

Popolazioni, individui e caratteri.

La statistica descrittiva si occupa dell'analisi di dati raccolti da una popolazione, ovvero da un insieme di individui. In sintesi, dato un insieme molto grande di dati, così grande che non è *utile* guardarlo dato per dati, si cerca di estrarne delle **informazioni sintetiche e tuttavia significative**.

- gli **individui** oggetto dell'indagine: ciascun individuo è un oggetto singolo dell'indagine.
- la **popolazione**, ovvero l'insieme degli individui oggetto dell'indagine.
- il **carattere** osservato o variabile, che è la quantità misurata o la qualità rilevata su ciascun individuo della popolazione.

Classificazione dei caratteri

- **caratteri numerici o quantitativi** come l'altezza, il reddito familiare, il numero dei componenti del nucleo familiare;
- **caratteri qualitativi** come il colore degli occhi.

I caratteri numerici a loro volta si possono suddividere in

- **caratteri numerici discreti** che possono assumere solo un insieme discreto di valori, come il numero dei componenti dei nuclei familiari;
- **caratteri numerici continui** che variano con continuità ovvero con una estrema accuratezza, eccessiva rispetto ai fini dell'indagine, come l'altezza delle persone o il reddito annuo familiare.

Campione statistico, modalità e classi modali

Osservo un certo carattere su una popolazione di n individui. Abbiamo un *vettore delle osservazioni*

$$x = (x_1, \dots, x_n)$$

che chiamiamo **campione statistico** di cardinalità n .

- Se il campione è relativo ad un carattere qualitativo o numerico discreto, chiamo **modalità** i valori che esso assume su un campione.
- Se il campione è relativo ad un carattere numerico continuo: sia $[a, b)$ un intervallo che contiene tutti i valori $x_i, i = 1, \dots, n$. Suddivido l'intervallo $[a, b)$ in N parti uguali (N sarà suggerito dall'esperienza). Otteniamo N intervalli

$$I_j := \left[a + (j-1) \frac{b-a}{N}, a + j \frac{b-a}{N} \right), \quad j = 1, \dots, N.$$

Ciascuno di questi intervalli, se contiene almeno una osservazione, è detto **classe di modalità**.

Frequenza assoluta e frequenza relativa 1/2

Consideriamo un campione $x = (x_1, \dots, x_n)$ relativo ad un carattere qualitativo o numerico discreto.

Il carattere osservato assume un certo numero di valori distinti

$$z_1, \dots, z_k, \quad 1 \leq k \leq n.$$

Per ogni $j = 1, \dots, k$ chiamo **effettivo** o **frequenza assoluta** della modalità z_j il numero

$$n_j := \# \{i \in \{1, \dots, n\} : x_i = z_j\}$$

mentre chiamo **frequenza relativa** della modalità z_j il numero

$$p_j := \frac{n_j}{n}.$$

Frequenza assoluta e frequenza relativa 2/2

Se il carattere osservato è numerico continuo, si considera ciascuna classe di modalità individuata

$$I_j := \left[a + (j-1) \frac{b-a}{N}, a + j \frac{b-a}{N} \right), \quad j = 1, \dots, N$$

e si chiama **frequenza assoluta o effettivo** della classe di modalità I_j il numero

$$n_j := \# \{i \in \{1, \dots, n\} : x_i \in I_j\}.$$

Come prima definiamo **frequenza relativa** della classe I_j il numero

$$p_j := \frac{n_j}{n}.$$

Moda e valori modalì

- $x = (x_1, \dots, x_n)$ campione statistico
 - z_1, z_2, \dots, z_k le modalità assunte (o l_1, \dots, l_k le classi di modalità assunte)
 - p_1, \dots, p_k le relative frequenze relative.
- 1 Se $\exists! \bar{j} \in \{1, 2, \dots, k\}$ tale che la modalità $z_{\bar{j}}$ (o la classe $l_{\bar{j}}$) ha frequenza massima, ovvero se esiste un unico $\bar{j} \in \{1, 2, \dots, k\}$ tale che $p_{\bar{j}} \geq p_j \forall j = 1, \dots, k$, allora la modalità $z_{\bar{j}}$ (o la classe $l_{\bar{j}}$) si dice **moda** del campione x .
 - 2 Se esistono due o più indici $\bar{j}_1, \bar{j}_2, \dots, \bar{j}_s$ tali che le modalità $z_{\bar{j}_1}, z_{\bar{j}_2}, \dots, z_{\bar{j}_s}$ (o le classi $l_{\bar{j}_1}, l_{\bar{j}_2}, \dots, l_{\bar{j}_s}$) hanno frequenza massima, allora queste modalità (o classi) si dicono **valori (o classi) modalì**.

Mediana

$x = (x_1, \dots, x_n)$ un campione relativo ad un carattere numerico.

Ordiniamo i dati del campione in ordine crescente:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$$

- n dispari: $n = 2m + 1$

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(m)} \leq x_{(m+1)} \leq x_{(m+2)} \leq \dots \leq x_{(2m)} \leq x_{(2m+1)}$$

$x_{(m+1)}$ è detto **mediana** del campione.

- n pari: $n = 2m$

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(m-1)} \leq x_{(m)} \leq x_{(m+1)} \leq x_{(m+2)} \leq \dots \leq x_{(2m)}$$

La media algebrica $\frac{x_{(m)} + x_{(m+1)}}{2}$ è detta **mediana** del campione

Media

$x = (x_1, \dots, x_n)$ campione relativo ad un carattere numerico

Media (aritmetica)

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i.$$

Supponiamo che nel campione siano presenti k modalità z_1, z_2, \dots, z_k con rispettive frequenze assolute N_1, N_2, \dots, N_k e frequenze relative p_1, p_2, \dots, p_k . Allora

$$\begin{aligned} \bar{x} &= \frac{1}{n} (x_1 + x_2 + \dots + x_n) = \frac{1}{n} (N_1 z_1 + N_2 z_2 + \dots + N_k z_k) = \\ &= p_1 z_1 + p_2 z_2 + \dots + p_k z_k = \sum_{j=1}^k p_j z_j. \end{aligned}$$

Varianza campionaria

Varianza campionaria

$$s_x^2 = \text{Var}[x] := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

È un *indice di dispersione*: la varianza è nulla se e solo se tutti i dati del campione sono uguali (e dunque coincidono con la media). Una varianza bassa indica che comunque i dati sono *vicini* al valore medio \bar{x} mentre una varianza alta indica una maggiore dispersione dei dati.

scarto quadratico medio (deviazione standard)

$$s_x = \text{Std}[x] := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Varianza campionaria 2

Anche per la varianza campionaria possiamo scrivere una formula che coinvolga solo le modalità e le rispettive frequenze.

$$\begin{aligned} s_x^2 &= \frac{1}{n-1} ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2) = \\ &= \frac{1}{n-1} (N_1(z_1 - \bar{x})^2 + N_2(z_2 - \bar{x})^2 + \dots + N_k(z_k - \bar{x})^2) = \\ &= \frac{n}{n-1} (p_1(z_1 - \bar{x})^2 + p_2(z_2 - \bar{x})^2 + \dots + p_k(z_k - \bar{x})^2) = \\ &= \frac{n}{n-1} \sum_{j=1}^k p_j(z_j - \bar{x})^2. \end{aligned}$$

Covarianza

Considero un **campione bivariato** cioè: rilevazione di due caratteri sugli individui di una medesima popolazione.

$$x = (x_1, \dots, x_n), \quad y = (y_1, \dots, y_n).$$

x_i e y_i sono le rilevazioni dei due caratteri sul medesimo individuo, l'individuo cioè che abbiamo etichettato come *individuo i*.

Covarianza di x e y

$$\text{Cov}[x, y] := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

dove \bar{x} e \bar{y} sono le medie dei campioni x e y , rispettivamente.

Coefficiente di correlazione 1.2

Nel caso in cui né x né y siano campioni costanti (ipotesi lavorativa che sarà sempre sottintesa), definiamo

coefficiente di correlazione di x e y

$$\rho [x, y] := \frac{\text{Cov} [x, y]}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right) \left(\sum_{i=1}^n (y_i - \bar{y})^2\right)}}.$$

Coefficiente di correlazione 2/2

Remark

$$\text{Cov}[x, x] = \text{Var}[x]; \rho[x, x] = 1.$$

- ① $-1 \leq \rho[x, y] \leq 1$;
- ② $\rho[x, y] = 1 \iff \exists a > 0, b \in \mathbb{R}$ tale che $y_i = ax_i + b \quad \forall i = 1, \dots, n$.
I campioni x e y si dicono *positivamente correlati*;
- ③ $\rho[x, y] = -1 \iff \exists a < 0, b \in \mathbb{R}$ tale che $y_i = ax_i + b \quad \forall i = 1, \dots, n$.
I campioni x e y si dicono *negativamente correlati*.
- ④ Se $\rho[x, y] = 0$ i campioni x e y si dicono *scorrelati*.

$y = ax + b$ equazione della retta cercata.

Per $x = x_i$ si ottiene il punto sulla retta $(x_i, ax_i + b)$.

Minimi quadrati

Cerchiamo la retta (ovvero i parametri a e b) che minimizza la somma degli errori quadratici nella direzione y

$$S(a, b) := \sum_{i=1}^n (y_i - (ax_i + b))^2 \rightarrow \min$$

$$\begin{aligned}
 S(a, b) &= \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - (ax_i - a\bar{x} + a\bar{x} + b))^2 = \\
 &= \sum_{i=1}^n ((y_i - \bar{y}) - a(x_i - \bar{x}) + (\bar{y} - a\bar{x} - b))^2 = \\
 &= \sum_{i=1}^n (y_i - \bar{y})^2 + a^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \\
 &\quad + n(\bar{y} - a\bar{x} - b)^2 - 2a \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\
 &= (n-1)(s_y^2 + a^2 s_x^2 - 2a \text{Cov}[x, y]) + n(\bar{y} - a\bar{x} - b)^2.
 \end{aligned}$$

- Unico punto di minimo

$$a = \frac{\text{Cov}[x, y]}{s_x^2}; \quad b = \bar{y} - \frac{\text{Cov}[x, y]}{s_x^2} \bar{x};$$

- il minimo dell'errore S vale

$$(n-1) \left(s_y^2 - \frac{(\text{Cov}[x, y])^2}{s_x^2} \right) = (n-1) s_y^2 \left(1 - (\rho[x, y])^2 \right)$$

- la retta ha equazione

$$y = \bar{y} + \frac{\text{Cov}[x, y]}{s_x^2} (x - \bar{x}).$$

È detta **retta di regressione del campione y sul campione x** .

Remark

Il punto (\bar{x}, \bar{y}) appartiene alla retta.