Lezione 12/10/2022

La statistica descrittiva

La statistica descrittiva studia i dati di una popolazione.

Una **popolazione** è un insime di entità.
Esempi: gli studenti di un corso di laurea, le automobili di una città, gli iscritti di una palestra, ...

Un **carattere** è una proprietà che si vuole studiare della popolazione.
Esempi: se la popolazione è un insieme di studenti allora dei caratteri possono essere il colore degli occhi, l'altezza, il voto di un esame, il colore preferoto, ...

Le **modalità** di un carattere sono tutti i possibili modi in cui può essere un carattere.
Ad esempio, le modalità del carattere *voto* sono tutti i numeri da 1 a 30; le modalità del carattere *colore degli occhi* sono *Verde, Marrone, Nero e Azzurro.*

I caratteri si distinguono in **qualitativi** e **quantitativi**. I caratteri qualitativi hanno le modalità espresse da parole. Ad esempio, il carattere *colore degli occhi* è qualitativo. I caratteri quantitativi, invece, hanno dei numeri come modalità. Ad esempio, il carattere *voto* è quantitativo.

I caratteri quantitativi possono essere **discreti** se le modalità sono numeri interi (il voto, l'anno di nascita) o **continui** se le modalità sono numeri reali (il peso, l'altezza e la temperatura).

Gli **indici statistici** forniscono informazioni di sintesi dei dati. Siamo interessati agli **indici di centralità** (media, mediana e moda) e agli **indici di variabilità** (varianza e deviazione standard). Gli indici di centralità indicano il centro dei dati, mentre gli indici di variabilità misurano quanto i dati sono lontani (o vicini) gli uni dagli altri.

Istogramma: classi di diversa ampiezza

Classi di età	Numero di impiegati	Ampiezza della classe	Densità di frequenza della classe
[20, 30)	17	10	$\frac{17}{10} = 1{,}7$
[30, 50)	24	20	$\frac{24}{20} = 1{,}2$
[50, 60)	22	10	$\frac{22}{10} = 2{,}2$
[60, 65)	6	5	$\frac{6}{5} = 1{,}2$

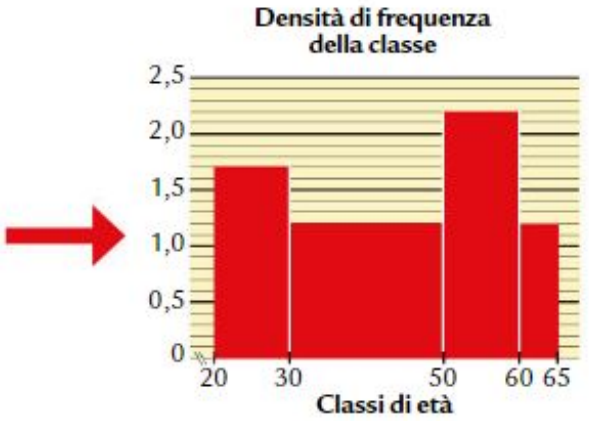

Figura 2

Media aritmetica

La media aritmetica (o media semplice) dei valori numerici $x_1, \ldots, x_n$ è il numero indicato con $\bar{x}$ (oppure con μ) e definito dalla seguente formula

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(la media aritmetica si usa solo per modalità relative a un carattere quantitativo)

Esempi (media aritmetica)

ESEMPI Media aritmetica semplice

a. Siano 2000 \$, 1600 \$, 1400 \$ e 1800 \$ i costi di quattro viaggi negli Stati Uniti. Il costo medio dei quattro viaggi è:

$$\frac{2000 + 1600 + 1400 + 1800}{4}\ \$ = 1700\ \$$$

Osserva che in questo caso il valore della media ottenuto potrebbe essere esso stesso un valore del carattere «costo del viaggio».

b. Sei famiglie hanno, rispettivamente, il seguente numero di figli: 0, 2, 1, 1, 3, 2. Il numero medio di figli per famiglia è:

$$\frac{0+2+1+1+3+2}{6} = \frac{9}{6} = 1{,}5$$

In questo caso la media aritmetica, evidentemente, **non** può rappresentare il numero dei figli di una famiglia.

Proprietà della Media aritmetica

Teorema

Siano $x_1, \ldots, x_n$ dati numerici e $a, b \in \mathrm{R}$. Se poniamo $y_i = ax_i + b$, allora la media $\bar{y} = a\bar{x} + b$.

Esempio

La media dei prezzi degli articoli venduti in un negozio è di 35 euro.

1. Se il negoziante decidesse di aumentare i prezzi di tutti gli articoli di 2 euro, quale diverrebbe la media dei prezzi degli articoli del negozio?
   $x_1 + 2, \ldots, x_n + 2$ implica che $\bar{y} = \bar{x} + 2 = 37$ euro.
2. E se invece il negoziante aumentasse il prezzo di tutti gli articoli del 10%?
   $y_i = x_i + \frac{10}{100}x_i = (1 + \frac{10}{100})x_i = 1,1 * x_i$, allora
   $\bar{y} = 1.1 * \bar{x} = 1,1 * 35 = 38,5$ euro

Mediana

Dati n valori numerici, ordinati in senso crescente o decrescente, si definisce loro mediana

- il numero che occupa la posizione centrale, se n è dispari;
- la media aritmetica dei due numeri che occupano le posizioni centrali, se n è pari

Sequenza di numeri	2, 6, 1, 11, 9, 8, 13 (n dispari)	3, 3, 4, 6, 7, 10 (n pari)
Ordinamento	I numeri **non** sono ordinati in senso crescente; per prima cosa vanno *riordinati*: 1, 2, 6, 8, 9, 11, 13	I numeri sono già ordinati in senso crescente.
Calcolo della mediana	1, 2, 6, **8**, 9, 11, 13 Posizione centrale 8 ← mediana	3, 3, **4**, **6**, 7, 10 Media delle posizioni centrali $\frac{4+6}{2} = 5$ ← mediana

- se n è dispari, il numero che occupa la posizione centrale (ossia la mediana) è quello di posto $\dfrac{n+1}{2}$
- se n è pari, i due numeri che occupano le posizioni centrali sono quelli di posto $\frac{n}{2}$ e $\frac{n}{2}+1$ e la mediana è la loro media aritmetica

Moda

In un'indagine statistica, si chiama moda una modalità che si presenta con la massima frequenza (assoluta, relativa o percentuale).

Esempi

- Supponiamo che i dati grezzi ottenuti in seguito a una rilevazione statistica siano i seguenti numeri: **2, 3, 4, 5, 5, 6, 6, 6**. Poiché la modalità che si presenta con maggiore frequenza è 6, questa è la moda dei dati raccolti.
- Supponiamo che i dati grezzi ottenuti in seguito a una rilevazione statistica siano i seguenti numeri: **2, 2, 3, 3, 4, 5**. Ci sono due modalità (2 e 3) che presentano lamassima frequenza, dunque i dati raccolti presentano due mode: 2 e 3.

Valori medi nel caso di una distribuzione di frequenze

Un'indagine effettuata su un campione di famiglie ha prodotto la distribuzione di frequenze rappresentata nella seguente tabella.
Determiniamo la media.

Numero di figli per famiglia	0	1	2	3	4	5
Frequenza	9	27	40	20	3	1

Il numero medio di figli per famiglia è dato dalla formula:

$$\bar{x} = \frac{0 * 9 + 1 * 27 + 2 * 40 + 3 * 20 + 4 * 3 + 5 * 1}{9 + 27 + 40 + 20 + 3 + 1} = 1,84$$

Valori medi nel caso di una distribuzione di frequenze

$$\bar{x} = \frac{0 * 9 + 1 * 27 + 2 * 40 + 3 * 20 + 4 * 3 + 5 * 1}{9 + 27 + 40 + 20 + 3 + 1}$$

$\bar{x}$ si può riscrivere nella seguente formula

$$\bar{x} = 0 * \frac{9}{100} + 1 * \frac{27}{100} + 2 * \frac{40}{100} + 3 * \frac{20}{100} + 4 * \frac{3}{100} + 5 * \frac{1}{100}$$

Osserva che i valori in blu sono le frequenze relative

Nel caso in cui sono rilevate $x_1, \ldots, x_k$ differenti modalità di un carattere, con frequenze assolute rispettivamente uguali a $f_1, \ldots, f_k$ e frequenze relative uguali a $f_{r_1}, \ldots, f_{r_k}$, per il calcolo della media si possono utilizzare le due formule

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_k f_k}{f_1 + f_2 + \ldots + f_k}$$

utilizzando le frequenze assolute

oppure

$$\bar{x} = x_1 f_{r_1} + x_2 f_{r_2} + \ldots + x_k f_{r_k}$$

utilizzando le frequenze relative

- Abbiamo introdotto tre valori medi: la *media aritmetica*, la *mediana* e la *moda*. In generale è buona pratica calcolare tutti e tre questi valori: infatti essi forniscono informazioni *complementari*, che descrivono aspetti differenti. Per esempio, in riferimento alla retribuzione annua netta dei dipendenti di una data azienda:
  - sapere che la *media* dei salari è 22 500 euro significa che, se il denaro complessivamente speso per gli stipendi venisse distribuito in modo che il salario sia uguale per tutti, allora ciascuno riceverebbe 22 500 euro all'anno;
  - sapere che la *mediana* dei salari è 20 500 euro, significa che circa la metà dei dipendenti percepisce uno stipendio superiore (o uguale) a 20 500 euro, e circa la metà uno stipendio inferiore (o uguale) a 20 500 euro;
  - sapere che la *moda* dei salari è 18 000 euro, significa che questo è il salario più frequente, ossia più comune.
- Sebbene la media aritmetica sia certamente il valore medio più noto e utilizzato, a seconda del particolare fenomeno in esame la *mediana* o la *moda* possono talvolta rivelarsi valori medi più idonei. Per esempio, se calcoliamo la *media aritmetica* dei numeri:

  3, 4, 4, 6, 7, 8, 9, 100

  troviamo come risultato $\overline{x} \simeq 17,6$ e possiamo notare che ben sette degli otto numeri sono più piccoli della media aritmetica. In questo caso la media aritmetica è *poco* rappresentativa dei dati, perché è eccessivamente influenzata dal valore anomalo 100; è più rappresentativa dei dati la *mediana*, che vale 6,5. Una situazione analoga si verifica, per esempio, nelle rilevazioni dei redditi o dei consumi, in cui i dati possono presentare valori «anomali» molto grandi o molto piccoli: in tali casi la *mediana* tende di solito a fornire un valore più rappresentativo della media aritmetica (troppo sensibile ai valori «anomali»). Se invece consideriamo, per esempio, il caso di un negoziante che deve scegliere la taglia di pantaloni di cui ordinare il maggior numero di capi, allora è chiaro che il valore di sintesi più rappresentativo risulta la *moda*: il negoziante sceglierà la taglia più comune, cioè quella acquistata più di frequente.

Indici di variabilità

Esempio Consideriamo, a questo proposito, un semplice esempio. In un gruppo, A, di dieci individui le retribuzioni nette annue pro capite sono (in euro).

- Gruppo A:
  2000; 2000; 2000; 3000; 4000; 4000; 5000; 5000; 5000; 168000
- Gruppo B:
  18000; 18000; 19000; 19000; 20000; 20000; 21000; 21000; 22000; 2200

La media è 20000 in entrambi i casi, ma i due gruppi di dati sono diversi tra loro. La media non può cogliere la variabilità del fenomeno.

Indici di variabilità

Una prima idea per cercare di costruire una misura più raffinata della variabilità, che tenga conto di tutte le modalità osservate, è la seguente: invece di cercare di confrontare fra loro le varie modalità (come nel caso del *campo di variazione*) si confronta ciascuna delle modalità osservate con la loro media.
Siano $x_1, x_2, \ldots, x_n$ le n modalità osservate di un carattere quantitativo; indichiamo come al solito con $\bar{x}$ la loro media aritmetica. Calcoliamo i cosiddetti **scarti** dalla media, cioè le differenze:

$$x_1 - \bar{x}, \qquad x_2 - \bar{x}, \qquad \ldots, \qquad x_n - \bar{x}$$

A questo punto, per ottenere un unico numero che esprima una misura della *variabilità* dei dati osservati, si potrebbe essere tentati di calcolare la media aritmetica degli scarti. Ma così facendo si otterrebbe sempre 0! Risulta infatti:

$$\frac{\overbrace{(x_1 - \bar{x}) + \ldots + (x_n - \bar{x})}^{n \text{ addendi}}}{n} = \frac{x_1 + \ldots + x_n - n \cdot \bar{x}}{n} = \frac{x_1 + \ldots + x_n}{n} - \bar{x} = \bar{x} - \bar{x} = 0$$

Il motivo di ciò sta nel fatto che gli scarti negativi si compensano esattamente con quelli positivi. Occorre dunque introdurre qualche modifica per eliminare l'influenza del *segno* degli scarti. Una possibilità consiste nel rimpiazzare gli scarti con i loro *valori assoluti*; un'altra nel considerare i *quadrati* degli scarti. Poiché i quadrati sono più semplici da trattare dal punto di vista matematico, si preferisce ricorrere alla seconda possibilità. In definitiva, si assume come misura della variabilità dei dati la media aritmetica dei quadrati degli scarti, cui si dà un nome particolare.

Varianza

Dati n valori numerici $x_1, \ldots, x_n$, di media aritmetica $\bar{x}$, si chiama loro varianza, e si indica con il simbolo σ^2, la media aritmetica dei quadrati degli scarti

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$$