Statistica descrittiva

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*

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

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, $i = 1, \ldots, n$. Suddivido l'intervallo $[a, b)$ in N parti uguali (N sarà suggerito dall'esperienza). Otteniamo N intervalli

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

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

$$z_1, \ \dots, \ z_k, \qquad 1 \leqslant k \leqslant n.$$

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

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

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

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

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

$$n_j := \#\left\{i \in \{1, \ldots, n\} \colon x_i \in I_j\right\}.$$

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

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

Moda e valori modali

- $x = (x_1, \dots, x_n)$ campione statistico
- $z_1, z_2, \dots, z_k$ le modalità assunte (o $I_1, \dots, I_k$ 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 $I_{\bar{j}}$) ha frequenza massima, ovvero se esiste un unico $\bar{j} \in \{1, 2, \dots, k\}$ tale che $p_{\bar{j}} \geqslant p_j \ \forall j = 1, \dots, k$, allora la modalità $z_{\bar{j}}$ (o la classe $I_{\bar{j}}$) si dice **moda** del campione x.
2. Se esistono due o più indici $\overline{j_1}, \ \overline{j_2}, \dots, \overline{j}_s$ tali che le modalità $z_{\overline{j_1}}, \ z_{\overline{j_2}}, \dots, z_{\overline{j}_s}$ (o le classi $I_{\overline{j_1}}, \ I_{\overline{j_2}}, \dots, I_{\overline{j}_s}$) hanno frequenza massima, allora queste modalità (o classi) si dicono **valori (o classi) modali**.

Mediana

$x = (x_1, \ldots, x_n)$ un campione relativo ad un carattere numerico.
Ordiniamo i dati del campione in ordine crescente:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n-1)} \leqslant x_{(n)}$$

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

  $$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(m)} \leqslant x_{(m+1)} \leqslant x_{(m+2)} \leqslant \ldots \leqslant x_{(2m)} \leqslant x_{(2m+1)}$$

  $x_{(m+1)}$ è detto **mediana** del campione.
- n pari: $n = 2m$

  $$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(m-1)} \leqslant x_{(m)} \leqslant x_{(m+1)} \leqslant x_{(m+2)} \leqslant \ldots \leqslant x_{(2m)}$$

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

Media

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

Media (aritmetica)

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

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

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots x_n) = \frac{1}{n}(N_1 z_1 + N_2 z_2 + \ldots N_k z_k) =$$

$$= p_1 z_1 + p_2 z_2 + \ldots p_k z_k = \sum_{j=1}^{k} p_j z_j .$$

Varianza campionaria

Varianza campionaria

$$s_x^2 = \mathrm{Var}\left[x\right] := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 .$$

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

scarto quadratico medio (deviazione standard)

$$s_x = \mathrm{Std}\left[x\right] := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

Varianza campionaria 2

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

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

Covarianza

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

$$x = (x_1, \ldots, x_n), \qquad y = (y_1, \ldots, y_n).$$

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

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

dove $\overline{x}$ e $\overline{y}$ 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

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

Coefficiente di correlazione 2/2

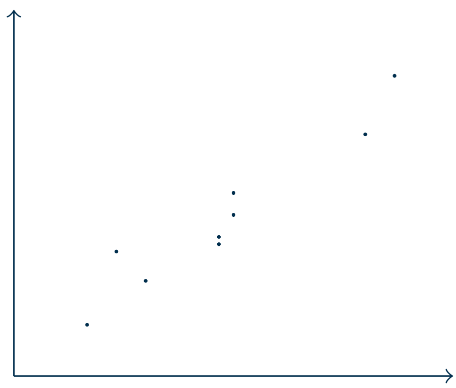
Remark

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

1. $-1 \leqslant \rho[x, y] \leqslant 1$;
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*.

Retta di regressione

$x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ campione bivariato.
Rappresentiamo i punti (x_i, y_i) sul piano cartesiano Oxy.



Se i punti sono *pressoché allineati* si cerca una retta che in qualche senso *approssimi* i punti (x_i, y_i).

$y = ax + b$ 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

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

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

- Unico punto di minimo

$$a = \frac{\mathrm{Cov}\,[x, y]}{s_x^2}; \qquad b = \overline{y} - \frac{\mathrm{Cov}\,[x, y]}{s_x^2}\overline{x};$$

- il minimo dell'errore S vale

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

- la retta ha equazione

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

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

Remark

Il punto $(\overline{x}, \overline{y})$ *appartiene alla retta.*