Concetti chiave e regole

Che cos'è la statistica

La statistica ha come scopo lo studio di **fenomeni collettivi** ed esegue le sue indagini su **popolazioni** di individui al fine di raccogliere e analizzare dati relativi a uno o più **caratteri** del fenomeno stesso; i vari modi con cui si può presentare un carattere si dicono **modalità** del carattere.
Di un carattere, oggetto di una indagine statistica, si dice che è:

- una **mutabile statistica** se le sue modalità sono di tipo qualitativo, cioè si possono esprimere mediante nomi o aggettivi (colore dei capelli, tipo di laurea);
- una **variabile statistica** se le sue modalità sono di tipo quantitativo, cioè si possono esprimere mediante numeri o intervalli di numeri (altezza, peso, numero di abitanti).

La rappresentazione dei dati

I dati di un'indagine statistica possono essere raccolti in una **distribuzione di frequenze** (assolute o relative) nelle quali ogni modalità x_i del carattere è associato a un numero f_i, la sua frequenza assoluta, che indica quante volte quel carattere compare.

Ad ogni modalità del carattere può anche essere associata la sua frequenza relativa p_i, intesa come rapporto fra la frequenza assoluta e il totale T delle osservazioni.

I dati relativi a modalità x_i e frequenze f_i oppure p_i sono di solito riassunti in una **tabella delle frequenze**.

Una distribuzione di frequenze può essere rappresentata graficamente in diversi modi:

- mediante un diagramma a rettangoli o circolare
- mediante un diagramma cartesiano o un istogramma
- mediante un ideogramma o un cartogramma.

Gli indici di posizione

Una distribuzione di frequenze può essere sintetizzata da alcuni indici di posizione:

- le **medie ferme**, fra cui:
  - la media aritmetica $M = \dfrac{\sum_{i=1}^{n} x_i}{n}$ oppure nel caso di medie ponderate $M = \dfrac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$
  - la media geometrica $M_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

    oppure nel caso di medie ponderate $M_G = \sqrt[F]{(x_1)^{f_1} \cdot (x_2)^{f_2} \cdot \ldots \cdot (x_n)^{f_n}}$ dove $F = \sum_{i=1}^{n} f_i$
  - la media quadratica $M_Q = \sqrt{\dfrac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$

    oppure nel caso di medie ponderate $M_Q = \sqrt{\dfrac{x_1^2 \cdot f_1 + x_2^2 \cdot f_2 + \ldots + x_n^2 \cdot f_n}{f_1 + f_2 + \ldots + f_n}}$
  - la media armonica $M_A = \dfrac{n}{\dfrac{1}{x_1} + \dfrac{1}{x_2} + \ldots + \dfrac{1}{x_n}}$

    oppure nel caso di medie ponderate $M_A = \dfrac{f_1 + f_2 + \ldots + f_n}{\dfrac{f_1}{x_1} + \dfrac{f_2}{x_2} + \ldots + \dfrac{f_n}{x_n}}$

  Le medie ferme si possono calcolare solo per caratteri di tipo quantitativo.

- le **medie lasche**, fra cui:
  – la moda, cioè il termine a cui corrisponde la massima frequenza
  – la mediana, cioè il termine che, disposti i dati in ordine crescente o decrescente, occupa il posto centrale.

  Le medie lasche, quando esistono, si possono calcolare per qualsiasi tipo di carattere.

Le proprietà della media e della mediana

La media aritmetica e la mediana di una distribuzione godono di alcune proprietà:

- la somma degli scarti dalla media aritmetica è sempre nulla, mentre la somma dei quadrati degli scarti è minima
- la somma dei valori assoluti degli scarti dalla mediana è minima.

Le misure di dispersione

Per avere informazioni su come i dati di una indagine statistica si distribuiscono attorno ai loro valori di sintesi e quindi per poter confrontare distribuzioni, si studiano gli indici di variabilità; i più importanti fra essi sono:

- lo **scarto quadratico medio** o **deviazione standard** σ, che è la media quadratica degli scarti dalla media aritmetica:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n}}$$ nel caso di dati semplici

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left\{(x_i - M)^2 \cdot f_i\right\}}{\sum_{i=1}^{n} f_i}}$$ nel caso di dati ponderati con pesi f_i

- lo **scostamento medio** ***S***, che è la media aritmetica dei valori assoluti degli scarti dalla mediana:

$$S = \frac{\sum_{i=1}^{n}|x_i - M_e|}{n}$$ nel caso di dati semplici

$$S = \frac{\sum_{i=1}^{n}\left\{|x_i - M_e| \cdot f_i\right\}}{\sum_{i=1}^{n} f_i}$$ nel caso di dati ponderati con pesi f_i

- la **varianza** σ^2, che è il quadrato della deviazione standard

  Per il calcolo della deviazione standard (e quindi della varianza) si può anche usare la formula

$$\sigma = \sqrt{\text{media dei quadrati delle osservazioni} - \text{quadrato della media}}$$

I coefficienti di variazione

Quando si devono confrontare distribuzioni molto diverse fra loro, con caratteri che hanno differenti unità di misura o ordini di grandezza dei dati molto diversi, si ricorre ai **coefficienti di variazione** che hanno il vantaggio di essere numeri puri, quindi sempre confrontabili; i coefficienti di variazione sono definiti come rapporto fra l'indice di variabilità della distribuzione ed il valore medio utilizzato per calcolarlo:

- $CV = \dfrac{\sigma}{M}$ se il polo è la media aritmetica
- $CV = \dfrac{S}{M_e}$ se il polo è la mediana