STATISTICA I

master COPY COPISTERIA

LEZIONE 1

Si consideri un insieme di punti dati - DATABASE

$\{x_1, x_2, \ldots, x_n\}$

dove $n \in \mathbb{N}$ è il numero di punti dato (o cardinalità del database).

Esempi

Compleanni: {04/05, 09/10, ...}

Salari (in migliaia di euro): {21, 26, 40, 29, 31, 32, 32, 32, ...}

Test PCR {-, -, -, -, +, -, -, +, -, ...}

2.2 Organizzazione e descrizione dei dati

Per leggere, interpretare e capire il contenuto e le caratteristiche globali di un database dobbiamo presentare il suo contenuto in maniera intelleggibile.

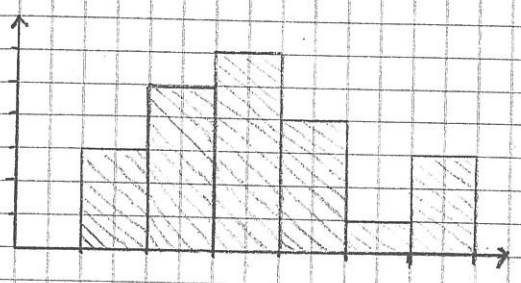
⇒ rappresentiamo le frequenze (freq. assolute) nel database, cioè il numero di volte che un valore è osservato. (efficiente se il numero di valori distinti è basso)

Queste possono essere rappresentate in diversi modi:

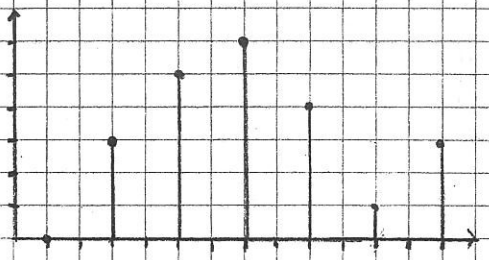
- TABELLE:

valore	v_1	v_2	v_3	v_4 ...
frequenza	f_1	f_2	f_3	f_4 ...

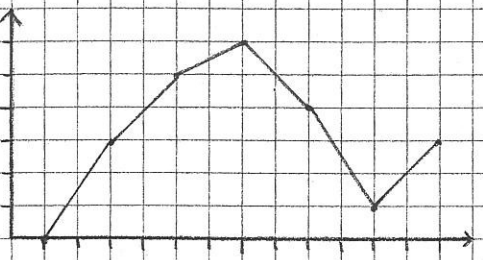
- GRAFICO A BARRE



- GRAFICO A BASTONCINI



- GRAFICO A LINEE

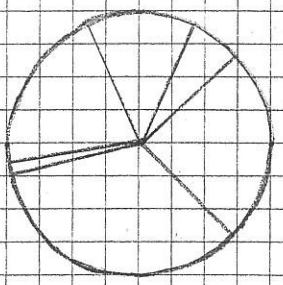


Talvolta il numero assoluto f di realizzazioni di un valore ci interessa meno della sua frequenza relativa (frazione di volte che un valore appare in un database di ampiezza n), data da f/n

Come nel caso delle frequenze assolute, le frequenze relative si possono rappresentare in:

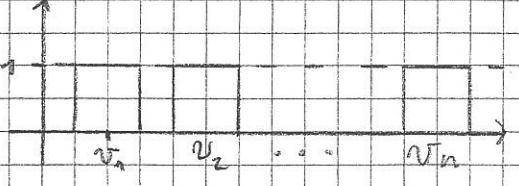
- tabelle
- grafici a barre
- grafici a bastoncini
- grafici a linee

Inoltre, un'altra rappresentazione possibile è quella tramite diagramma a torta.



I metodi esposti sopra perdono il loro senso se il numero di valori è comparabile al numero di elementi nel campione.

Controesempio



Se questo è il caso, possiamo dividere i dati in gruppi di valori contigui (CLASSI) e associare ogni valore a una di esse.

⚠ la DEFINIZIONE DELLE CLASSI gioca un ruolo fondamentale:
- se troppo poche → una colonna
- se troppe → troppe colonne "basse"

Ideale → 5-20 classi

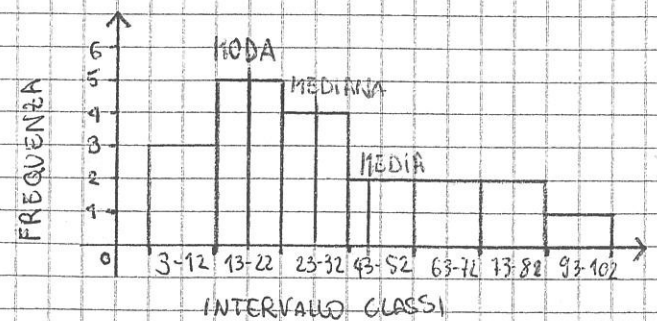
Infine, è possibile rappresentare i dati con un GRAFICO CUMULATIVO

2.3 le grandezze che sintetizzano i dati

Spesso i database a disposizione sono così grandi (milioni di punti dati, centinaia di categorie...) che risulta necessario rappresentarli/sintetizzarli ulteriormente.
In questo caso vale la pena sintetizzare i dati tramite STATISTICHE → grandezze calcolate a partire dai dati.
Un tipo importante di statistica è quella che mi identifica il valore ("tipico") del campione, attorno al quale si "accumulano" i dati.

⚠ Non esiste una definizione unica di tale statistica ma dipende dalla nostra interpretazione di "tipico" e "accumulano".

Infatti esistono (almeno) 3 tali statistiche:



Troviamo quindi, per un campione $\{x_1, \dots, x_n\}$ di n elementi

DEFINIZIONE 1

Si dice MEDIA CAMPIONARIA (denotata con $\bar{x}$)

$$\boxed{\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n}\sum_{j=1}^{n} x_j}$$

RICHIAMO SOMMATORIA

$$\sum_{j=1}^{10} x_j = x_1 + x_2 + x_3 + \dots + x_{10}$$

Esempio

$x_j = 2^j \qquad \sum_{j=0}^{5} 2^j = 1 + 2 + 3 + 4 + 5 + 0 = 63$

NOTA utile: se $a, b \in \mathbb{R}$

- $\sum_{j=1}^{10} a x_j = (a x_1 + a x_2 + \dots + a x_{10}) = a(x_1 + x_2 + \dots + x_{10}) = a\left(\sum_{j=1}^{10} x_j\right)$
- $\sum_{j=1}^{10} (x_j + b) = (x_1 + b) + (x_2 + b) + \dots + (x_{10} + b) = (x_1 + x_2 + \dots + x_{10}) + (b + b \dots b) =$

$$= \sum_{j=1}^{10} x_j + 10b$$

---

le proprietà della sommatoria possono essere usate per calcolare la media campionaria più facilmente.

Esempio

(Punteggio medio al campionato di golf)

$= \{284, 280, 277, 282, 279, 285, 281, 283, 278, 277\}$

Possiamo scrivere $x_i = y_i + 280$

$= \{4, 0, -3, 2, -1, 5, 1, 3, -2, -3\}$

e usare che: $\bar{x} \sum_{j=1}^{10} x_j = \frac{1}{10}\sum_{j=1}^{10}(y_j + 280) = \frac{1}{10}\sum_{j=1}^{10} y_i + \frac{1}{\cancel{10}} \cdot \cancel{10} \cdot 280 = \bar{y} + 280$

dove $\bar{y} = \frac{1}{10}(4+0-3+2-1+5+1+3-2-3) = \frac{6}{10}$

$\Rightarrow \bar{x} = \bar{y} + 280 = \frac{6}{10} + 280 = 280{,}6$

In generale, se $\boxed{y_j = ax_j + b}$, con $a, b \in \mathbb{R}$ allora

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n}(\underbrace{ax_j+b}_{y_j}) = \frac{1}{n}\left(\sum_{j=1}^{n} ax_j + \sum_{j=1}^{n} b\right) = \frac{1}{n}\left(a\sum_{j=1}^{n} x_j + \sum_{j=1}^{n} b\right)$$

$$= a \cdot \frac{1}{n}\cdot \underbrace{\sum_{j=1}^{n} x_j}_{\bar{x}} + \frac{1}{\cancel{n}} \cdot \cancel{n} \cdot b = a\bar{x} + b$$

RIASSUMENDO $\boxed{\bar{y} = \frac{1}{n}\sum_{j=1}^{n}(ax_j+b) = a\bar{x}+b}$

NOTA: nel caso in cui l'insieme di dati è fornito tramite le frequenze (assolute) dei suoi valori, si calcola la media campionaria pesando i valori per la loro frequenza.

Esempio

Età	15	16	17	18	19	20
freq	2	5	11	9	14	13

dove $n = f_1 + f_2 + f_3 + \dots = 54$

Allora abbiamo $\bar{x} = \frac{1}{54}(2\cdot 15 + 5\cdot 16 + 11\cdot 17 + 9\cdot 18 + 14\cdot 19 + 13\cdot 20) = 18{,}24$

In generale: per un insieme di dati con valori $\{v_1, \dots, v_h\}$ e relative frequenze $\{f_1, \dots, f_h\}$ (il valore v_j ha frequenza f_j)

- il numero complessivo di dati è $\boxed{n = f_1 + \dots + f_h = \sum_{j=1}^{h} f_j}$

- la media campionaria è

$$\boxed{\bar{x} = \frac{1}{n}\sum_{j=1}^{h} f_j \cdot v_j = \underbrace{\frac{f_1}{n}}_{\text{peso}} \cdot \underbrace{v_1}_{\text{valore}} + \dots + \frac{f_h}{n} v_h}$$ Media pesata