ANALISI 2 | ma 5/5/20

400

INTRODUZIONE ALLA STATISTICA

Nelle scienze applicate (chimica, fisica, biologia, medicina, economia, ecc.) ci si trova spesso in presenza di una grande quantità di dati, provenienti da rilevazioni o da misure, che occorre elaborare ed interpretare, costruendo opportune regole d'inferenza che permettano di trarre le deduzioni che portano alla costruzione di un modello matematico per il problema in questione. Questo è il compito della Statistica: la statistica descrittiva organizza e riassume in modo significativo i dati raccolti; la statistica inferenziale, utilizzando i metodi e le nozioni del calcolo delle probabilità, costruisce le regole d'inferenza da applicare ai dati raccolti.

Possiamo descrivere un problema statistico in questi termini: vi è un insieme molto grande di oggetti, che chiameremo popolazione, al quale sono associate delle quantità misurabili. L'approccio statistico a questo problema consiste nel selezionare un campione, cioè un sottoinsieme ridotto della popolazione, e analizzarlo, sperando di poter trarre delle conclusioni riguardo all'intera popolazione.

Iniziamo a dare un cenno sulla statistica descrittiva.

Statistica descrittiva

Vogliamo descrivere le metodologie usate per raccogliere, rappresentare ed elaborare i dati osservati nell'esame di un certo fenomeno. Il primo problema che si presenta è quello di sintetizzare la massa di dati grezzi in pochi numeri o indicatori particolarmente significativi, attraverso metodi grafici o numerici che descrivano la massa dei dati senza alterarne il senso complessivo.

La statistica descrittiva analizza le caratteristiche di un campione di popolazione, i cui elementi (persone, numeri, oggetti) si chiamano unità statistiche; le caratteristiche sono di tipo qualitativo o numerico.

Esempio:

popolazione	unità statistica	caratteristica	qualità/numero
nati a Pisa nel 2018	bambino	sesso	qualità
studenti di ingegneria	persona	età, media voti	numero
giorni dell'anno	giorno	temperatura	numero.

Si parla di caratteristiche, o attributi, presenti in un certo grado nella popolazione; ci occuperemo solo di caratteristiche di tipo numerico. I dati relativi a tali caratteristiche sono le variabili statistiche.

Fase 1 (rilevazione dei dati). Dopo aver descritto il fenomeno che è l'oggetto dell'indagine, e aver individuato la popolazione e le sue unità statistiche, si scelgono le caratteristiche che si vogliono analizzare e si raccolgono i dati. A questo punto occorre fare lo spoglio dei dati, ossia contarli, ordinarli e classificarli.

Fase 2 (rappresentazione dei dati). Raccolti i dati, essi vanno descritti mediante opportuni grafici e tabelle.

Fase 3 (elaborazione dei dati). Si utilizzano vari metodi per ottenere indici di sintesi sui dati rilevati e relazioni statistiche fra i dati stessi.

Ordinamento dei dati

Siano $z_1, \dots, z_N$ i dati grezzi raccolti. La loro interpretazione è in genere assai difficile a prima vista. Per prima cosa è utile ordinarli in modo crescente:

$$y_1 \le y_2 \le \dots \le y_{N-1} \le y_N .$$

Il range di un insieme di dati $\{z_1, \dots, z_N\}$ è la differenza fra il massimo ed il minimo di essi:

$$\max_{1 \le i \le N} z_i - \min_{1 \le i \le N} z_i = y_N - y_1 .$$

I dati raccolti possono essere in parte coincidenti: quindi, se $x_1, \dots, x_n$ $(n \le N)$ sono i valori distinti delle z_i, si ha

$$y_j \in \{x_1, \dots, x_n\} \quad \forall j = 1, 2, \dots, N .$$

Se denotiamo con n_j il numero dei dati z_i uguali a x_j $(j = 1, \dots n)$, si ha $\sum_{j=1}^{n} n_j = N$. Gli n_j sono le frequenze assolute con cui si presentano i dati $x_1, \dots, x_n$, mentre i valori

$$f_j = \frac{n_j}{N}, \quad j = 1, \dots, n,$$

sono le frequenze relative dei dati $x_1 \dots x_N$. Ovviamente, $\sum_{j=1}^{N} f_j = 1$.

Se i dati $x_1, \dots, x_n$ sono ordinati, si definiscono anche la frequenza cumulata assoluta rispetto a x_k:

$$R_k = \sum_{j=1}^{k} n_j \,, \quad k = 1, \dots, n,$$

e la frequenza cumulata relativa, sempre rispetto a x_k,

$$F_k = \sum_{j=1}^{k} f_j \,, \quad k = 1, \dots, n.$$

Esempio In una classe di 28 ragazzi, alla domanda "Quale sport preferisci?", 10 ragazzi hanno scelto il calcio, 4 il tennis, 6 la pallavolo, 3 il basket, 5 hanno dato altre risposte. Si ha

sport	freq. assoluta	freq. relativa
calcio	10	0.36
tennis	4	0.14
pallavolo	6	0.21
basket	3	0.11
altro	5	0.18

Qui non si possono ordinare i dati, quindi non ha senso parlare di frequenze cumulative.
Però, alla stessa classe viene chiesto "Quanti anni hai?", e le risposte sono: 18 anni (4 persone), 19 (10 persone), 20 (in 7), 21 (in 3), 22 (1 persona), si ha:

età	freq. ass.	freq. cum. ass.	freq. rel.	freq. rel. ass.
18	4	4	0.14	0.14
19	13	17	0.46	0.60
20	7	24	0.25	0.85
21	3	27	0.11	0.96
22	1	28	0.04	1.00

Se l'insieme dei dati da studiare è troppo grande, si possono raggruppare in classi. Se X è una variabile che assume valori in $[a,b]$, possiamo dividere $[a,b]$ in m intervalli disgiunti (di uguale ampiezza, o anche no)

$[a_0,a_1[\,,\ [a_1,a_2[\,,\ \dots,\ [a_{m-1},a_m]$, con $a=a_0<a_1<\dots<a_m=b$.

I dati vengono raggruppati nelle classi di appartenenza, calcolando le frequenze assolute di classe $n_1,\dots,n_m$ (n_k è il numero di valori compresi fra a_{k-1} e a_k) e le frequenze relative di classe $p_1,\dots,p_m$ (p_k è uguale a $\frac{n_k}{N}$, se N è il n° dei dati osservati).

Il numero delle classi non deve essere troppo alto, altrimenti ciascuna classe conterrebbe pochi dati, né troppo basso, perché si perderebbe troppa informazione sulla distribuzione reale dei dati. Per prassi consolidata, il numero di classi m deve soddisfare

$$m \simeq 1+\frac{10}{3}\log_{10}N.$$

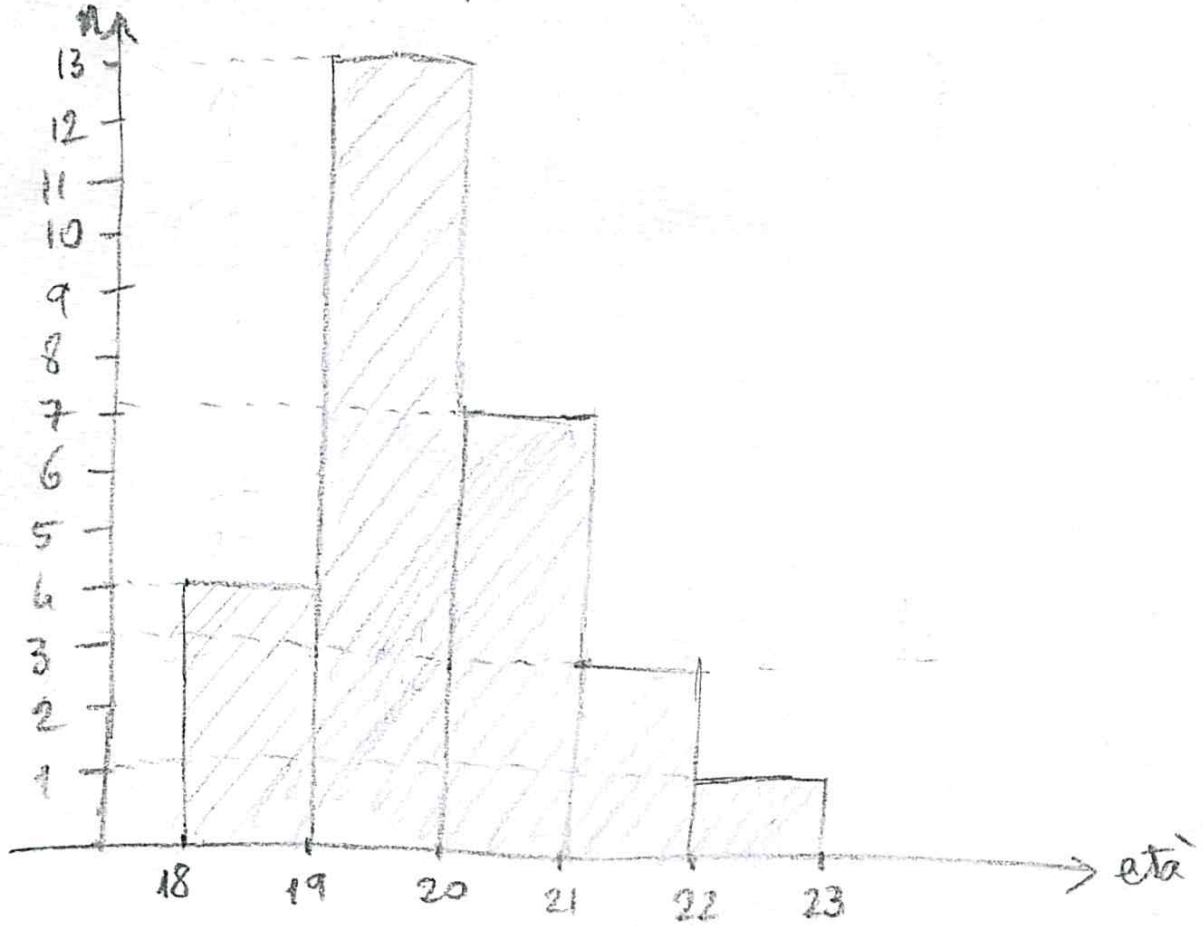
Non facciamo un esempio esplicito, che però si trova a pag. 137-138 degli appunti di probabilità e statistica.

Spesso accanto alle tabelle si usano rappresentazioni quali grafici a torta, in cui ogni spicchio ha ampiezza proporzionale alla frequenza relativa della corrispondente classe, oppure grafici a barre, con lunghezze di ciascuna barra proporzionale alla corrispondente frequenza relativa o assoluta. Ancora, si hanno

istogrammi, in cui si divide l'asse delle ascisse in intervalli contigui di ampiezza proporzionale alle classi, e si riporta su ogni intervallo un rettangolo di altezza proporzionale alla frequenza della classe corrispondente.

Infine ci sono i poligoni di frequenza, in cui i dati sono rappresentati da una linea spezzata. Ogni classe è rappresentata dal valore centrale dell'intervallo ad essa relativo; l'ordinata corrispondente è la frequenza di tale classe. I punti del piano così ottenuti si uniscono con una spezzata. Se i poligoni di frequenza si riferiscono a frequenze cumulate, la spezzata sarà crescente.

Qui sotto l'istogramma delle frequenze assolute delle età relative all'esempio precedente.



Misure quantitative

Sono di due tipi: misure di posizione e misure di dispersione.

a) misure di posizione.

Dato un insieme di dati numerici $Z=\{z_1,\dots,z_N\}$, la loro media aritmetica è

$$\mathbb{E}[Z]=\bar{z}=\frac{1}{N}\sum_{i=1}^{N} z_i;$$

utilizzando i valori effettivamente distinti $\{x_1\dots,x_n\}$ e le rispettive frequenze relative $f_1,\dots,f_n$, si ha

$$\bar{z}=\frac{1}{N}\sum_{k=1}^{n} n_k x_k = \sum_{k=1}^{n} f_k x_k .$$

Notiamo che la media aritmetica è lineare: se $Y=\{y_1,\dots,y_N\}$ è un altro insieme di N dati, e se $a,b\in\mathbb{R}$, allora la media aritmetica di $U=\{ay_1+bz_1,\dots,ay_N+bz_N\}$ è

$$\bar{u}=a\bar{y}+b\bar{z}.$$

La media geometrica dei numeri positivi $Z=\{z_1,\dots,z_N\}$ è

$$\bar{z}_g=\left(\prod_{j=1}^{N} z_j\right)^{1/N}$$

Essa è la più appropriata in molte situazioni.

Esempi (1) Una banca ha applicato negli ultimi 8 anni i seguenti tassi di interesse composto: 1.1% nei primi 2 anni, poi 1.9% per 3 anni, 1.5% per 2 anni e infine 1.4% nell'ultimo

anno. Qual è stato il tasso medio annuo?
Detti r_i i tassi applicati ogni anno e r il tasso medio, si ha:

$$r_1 = r_2 = 0.011, \quad r_3 = r_4 = r_5 = 0.019, \quad r_6 = r_7 = 0.015, \quad r_8 = 0.014.$$

Essendo r il tasso medio, deve essere

$$(1+r)^8 = \prod_{i=1}^{8} (1+r_i),$$

e dunque

$$1+r = \left(\prod_{i=1}^{8} 1+r_i \right)^{\frac{1}{8}} \simeq 1.01537.$$

(2) Un parallelepipedo ha lati lunghi 8, 5 e 25 cm; qual è la lunghezza del lato del cubo di uguale volume?
Dev'essere $\ell^3 = 8 \cdot 5 \cdot 25 \text{ cm}^3 = 1000 \text{ cm}^3$, ossia $\ell = 10$ cm: ℓ è la media geometrica di 8, 5, 25.

La <u>media armonica</u> dei numeri positivi $Z = \{z_1, \dots, z_N\}$ è

$$\bar{z}_a = \frac{N}{\sum_{i=1}^{N} \frac{1}{z_i}}.$$

Anche questa media è utile in certi casi.

Esempio. Un'automobile percorre avanti e indietro un tratto di strada, alla velocità costante di 80 km/h all'andata e 120 km/h al ritorno. Qual è la velocità media sull'intero percorso?
Se il tratto è lungo s, i tempi impiegati sono (in ore)

all'andata, $t_a = \frac{s}{80}$; al ritorno, $t_r = \frac{s}{120}$;
quindi la velocità media sull'intero percorso è

$$V = \frac{2s}{t_a + t_r} = \frac{2}{\frac{1}{80} + \frac{1}{120}} = 96 \text{ km/h}.$$

Notiamo che

$$\bar{z}_a \leq \bar{z}_g \leq \bar{z} \qquad \forall \{z_1, \dots, z_N\} \subseteq]0, +\infty[.$$

La mediana di un insieme di dati $\{z_1, \dots, z_N\}$, ordinati in modo crescente o decrescente, è il valore centrale se N è dispari, mentre è la media aritmetica dei due valori centrali se N è pari. Se, ad esempio,

$$Z = \{3, 85, 96, 97, 99, 100, 103, 105, 106\}$$

la mediana è 99; se aggiungiamo ai dati il numero 107, la mediana è 99.5. Invece la media aritmetica dei 9 dati è 88; con in più il 10° numero essa diventa 89.9.
La media aritmetica è fortemente influenzata dai valori estremi, mentre la mediana non ne risente. Se ad esempio sostituiamo il 1° dato 3 con 86, la mediana non cambia, ma la media sale a 96 (con 9 numeri) e a 98 (con 10 numeri).
Dunque l'uso della mediana (in luogo della media aritmetica) è consigliabile quando i dati sperimentali sono pochi, o sono in gran parte eccentrici verso un estremo.

Dato un insieme di dati ordinati $Z = \{z_1, \dots, z_N\}$, il quantile di ordine p ($p \in]0,1[$) è l'unico valore alla sinistra del quale vi è una frazione del totale dei dati pari a p.

Dunque il quantile di ordine p è quel valore q tale che la frequenza relativa cumulata F_q verifica

$$F_q \geq p \qquad F_k \leq p \quad \forall k = 1, \dots, q-1.$$

Ad esempio, il quantile di ordine $7/9$ dell'insieme di dati di pag. 408 è 103. Si noti che la mediana è il quantile di ordine $1/2$.

I quartili di un insieme di dati sono

Q_1 (alla sua sinistra il 25% dei dati)

Q_2 (la mediana)

Q_3 (alla sua sinistra il 75% dei dati)

In modo analogo si definiscono i decili e i centili (o percentili).

La moda di un insieme di dati $\{z_1, \dots z_N\}$, con valori effettivi $\{x_1, \dots, x_n\}$ e frequenze assolute $\{n_1, \dots, n_n\}$ è un valore che compare con frequenza massima, cioè è un x_k tale che $n_k = \max_{1 \leq j \leq n} n_j$. Tale x_k può non essere unico. Se i dati sono raggruppati in classi, la classe modale è quella (non necessariamente unica) che ha

frequenza massima.
La moda può risultare utile quando i dati non sono di tipo numerico (esempio: professione, scelta politica, luogo di nascita, ecc.). Però è poco utile nel caso non sia unica.

Osserviamo che se una distribuzione di dati è unimodale (cioè ha moda unica) e simmetrica (cioè $x_j = x_{n-j}$ per ogni j), allora media aritmetica, mediana e moda coincidono.

b) misure di dispersione

Le misure di posizione non danno informazioni sul modo in cui i dati di una variabile numerica $Z = \{z_1, \dots, z_N\}$ sono distribuiti intorno al valore centrale: due distribuzioni numeriche possono avere lo stesso valore centrale e valori molto vicini ad esso oppure molto dispersi. È inutile naturalmente fare la media aritmetica delle deviazioni dalla media $z_1 - \bar{z}, \dots, z_N - \bar{z}$, che sarà nulla; conviene invece considerare la deviazione media

$$\frac{1}{N} \sum_{j=1}^{N} |z_j - \bar{z}|,$$

che però è scomoda da maneggiare, a causa dei valori assoluti. Si considera invece la media dei quadrati delle deviazioni, cioè la varianza

$$\mathrm{Var}[Z] = (\sigma_Z)^2 = \frac{1}{N} \sum_{j=1}^{N} (z_j - \bar{z})^2,$$

che è nulla se e solo se $z_j = \bar{z}$ per ogni $j = 1 \dots N$, mentre è

strettamente positiva se esistono almeno due indici i, j tali che $z_j \neq z_i$. Il numero

$$N(\sigma_z)^2 = \sum_{j=1}^{N} (z_j - \bar{z})^2$$

è la devianza di Z. Invece la radice quadrata della varianza è lo scarto quadratico medio o deviazione standard

$$\sigma_z = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (z_j - \bar{z})^2},$$

il quale, a differenza della varianza, è espresso nella stessa unità di misura dei dati.

Il rapporto fra la deviazione standard e il modulo della media, $\sigma_z / |\bar{z}|$, è il coefficiente di variabilità di Z ed è spesso espresso in forma percentuale, anche se esso non è in generale un numero compreso fra 0 e 1.

Elenchiamo le proprietà della varianza:

$$Var[z] = \sigma_z^2 = \overline{z^2} - \bar{z}^2 = E[z^2] - E[z]^2,$$

se $W = \{z_1 + a, \dots, z_N + a\}$, allora $\sigma_z^2 = \sigma_W^2$;

se $Y = \{a z_1, \dots, a z_N\}$, allora $\sigma_Y^2 = a^2 \sigma_z^2$.

se $X = \left\{ \frac{z_1 - \bar{z}}{\sigma_z}, \dots, \frac{z_N - \bar{z}}{\sigma_z} \right\}$, allora $\bar{x} = 0$ e $\sigma_X = 1$.

Il passaggio dai dati Z ai dati X è la standardizzazione.