Calcolo delle Probabilità e Statistica

Luca De Paulis

16 settembre 2020

INDICE

1 | STATISTICA DESCRITTIVA

1.1 CONCETTI BASE

La statistica descrittiva è la branca della statistica che descrive fenomeni statistici senza sfruttare nozioni di probabilità. I concetti fondamentali della statistica descrittiva sono il concetto di *popolazione* e di *campione*: la popolazione è l'insieme delle entità e dei dati che vogliamo studiare, mentre il campione è un piccolo sottoinsieme della popolazione che verrà analizzato per fini statistici.

Altri concetti base sono il concetto di *frequenza assoluta* e *relativa*: si dice frequenza assoluta di un evento A il numero di volte che l'evento accade, senza considerare il numero di eventi (anche di tipo diverso) che accadono; invece si dice frequenza assoluta di un evento A il numero di volte che l'evento accade diviso il numero di eventi totali.

1.2 ANALISI NUMERICA DEI DATI

Supponiamo di avere un vettore $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ che rappresenta i nostri dati. Possiamo definire alcune operazioni fondamentali su questi dati.

Definizione 1.2.1 **Media (empirica.)** Dato $\boldsymbol{x}$ vettore di dati, si dice *media (empirica)* il valore

$$\bar{\boldsymbol{x}} := \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}. \tag{1}$$

Per descrivere quanto i dati contenuti in $\boldsymbol{x}$ si discostano dalla media $\bar{\boldsymbol{x}}$ si usa il concetto di varianza:

Definizione 1.2.2 **Varianza.** Dato $\boldsymbol{x}$ vettore di dati, si dice *varianza campionaria* il valore

$$\operatorname{var}(\boldsymbol{x}) := \sum_{i=1}^{n} \frac{(x_i - \bar{\boldsymbol{x}})^2}{n-1}; \tag{2}$$

si dice invece *varianza empirica* il valore

$$\operatorname{var}_e(\boldsymbol{x}) := \sum_{i=1}^{n} \frac{(x_i - \bar{\boldsymbol{x}})^2}{n}. \tag{3}$$

La varianza campionaria verrà usata quando i dati si riferiscono ad un campione, mentre la varianza empirica sarà più utile per trattare dati riferiti alle popolazioni.

In alcuni casi è utile conoscere la radice quadrata della varianza, quindi definiamo lo *scarto quadratico medio* (o *deviazione standard*) nel seguente modo:

$$\sigma(\boldsymbol{x}) := \sqrt{\operatorname{var}(\boldsymbol{x})}, \quad \sigma_e(\boldsymbol{x}) := \sqrt{\operatorname{var}_e(\boldsymbol{x})}. \tag{4}$$

Proposizione 1.2.3 *Vale la seguente uguaglianza:*

$$\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{\boldsymbol{x}}^2. \tag{5}$$

Dimostrazione.

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}})^2 &= \sum_{i=1}^{n}(x_i^2-2x_i\bar{\boldsymbol{x}}+\bar{\boldsymbol{x}}^2)\\ &= \sum_{i=1}^{n}x_i^2-2\bar{\boldsymbol{x}}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{\boldsymbol{x}}^2\\ &= \sum_{i=1}^{n}x_i^2-2\bar{\boldsymbol{x}}\sum_{i=1}^{n}x_i+n\bar{\boldsymbol{x}}^2\end{aligned}$$

Ricordando che $\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}x_i$, ovvero $\sum_{i=1}^{n}x_i=2\bar{\boldsymbol{x}}$:

$$\begin{aligned}&= \sum_{i=1}^{n}x_i^2-2n\bar{\boldsymbol{x}}^2+n\bar{\boldsymbol{x}}^2\\ &= \sum_{i=1}^{n}x_i^2-n\bar{\boldsymbol{x}}^2.\end{aligned}$$

□

Usando la (5) otteniamo che

$$\operatorname{var}_e(\boldsymbol{x})=\sum_{i=1}^{n}\frac{x_i^2}{n}-\bar{\boldsymbol{x}}^2.$$

Proposizione 1.2.4 *Dato $\boldsymbol{x}$ vettore dei dati, la varianza (campionaria o empirica) di $\boldsymbol{x}$ è 0 se e solo se*

$$x_1=\cdots=x_n=\bar{\boldsymbol{x}}.$$

La dimostrazione è ovvia: essendo la varianza definita come la somma di termini non negativi, essa è uguale a 0 se e soltanto se ogni termine è uguale a 0, ovvero se e solo se $x_i=\bar{\boldsymbol{x}}$ per ogni $i=1,\dots,n$. La varianza quindi rappresenta la "dispersione" dei dati: più è alta, più i dati sono diversi tra loro; più è bassa e più sono vicini (nel caso limite in cui sono tutti uguali la varianza è 0).

Possiamo generalizzare questa idea: fissata una soglia $d\in\mathbb{R}$ consideriamo il numero di elementi x_i la cui distanza dalla media $\bar{\boldsymbol{x}}$ è maggiore o uguale a d:

$$\#\{\,x_1\,:\,|x_i-\bar{\boldsymbol{x}}|\geqslant d\,\}.$$

È possibile dimostrare che vale la seguente disuguaglianza:

$$\#\{\,x_1\,:\,|x_i-\bar{\boldsymbol{x}}|\geqslant d\,\}\leqslant\frac{1}{d}\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}})^2,$$

da cui, dividendo entrambi i membri per n, segue un caso particolare della cosiddetta *disuguaglianza di Chebyshev*:

$$\frac{\#\{\,x_1\,:\,|x_i-\bar{\boldsymbol{x}}|\geqslant d\,\}}{n}\leqslant\frac{\operatorname{var}_e(\boldsymbol{x})}{d}. \tag{6}$$

Il membro sinistro rappresenta la *percentuale* dei dati che si discostano dalla media per un valore superiore alla soglia d.

Un altro metodo utile per ripartire i dati è utilizzare la cosiddetta *funzione di ripartizione empirica*.

Definizione 1.2.5 Dato $\boldsymbol{x}\in\mathbb{R}^n$ vettore dei dati, la funzione di ripartizione empirica è una funzione $F_e:\mathbb{R}\to[0,1]$ tale che

$$F_e(t)=\frac{\#\{\,x_i\,:\,x_i\leqslant t\,\}}{n}.$$

Dunque per ogni soglia t la funzione di ripartizione empirica restituisce la percentuale dei dati che sono minori o uguali a t.

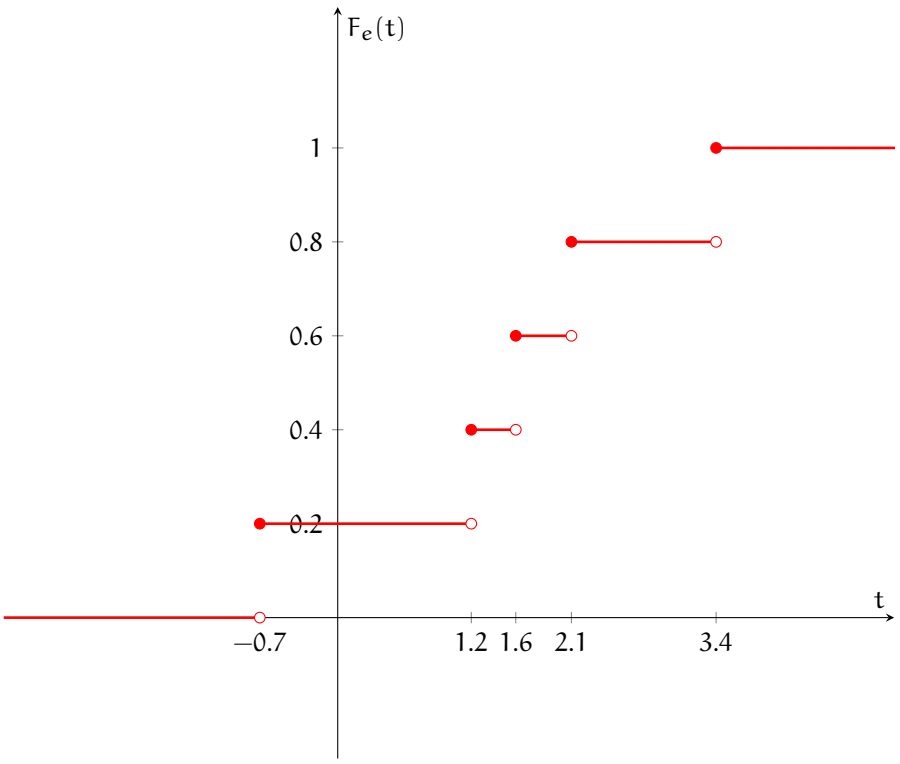
ESEMPIO 1.2.6. Se il vettore dei dati è $\boldsymbol{x} = (1.2, -0.7, 3.4, 1.6, 2.1)$ per trovare la funzione di ripartizione empirica F_e mi conviene innanzitutto ordinarli, ottenendo

$$\boldsymbol{x} = (-0.7, 1.2, 1.6, 2.1, 3.4).$$

A questo punto posso descrivere molto semplicemente la funzione di ripartizione empirica:

- se $t < -0.7$ allora $F_e(t) = 0$: tutti i dati sono maggiori della soglia t;
- se $t \in [-0.7, 1.2)$ allora $F_e(t) = 1/5$: un solo dato è sicuramente minore o uguale a t (ovvero $x_1 = -0.7$), da cui dividendo per $n = 5$ si ottiene $1/5$;
- se $t \in [1.2, 1.6)$ allora $F_e(t) = 2/5$: due dati sono minori o uguali a t (ovvero -0.7 e 1.2), da cui dividendo per $n = 5$ si ottiene $2/5$;
- se $t \in [1.6, 2.1)$ allora $F_e(t) = 3/5$;
- se $t \in [2.1, 3.4)$ allora $F_e(t) = 4/5$;
- se $t \geqslant 3.4$ allora $F_e(t) = 1$ (tutti i dati sono minori o uguali a t, dunque la percentuale è 1);

Il grafico di questa funzione è quindi:



Percentili e quantili

Definizione 1.2.7

Percentile. Sia $k \in \mathbb{R}$ con $0 \leqslant k \leqslant 100$. Allora, dato un vettore dei dati $\boldsymbol{x}$ il k-esimo percentile è un qualsiasi numero $t \in \mathbb{R}$ tale che

- almeno $k/100$ dei dati sono minori o uguali di t,
- almeno $1-k/100$ dei dati sono maggiori o uguali a t.

Intuitivamente un numero reale t è il k-esimo percentile del nostro vettore di dati $\boldsymbol{x}$ se t è il più piccolo numero che è maggiore o uguale al k percento dei dati. Dato che preferiamo trattare numeri compresi tra 0 e 1 invece che tra 0 e 100 introduciamo il concetto di β-quantile: se t è un k-esimo percentile, allora t è un β-quantile per $\beta = k/100$.

Detto più direttamente, un numero t è un β-quantile se

- almeno β dei dati sono minori o uguali a t,
- almeno $1-\beta$ dei dati sono maggiori o uguali a t.

Esempio 1.2.8. Dato il vettore $\boldsymbol{x} = (10, 20, 40, 60, 100)$, il dato $x_4 = 60$ corrisponde all'80-esimo percentile, o equivalentemente allo 0.80-quantile.

Alcuni quantili particolari hanno dei nomi specifici:

- lo 0.25-quantile è anche chiamato *primo quartile*,
- lo 0.50-quantile è anche chiamato *mediana*,
- lo 0.75-quantile è anche chiamato *terzo quartile*.

Dati multipli

In alcuni casi è necessario fare indagini statistiche su dati multipli: rappresentiamo i nostri dati come un vettore di coppie (o triple, o n-uple) di dati:

$$(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \dots, (x_n, y_n)).$$

Per studiare la *correlazione* tra i dati delle x e i dati delle y abbiamo bisogno di alcuni strumenti:

Definizione 1.2.9 **Covarianza.** Dato un vettore di coppie di dati $(\boldsymbol{x}, \boldsymbol{y})$ si dice *covarianza campionaria* il numero

$$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}); \tag{7}$$

si dice invece *convarianza empirica* il numero

$$\operatorname{cov}_e(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}). \tag{8}$$

Definizione 1.2.10 **Coefficiente di correlazione.** Dato un vettore di coppie di dati $(\boldsymbol{x}, \boldsymbol{y})$, se $\sigma(\boldsymbol{x}), \sigma(\boldsymbol{y}) \neq 0$, si dice *coefficiente di correlazione* il numero

$$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma(\boldsymbol{x})\,\sigma(\boldsymbol{y})} = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{\sqrt{\displaystyle\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2}\sqrt{\displaystyle\sum_{i=1}^{n} (y_i - \bar{\boldsymbol{y}})^2}}.$$

Proposizione 1.2.11 *Dato un vettore di coppie di dati $(\boldsymbol{x}, \boldsymbol{y})$ con $\sigma(\boldsymbol{x}), \sigma(\boldsymbol{y}) \neq 0$, vale che*

$$0 \leqslant \left| r(\boldsymbol{x}, \boldsymbol{y}) \right| \leqslant 1.$$

Dimostrazione. Viene dalla disuguaglianza di Cauchy-Schwartz:

$$\sum_{i=1}^{n} \left| (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}) \right| \leqslant \sqrt{\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{\boldsymbol{y}})^2},$$

da cui segue che

$$\left|r(\boldsymbol{x},\boldsymbol{y})\right| = \frac{\sum_{i=1}^{n}\left|(x_i-\bar{\boldsymbol{x}})(y_i-\bar{\boldsymbol{y}})\right|}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{\boldsymbol{y}})^2}} \leqslant 1. \qquad \square$$

Intuitivamente il coefficiente di correlazione misura quanto è semplice approssimare la relazione tra le x e le y con una funzione lineare affine, ovvero con una retta: per vedere ciò cerchiamo di capire quale retta approssima meglio i nostri dati.

Per approssimare linearmente $(\boldsymbol{x},\boldsymbol{y})$ dobbiamo fare in modo che la nostra retta $a+bx$ sia il più vicino possibile ai punti (x_i,y_i) che formano i dati: vogliamo quindi che per ogni punto x_i la distanza tra y_i e $a+bx_i$ sia la minima possibile. Possiamo ottenere quello che vogliamo calcolando il seguente valore:

$$\min_{a,b\in\mathbb{R}^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2. \tag{9}$$

(Eleviamo le distanze al quadrato in modo da renderle tutte positive e le sommiamo insieme poiché vogliamo che la distanza *complessiva* della retta sia minima.)

Teorema 1.2.12 *Il valore minimo della quantità in (9) si ottiene scegliendo*

$$b^* = \frac{\operatorname{cov}(\boldsymbol{x},\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})}, \quad a^* = -b\bar{\boldsymbol{x}}+\bar{\boldsymbol{y}}.$$

Inoltre vale che

$$\min_{a,b\in\mathbb{R}^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2 = \sum_{i=1}^{n}(y_i-\bar{\boldsymbol{y}})^2\left(1-r(\boldsymbol{x},\boldsymbol{y})^2\right).$$

"Dimostrazione". Sia $Q:\mathbb{R}^2\to\mathbb{R}$ la funzione definita da

$$Q(a,b) := \sum_{i=1}^{n}(y_i-a-bx_i)^2.$$

Per dimostrare che questa funzione ha minimo calcoliamo i limiti all'infinito: siccome vale che

$$\lim_{|a|,|b|\to+\infty}\sum_{i=1}^{n}(y_i-a-bx_i)^2 = +\infty$$

per il teorema di Weierstrass generalizzato questa funzone ha minimo. Per calcolarlo, imponiamo che le derivate parziali $\dfrac{\partial Q}{\partial a}$ e $\dfrac{\partial Q}{\partial b}$ siano uguali a 0, da cui ricaviamo le espressioni per a^* e b^*. Sostituendole in Q otteniamo l'espressione per il minimo di Q, che è la seconda parte della tesi. $\square$

La retta a^*+b^*x viene detta *retta di regressione* ed è la funzione lineare affine che meglio approssima i dati che abbiamo a nostra disposizione. Dato che la minima distanza tra la retta e il vettore dei dati (nel senso dato dalla formulazione in (9)) è proporzionale a $\left(1-r(\boldsymbol{x},\boldsymbol{y})^2\right)$, avremo che:

- più $r(\boldsymbol{x},\boldsymbol{y})^2$ si avvicina a 1, più la distanza minima si avvicina a 0 e quindi i dati sono correlati linearmente;

- più $r(\boldsymbol{x}, \boldsymbol{y})^2$ si avvicina a 0, più la distanza minima cresce e quindi i dati sono dispersi e non seguono una correlazione lineare.

Inoltre il coefficiente angolare della retta di regressione b^* può essere riscritto come

$$b^* = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})} = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma(\boldsymbol{x})\,\sigma(\boldsymbol{y})} \frac{\sigma(\boldsymbol{x})\,\sigma(\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})} = r(\boldsymbol{x}, \boldsymbol{y}) \frac{\sigma(\boldsymbol{x})\,\sigma(\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})}.$$

Dato che $\frac{\sigma(\boldsymbol{x})\sigma(\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})} \geqslant 0$ il segno del coefficiente angolare dipende solamente dal coefficiente di correlazione:

- se $r(\boldsymbol{x}, \boldsymbol{y}) > 0$ la retta di regressione ha coefficiente angolare positivo, dunque è crescente e al crescere delle x tendenzialmente crescono anche le y;
- se $r(\boldsymbol{x}, \boldsymbol{y}) < 0$ la retta di regressione ha coefficiente angolare negativo, dunque è decrescente e al crescere delle x tendenzialmente le y decrescono.

Il coefficiente di correlazione dunque ci dice quanto sono correlate le due quantità che stiamo esaminando (più è vicino ad 1 e più sono correlate) e se al crescere della prima cresce anche la seconda (se è di segno positivo), oppure al crescere della prima la seconda diminuisce (se è di segno negativo).