01/03/2022

statistica

Si considera un insieme di punti dati → DATABASE

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

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

es.
- stipendi (in migliaia di euro) $\{21, 26, 30, \dots\}$
- compleanni $\{01.05, 09.10, \dots\}$
- Test PCR $\{-, -, +, -, +, -\}$
- Instagram: network

ORGANIZZAZIONE E DESCRIZIONE DEI DATI

Per descrivere un insieme di dati bisogna sintetizzarli in due-tre numeri o una rappresentazione leggibile.

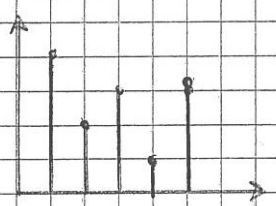
→ rappresentiamo le FREQUENZE (freq. ASSOLUTE) nel database cioè il numero di volte che il valore si ripete.

valore	v_1	v_2	v_3	…
frequenza	f_1	f_2	f_3	…

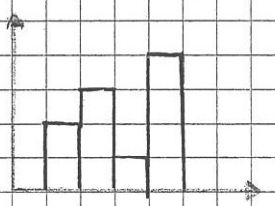
→ TABELLE DI FREQUENZA

Come è possibile rappresentare i dati:

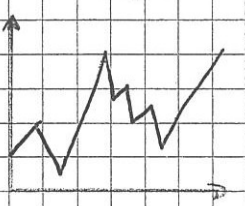
* GRAFICO A BASTONCINI



* GRAFICO A BARRE



* GRAFICO A LINEE

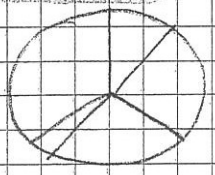


Il numero assoluto f di realizzazione di un valore ci interessa meno della sua FREQUENZA RELATIVA

↓

frazione di volte che un valore appare in un database

* DIAGRAMMA A TORTA



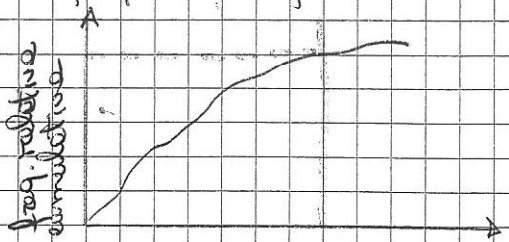
Se ogni numero ha valore diverso per ogni valore nel database non ha senso usare i metodi prima indicati.

Possiamo dividere i dati in gruppi di valori contigui (CLASSI) ed associare ogni valore ad una di esse.

⚠ la definizione delle classi gioca un ruolo fondamentale, se sono troppe poche avremo una sola colonna, se sono troppe avremo troppe colonne "basse".

Ideale è avere tra le 15 e le 20 classi

È possibile rappresentare i dati con un grafico cumulativo sommando le frequenze fino ad un certo valore



master copy COPISTERIA

GRANDEZZE CHE SINTETIZZANO I DATI

Spesso i database sono così grandi che è necessario rappresentare o sintetizzarli ulteriormente.
In questi casi vale la pena sintetizzare i dati tramite STATISTICHE → grandezze calcolate a partire dai dati.

un tipo importante è quello che mira ad identificare 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 "Tipico" e "accumulo"

Esistono TRE TIPI DI STATISTICHE: • MEDIA
• MEDIANA
• MODA

DEFINIZIONE 1: Si dice MEDIA CAMPIONARIA ($\bar{x}$)

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

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

25 punteggio medio in un torneo di golf.

$\{\underbrace{284}_{x_1}, 280, 277, 288, 279, 285, 281, 278, 277\}$

possiamo scrivere $x_i = y_i + 280$

$\{4, 0, -3, 8, -1, 5, 1, -2, -3\}$

e notare che $\bar{x} = \frac{1}{9}\sum_{j=1}^{9} x_j = \frac{1}{9}\sum_{j=1}^{9}(x_j + 280) = \frac{1}{9}\sum_{j=1}^{9} y_j + \frac{1}{9}\cdot 9 \cdot 280 =$

$= \bar{y} + 280$

dove $\bar{y} = \frac{1}{9}(4 + 0 - 3 + 8 - 1 + 5 + 1 - 2 - 3) = \frac{9}{9} = 1$

$\Rightarrow \bar{x} = \bar{y} + 280 = 1 + 280 = 281$

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

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n}(ax_j + b) = \frac{1}{n}\left(\sum ax_j + \sum b\right) = \frac{1}{n}\left(a\sum x_j + \sum b\right) = a\frac{1}{n}\sum x_j + \frac{1}{n}\cdot nb =$$

$$= a\bar{x} + b$$

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

es.

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

$n = f_1 + f_2 + f_3 + \dots + f_n$

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

$= 1824$

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

- il numero complessivo di dati è $n = f_1 + \dots + f_h = \sum_{j=1}^{h} f_j$
- la media campionaria è $\bar{x} = \frac{1}{n}\sum_{j=1}^{h} f_j \cdot v_j = \frac{f_1}{n}\cdot v_1 + \dots + \frac{f_h}{n} v_h$

($\frac{f_1}{n}$: peso; v_1: valore; media ponderata)

DEFINIZIONE 2: Assegnato un insieme di dati di ampiezza n e li si ordini dal minore al maggiore, la MEDIANA campionaria è data da:

- il valore del dato in posizione $\frac{n+1}{2}$ se n è DISPARI
- la media aritmetica tra i valori dei dati in posizione $\frac{n}{2}$ e $\frac{n}{2}+1$ se n è PARI

es.

1)

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

(27 e 28 sopra 18 e 19)

$n = 54$ (pari)

$\Rightarrow$ consideriamo la media tra $\frac{54}{2} = 27$ e $\frac{54}{2}+1 = 28$ numero più piccolo cioè mediana $= \frac{18+19}{2} = 18{,}5$

27 corrisponde a 18 perché $2+5+11+9 = 27$ (somme delle frequenze)

2)

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 16, 28, 32

1	58, 92, 93, 96, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

Troviamo che le medie sono: $\begin{cases} \bar{x}_{\text{sterile}} = 344{,}07 \\ \bar{x}_{\text{normale}} = 292{,}32 \end{cases}$

Per le mediane invece abbiamo $\begin{cases} n_{\text{sterile}} = 29 \to \text{dispari } j = 15 \\ n_{\text{normale}} = 19 \to \text{dispari } j = 10 \end{cases}$

$\begin{cases} \text{mediana sterile} = 259 \\ \text{mediana normale} = 265 \end{cases}$

⚠ la media campionaria sono molto diverse a causa di pochi valori grandi nel primo campione (outliers)! La mediana non percepisce questi dati

DEFINIZIONE 3: La MODA CAMPIONARIA di un insieme di dati (se esiste) è l'unico valore che ha frequenza massima. Se non esiste un unico valore con frequenza massima ci riferiamo di essi e detto VALORE MODALE.

es.

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

↳ MODA

03/03/2022

VARIANZA & VARIAZIONE STANDARD

Invece di cercare il "centro" o valore "tipico" di un campione, a volte vogliamo sintetizzare quanto il campione sia "disperso attorno a tale valore, ovvero vogliamo quantificare la "distanza tipica" dei punti dati dal loro centro!

IDEA 1) media delle differenze $\frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})$?

↓

SBAGLIATO ∀ perché ci sono valori positivi e negativi:

$$\frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x}) = \frac{1}{n}\sum_{j=1}^{n}x_j - \frac{1}{n}\sum_{j=1}^{n}\bar{x} = \underbrace{\frac{1}{n}\sum_{j=1}^{n}x_j}_{\bar{x}} - \frac{1}{n}\cdot n\bar{x} = 0$$

DEFINIZIONE: Assegnato un insieme di dati $\{x_1,\dots,x_n\}$ si dice VARIANZA CAMPIONARIA la quantità

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2$$

(S al quadrato) con $\bar{x}$ = media campionaria

es. Consideriamo i campioni $A=\{3,4,6,7,10\}$, $B=\{-20,5,15,24\}$

$$\bar{x}_A = \frac{1}{5}\cdot(3+4+6+7+10) = 6 \qquad \bar{x}_B = \frac{1}{4}\cdot(-20+5+15+24) = 6$$

$$S_B^2 = \frac{1}{4-1}\cdot\left[(-20-6)^2+(5-6)^2+(15-6)^2+(24-6)^2\right] = 360{,}67$$

$$S_A^2 = \frac{1}{5-1}\cdot\left[(3-6)^2+(4-6)^2+(6-6)^2+(7-6)^2+(10-6)^2\right] = 7{,}5$$

proposizione: Per un insieme $\{x_1,x_2,\dots,x_n\}$ con media $\bar{x}$ abbiamo

$$\sum_{j=1}^{n}(x_j-\bar{x})^2 = \left(\sum_{j=1}^{n}x_j^2\right) - n\bar{x}^2$$

Dimostrazione: $\sum_{j=1}^{m}(x_j-\bar{x})^2 = \sum_{j=1}^{m}(x_j^2 - 2x_j\bar{x} + \bar{x}^2) =$

$= \sum_{j=1}^{m} x_j^2 + \sum_{j=1}^{m}(-2x_j\bar{x}) + \sum_{j=1}^{m}\bar{x}^2 =$

$= \sum_{j=1}^{m} x_j^2 - 2\bar{x}\sum_{j=1}^{m} x_j + \sum_{j=1}^{m}\bar{x}^2 =$

per definizione $= \sum_{j=1}^{m} x_j^2 - 2m\bar{x}\cdot\frac{1}{m}\sum_{j=1}^{m} x_j + \sum_{j=1}^{m}\bar{x}^2 =$ ($\frac{1}{m}\sum_{j=1}^{m} x_j = \bar{x}$)

$= \sum_{j=1}^{m} x_j^2 - 2m\bar{x}\cdot\bar{x} + m\bar{x}^2 =$

$= \sum_{j=1}^{m} x_j^2 - m\bar{x}^2$

Infine come nel caso della media, l'espressione per la varianza campionaria di dati trasformati con:

$y_i = ax_i + b \qquad a, b \in \mathbb{R}$

prende una forma semplice:

$$s_y^2 = \frac{1}{m-1}\sum_{j=1}^{m} a^2(x_j-\bar{x})^2 = a^2 s_x^2$$

DEFINIZIONE: lo ~~VARIANZA~~ STANDARD CAMPIONARIA di un insieme di dati $\{x_1, \dots, x_m\}$ è:

$$s := \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(x_j-\bar{x})^2}$$

PERCENTILI CAMPIONARI e BOX PLOT

È utile identificare la parte del campione che ha "voti" nel "migliore $x^{\%}$".

In statistica identifichiamo il valore che separa il $k\%$ più basso (il "peggior" $k\%$) dei dati dal $(100-k)\%$ più alto (il "migliore" $100k\%$) come il K-PERCENTILE.

DEFINIZIONE: sia K un numero intero con $0 \le K \le 100$.
Il K-PERCENTILE CAMPIONARIO è il valore del punto dato x_j tale che almeno:
$\{$ il $k\%$ del database è $\le x_j$ (il $100k\%$ del database) $\ge x_j$

Se due punti dati soddisfano questa condizione allora il K-percentile è la loro MEDIA ARITMETICA

es. il 93-percentile è 98,3

In pratica per un campione di ampiezza m, il K-percentile si ottiene identificando i valori (al massimo 2) che soddisfano:
1) almeno $m \cdot \frac{k}{100}$ è $\le$ ad essi
2) almeno $\frac{100-k}{100}\,m$ valori è $\ge$ ad essi.

Es. $\{x_1, x_2, x_3\}$ e $\{3, 5, 9\}$ (m) Calcoliamo il 50-percentile per questo numero abbiamo
$\{$ almeno $3 \cdot \frac{50}{100} = 1,5 \to 2$ numero $\le$ ad esso
$\{$ almeno $3 \cdot \frac{100-50}{100} = 1,5 \to 2$ numero $\ge$ ad esso