbf{y_i} = \mathbf{N}$ indica il "totale dei casi"
Frequenza relativa $\mathbf{f_i = y_i / \sum_{i=1}^{n} y_i}$	"*frazione di casi*" in cui è stata rilevata la modalità corrispondente x_i $\sum_{i=1}^{n} \mathbf{f_i = 1}$
Percentuale $\mathbf{p_i}$	frazione di **100 soggetti**, su cui è stata rilevata la modalità x_i . $\mathbf{p_i = f_i \times 100}$

TIPOLOGIE di FREQUENZE

♣ Sulla base dei dati riportati nella tabella seguente, si chiede di calcolare le percentuali di soggetti corrispondenti a ciascuna delle classi di età considerate; si chiede inoltre di calcolare i valori accumulati delle frequenze assolute e spiegarne il significato .

Sulla base di quanto richiesto utilizziamo il seguente "prospetto di calcolo":

Tav. 1

x_i	y_i	$f_i = \frac{y_i}{\sum_{i=1}^{n} y_i}$	$p_i = f_i \times 100$	$F_i = \sum_{h=1}^{i} y_h$
22 ⊢ 26	13	**0,1857**	18,57	13
26 ⊢ 30	22	**0,3143**	31,43	35
30 ⊢ 34	12	**0,1714**	17,14	47
34 ⊢ 38	10	**0,1429**	14,29	57
38 ⊢ 42	6	**0,0857**	8,57	63
42 ⊢ 46	2	**0,0286**	2.86	65
46 ⊢ 50	2	**0,0286**	2,86	67
50 ⊢ 54	3	**0,0428**	4,28	70
Totali	**70**	**1,0000**	**100,00**	//

Ricordando che le ***percentuali*** vanno calcolate moltiplicando le frequenze relative per 100, i valori richiesti sono riportati nella quarta colonna della tabella, dove :

$$p_i = f_i \times 100$$

e $$\sum_{i=1}^{n} p_i = 100$$

Sul significato dei risultati ottenuti si può dire che:

- Le **frequenze assolute** (y_i) esprimono l'ammontare dei soggetti che possiedono la "caratteristica" espressa dalla corrispondente modalità (x_i).

- Le **frequenze assolute cumulate** (F_i) esprimono l'ammontare dei soggetti che possiedono la "caratteristica" fino alla modalità posta al limite superiore della classe corrispondente:
  F_2 = 35 indica che sono 35 gli individui che non superano i 30 anni di età .
- Le **frequenze relative** (f_i) , esprimono la frazione di soggetti che possiedono il "carattere in esame" con modalità compresa nell'intervallo corrispondente [o pari alla modalità corrispondente, se tali modalità sono espresse in "valori singoli"]:
  f_4 = 0,1429 individua la frazione di soggetti con "età" compresa tra 34 e 38 anni.

♣ Le frequenze relative possono essere sfruttate anche nel caso di variabili qualitative, ma è *scorretto* considerarne le cumulate nel caso in cui il carattere considerato è di tipo *nominale* (ossia senza un ordine).
Infatti, se si utilizzano i dati dell'esempio seguente (relativo ad una variabile qualitativa nominale), è lecito calcolare le **frequenze relative** e le **percentuali** (di cui si riporta il significato); ma non si possono prendere in considerazione le frequenze cumulate (F_i):

Tab.2

Sesso	**y_i**	**f_i**	**p_i**
Maschio	**7**	0,470	47 ,0
Femmina	**8**	0,530	53,0
Totali	**15**	**1,000**	**100,0**

Considerando le *frequenze relative*, diremo che la frazione dei casi in cui è stata rilevata la modalità *"femmina"* del carattere ***sesso*** è pari a 0,53 [**f_2 = 0,53**]; ed, usando le percentuali, si dirà che è del 53% la percentuale di soggetti di sesso femminile, all'interno del gruppo considerato.

VALORI MEDI

> ***Valore medio di una serie ordinata di valori è un valore compreso tra il più piccolo ed il più grande dei valori osservati***.

È una definizione molto generale.
Per rispettare la natura del fenomeno in esame è preferibile fare riferimento ad una classificazione più vincolante.
Si può fare una prima distinzione tra "***valori medi algebrici***" e "***valori medi di posizione***".

1 - *I Valori Medi Algebrici*

Sono sintetizzati tramite delle formule matematiche, possono essere usati solo per variabili di tipo quantitativo e prendono in considerazione tutta la distribuzione dei valori osservata.

La Media Aritmetica

Se le modalità di un fenomeno X, sono legate da una relazione di tipo additivo

$$x_1 + x_2 + \ldots + x_i + \ldots + x_n = \sum_{i=1}^{n} x_i$$

La Media Aritmetica (M) è quel valore costante che, sostituito ad ognuno dei valori osservati, ne lascia inalterata la somma:

$$\underbrace{M + M + \ldots + M}_{n \text{ volte } M} + \ldots + M = \sum_{i=1}^{n} x_i$$

Da cui si ricava la formula finale [1] :

$$\mathbf{M = \frac{\sum_{i=1}^{n} x_i}{n}}$$

- Nel caso in cui i <u>dati osservati possono essere sintetizzati in una</u> "distribuzione di frequenza":

con modalità : $x_1 , x_2 \ldots\ldots x_i \ldots\ldots x_n$
e con frequenze : $y_1 , y_2 \ldots\ldots y_i \ldots\ldots y_n$

la media aritmetica diventa:

$$\mathbf{M} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i} \qquad [2]$$

ma, anche in questa formulazione « ponderata » , la **media aritmetica** è sempre esprimibile come il <u>rapporto tra</u> l'ammontare totale del fenomeno**,** ossia $\sum x_i y_i$, diviso il numero totale dei casi **,** $\sum y_i$ **.**

- Tra le caratteristiche della media aritmetica vi è quello di <u>minimizzare</u> le *distorsioni* o *<u>gli errori accidentali</u>* ε_i <u>contenuti nei dati</u> x_i , ottenuti come <u>misure ripetute di un carattere X.</u>

a) ***Proprietà del Baricentro: l***a ***<u>somma</u>*** degli ***<u>scarti</u>*** tra i singoli <u>valori e la propria</u> *<u>Media aritmetica</u>* è ***<u>nulla</u>*** :

$$\sum_{i=1}^{n} (x_i - M) = 0$$

e

$$\sum_{i=1}^{n} (x_i - M)\, y_i = 0$$ per distribuzioni di frequenza

b) ***Proprietà del Minimo: l***a ***<u>somma</u>*** dei ***<u>quadrati degli scarti</u>*** <u>tra i singoli valori e la propria *media aritmetica* ***è un minimo,***</u> rispetto alla somma del quadrato degli scarti dei valori da un "valore medio" qualsiasi ***:***

$$\sum_{i=1}^{n} (x_i - M)^2 = \text{minimo}$$

e

$$\sum_{i=1}^{n} (x_i - M)^2 y_i = \text{minimo}$$ per distribuzioni di frequenza

2- I Valori Medi di Posizione

I "**valori medi di posizione**" sono indici che **individuano le modalità che occupano una posizione** ben precisa all'interno della serie ordinata di modalità. Rispondono all'esigenza di ripartire la distribuzione i parti di eguale numerosità e possono essere utilizzate su dati ***qualitativi ordinati*** e su dati ***quantitativi.***

- La Mediana

Il "valore mediano" o *Mediana* , fa parte dei valori medi di posizione e pertanto può essere utilizzata in presenza di ***caratteri ordinabili***, anche se qualitativi, ma solo dopo averli ordinati in ordine crescente: essa bipartisce la distribuzioni in due parti di eguale numerosità.
Data una serie ordinata di modalità:

$$x_1 \leq x_2 \leq x_3 \leq \ldots\ldots\ldots \leq x_i \leq \ldots\ldots\ldots \leq x_n$$

> ***La Mediana è la modalità che è preceduta e seguita dallo stesso numero di osservazioni***

Per individuarla si utilizza la "posizione" occupata all'interno della serie ordinata di modalità (o valori):
- se N è dispari , il "***posto***" occupato dalla Mediana è dato da **[(N + 1) / 2]** e la mediana è l'unica modalità che occupa quel "posto";

- se N è pari , i " ***posti***" sono due : **[N / 2]** e **[(N / 2) + 1]** e le mediane sono due modalità: una preceduta da metà delle osservazioni e la seconda seguita dall'altra metà delle osservazioni [possono essere due "modalità" (o valori) coincidenti].

- Nel caso di distribuzione di frequenza:

	x_1	$\leq$	$x_2 \leq$	$x_3 \leq \ldots\ldots\ldots \leq$	x_i	$\leq \ldots\ldots\ldots \leq$	x_n
con frequenze:	y_1		y_2	y_3		y_i		y_n

valgono le stesse regole per individuare il "posto" occupato dalla Mediana , ma occorre utilizzare le "frequenze cumulate" F_i , ossia:

la F_i maggiore o uguale al *posto* occupato dalla mediana , la individua

Se $F_k \geq$ "Posto" , allora $\mathbf{M_e = x_k}$

Esempi

Es. Calcolare la media aritmetica e la mediana dei dati riportati nella seguente distribuzione di frequenza del "consumo di gasolio" sostenuto in un anno da 155 "Hotel".

Tab.3

Consumo Gasolio (in quintali)	***Numero di Hotel***
148	14
152	26
156	31
160	37
164	38
168	9
Totale	***155***

In presenza di una distribuzione di frequenza, per il calcolo della media aritmetica si utilizza la formula ponderata.

Prospetto di calcolo

Consumo Gasolio x_i	y_i	$x_i\, y_i$	F_i
148	14	2072	14
152	26	3952	40
156	31	4836	71
160	37	5920	108
164	38	6232	146
168	9	1512	155
Totale	***155***	***24524***	*//*

• Utilizzando la formula di seguito riportata ed i calcoli della terza colonna della tabella precedente, si perviene al valore finale:

$$\mathbf{M} = \sum_{i=1}^{n} \mathbf{x_i y_i} \Big/ \sum_{i=1}^{n} \mathbf{y_i}$$

M = 24524 / 155 = 158,2 quintali

Da notare che:

- la media è espressa nella stessa unità di misura del carattere;
- il valore ottenuto è compreso tra il più piccolo (148) ed il più grande (168) delle modalità riportate in tabella.

• Per il calcolo della **mediana**, essendo in presenza di una distribuzione di frequenza, si utilizzano le frequenze cumulate, riportate nell'ultima colonna della seconda tabella.

La mediana è la modalità preceduta e seguita dallo stesso numero di casi; tale modalità, nel caso in cui il "totale frequenze" (N) è dispari, occupa il posto:

$$(N + 1) / 2 = (155 + 1) / 2 = 78° \text{ posto}$$

La prima frequenza cumulata che supera questo valore è $N_4 = 108$ a cui corrisponde la modalità 160, che è la mediana della distribuzione:

$$\mathbf{M_e = 160 \text{ quintali}}$$

Da notare che, anche in questo caso:

- la mediana è espressa nella stessa unità di misura del carattere;
- anche la mediana è compresa tra il più piccolo ed il più grande dei valori osservati.

Es. Sui dati della Tab.1, calcolare la mediana della distribuzione.

• Poiché il totale delle frequenze è un numero pari, vi sono due Mediane: la 1° mediana occupa il **posto (N / 2) = (70 /2) = 35** che individua **la classe mediana [26 —| 30]** , corrispondente alla frequenza assoluta cumulata F_2 =35; mentre la 2° mediana occupa **il posto (N / 2) + 1 = 35 + 1 = 36°**, che individua la classe mediana adiacente **[30 —| 34],** corrispondente alla frequenza assoluta cumulata F_3 =47.

Per individuare **una modalità mediana all'interno della classe** si utilizza la nota proporzione:

$$[L - l] : [M_e - l] = [F_s - F_{s-1}] : [\text{Posto}^\circ - F_{s-1}]$$

- Nell'esempio considerato $M_e^{(I)}$:

L = 30 l = 26 F_s = 35 F_{s-1} = 13 Posto = 35°

$$(30 - 26) : (M_e^{(I)} - 26) = (35 - 13) : (35 - 13)$$

$$M_e^{(I)} = 26 + [(30 - 26) \times (35 - 13)] / (35 - 13)$$

$\mathbf{M_e^{(I)}}$ = 26 + 4 = **30 anni** (età massima, raggiunta dalla prima metà dei soggetti)

- Analogamente si calcola il 2° valore mediano (in questo caso non coincidente con il 1°): $M_e^{(II)}$

L = 34 l = 30 F_s = 47 F_{s-1} = 35 Posto = 36°

$$(34 - 30) : (M_e^{(I)} - 30) = (47 - 35) : (36 - 35)$$

$$M_e^{(II)} = 30 + [(34 - 30) \times (36 - 35)] / (47 - 35)$$

$\mathbf{M_e^{(II)}}$ = 30 + (4 / 12) = 33,3 $\overline{3}$ ≅ **33 anni**

Ossia metà degli individui hanno un'età inferiore a 30 anni e l'altra metà un'età superiore a 33 anni .