D. Statistica descrittiva

La statistica descrittiva si occupa della presentazione e della sintesi di insiemi di dati (di solito numerici).

Presentazione dei dati in forma di tabelle o grafica (diagrammi); spesso utile classificazione (raggruppamenti in classi) dei dati.

Sintesi dei dati attraverso le statistiche, cioè funzioni numeriche definite sull'insieme dei dati ("calcolabili dai dati").

Consideriamo prima dati univariati, cioè ~~insiemi~~ successioni finite di numeri reali:

"insieme ordinato"/successione $\mathcal{X} = \{(i, x_i) : i \in \{1, \dots, n\}\}$

Sia $(x_i)_{i \in \{1, \dots, n\}} \subset \mathbb{R}$ l'insieme dei dati (il campione). La ~~cardinalità~~ lunghezza n è la numerosità del campione.

Statistiche elementari per la sintesi dei dati:

a) "centro dei dati":

media campionaria, mediana campionaria,

moda o valori modali

b) "dispersione dei dati":

varianza campionaria e deviazione standard campionaria

c) "distribuzione dei dati": percentili campionari

(casi particolari: quartili, mediana)

2) Statistiche del "centro dei dati":

Sia $(x_i)_{i\in\{1,\dots,n\}} \subset \mathbb{R}$ il campione.

Def.: La media campionaria del campione è data da $\bar{x} \doteq \frac{1}{n}\sum_{i=1}^{n} x_i$. (media aritmetica della successione)

Esempio: $x_i = i, \quad i \in \{1,\dots,n\}$

$$\rightsquigarrow \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\rightsquigarrow \bar{x} = \frac{n+1}{2}$$

In questo caso, la media campionaria è il "valore centrale".

La media campionaria è ~~molto~~ fortemente influenzata da valori estremi, anche poco frequenti: Ad esempio,

$n = 100, \quad x_i = i$ per $i \in \{1,\dots,99\}$,
$x_{100} = 10^6$

$$\rightsquigarrow \bar{x} = \frac{1}{100}\left(\frac{100}{2}\cdot 99 + 10^6\right) = 10049.5 .$$

La media campionaria si può riscrivere come media pesata dei valori assunti dai dati:

↓

Sia $v_1, \dots, v_K$ un'enumerazione dell'insieme dei valori dei dati, cioè

$$\{v_j : j \in \{1, \dots, K\}\} = \{x_i : i \in \{1, \dots, n\}\} \quad e \quad v_j \neq v_l \text{ per } j \neq l.$$

Sia f_j la <u>frequenza assoluta</u> del valore v_j:

$$f_j \doteq \#\{i \in \{1, \dots, n\} : x_i = v_j\}.$$

La media campionaria $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ si può allora scrivere come ~~la media pesata~~

$$\bar{x} = \sum_{j=1}^{K} \underbrace{\frac{f_j}{n}}_{\text{frequenza relativa del valore } v_j} \cdot v_j$$

media dei valori pesata (ponderata) dalle corrispondenti frequenze relative

Def.: I valori che hanno frequenza massima, cioè i valori v_j con $j \in \operatorname{argmax}\{f_l : l \in \{1, \dots, K\}\}$ si dicono <u>valori modali</u>. Se ~~il~~ esiste ~~valore~~ un unico valore di frequenza massima, cioè $\operatorname{argmax}\{f_l : l \in \{1, \dots, K\}\} = \{j_*\}$ per un (unico) $j_* \in \{1, \dots, K\}$, allora v_{j_*} si dice <u>moda campionaria</u>.

Sia ~~ξ~~ σ una permutazione degli indici $\{1,..,n\}$ (cioè σ una biiezione $\{1,..,n\} \to \{1,..,n\}$)

tale che $X_{\sigma(1)} \leq X_{\sigma(2)} \ldots \leq X_{\sigma(n)}$

(ovvero $X_{\sigma(i)} \leq X_{\sigma(i+1)}$ per ogni $i \in \{1,..,n-1\}$).

La successione $(X_{\sigma(i)})_{i\in\{1,..,n\}}$ è quindi un riordinamenti del campione in ordine crescente.

Def.: La <u>mediana campionaria</u> del campione è data dal valore

$$\bar{m} \doteq \begin{cases} X_{\sigma(\frac{n+1}{2})} & \text{se } n \text{ è dispari,} \\ \frac{1}{2}\left(X_{\sigma(\frac{n}{2})} + X_{\sigma(\frac{n+2}{2})}\right) & \text{se } n \text{ è pari.} \end{cases}$$

Esempio (cf. media campionaria):

dati già in ordine crescente

1) $x_i = i$, $i \in \{1,..,n\}$, allora

$$\bar{m} = \frac{n+1}{2} \quad \text{(sia per } n \text{ pari che dispari).}$$

In questo caso, $\bar{m} = \bar{x}$, media e mediana campionaria coincidono!

2) $n = 100, \quad x_i = i, \quad i \in \{1, \dots, 99\},$

$x_{100} = 10^6$

dati già in ordine crescente

$\rightsquigarrow \quad \bar{m} = \frac{1}{2}(50 + 51) = 50.5,$

mentre $\bar{x} = 10049.5$!

Infatti, $\bar{m}$ non cambierebbe se x_{100} assumesse un qualsiasi altro valore ≥ 51.

La mediana divide i dati in tre parti;

$\{i \in \{1, \dots, n\} : x_i < \bar{m}\}, \quad \{i \in \{1, \dots, n\} : x_i = \bar{m}\}, \quad \{i \in \{1, \dots, n\} : x_i > \bar{m}\}$

(possibilmente vuoto)

in modo che $\# \{i \in \{1, \dots, n\} : x_i < \bar{m}\} = \# \{i \in \{1, \dots, n\} : x_i > \bar{m}\}.$

"metà dei dati sotto, metà sopra la mediana".

Attenzione: conta anche la frequenza dei valori.

b) Statistiche della "dispersione dei dati"

Sia $(x_i)_{i\in\{1,\dots,n\}} \in \mathbb{R}$ il campione.

Vogliamo avere una misura ~~~~ per la dispersione dei dati intorno alla media campionaria; si dà più peso a deviazioni grandi, meno a deviazioni piccole.

Def.: Per $n \geq 2$, la varianza campionaria del campione è data da

$$s^2 \doteq \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

La quantità $s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)}$

si dice deviazione standard campionaria.

Osservazioni:

1) La varianza campionaria si basa ~~~~ sullo scarto quadratico tra media campionaria e valore dei dati; è "~~approssimata~~ quasi" uguale alla media campionaria degli scarti quadratici (asintoticamente uguale, $n \to \infty$).

2) Il prefattore $\frac{1}{n-1}$ invece di $\frac{1}{n}$ serve per ottenere uno stimatore "corretto" o "non distorto" quando i dati sono quantità aleatorie (cf. infra).

↓

3) Siano ~~...~~ $a, b \in \mathbb{R}$. Poniamo

$$y_i = a \cdot x_i + b, \quad i \in \{1, \dots, n\}.$$

Sia s_y^2 la varianza campionaria di (y_i),
s_x^2 " " " " " (x_i).

Allora $s_y^2 = a^2 \cdot s_x^2$, e $s_y = |a| \cdot s_x$

Il campione (y_i) è una trasformazione linear-affine di (x_i). La varianza campionario non dipende dalla traslazione b, e si trasforma con a^2, mentre la deviazione standard si trasforma con ~~...~~ il fattore $|a|$, ~~(a se a>0), mentre ... quindi~~ (oppure a se $a > 0$).
La deviazione standard ha quindi la stessa unità di misura dei dati.

4) Sviluppando i quadrati, la varianza campionaria si riscrive come

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} \bar{x}^2.$$

utile per i calcoli

Esempio: ↗ Foglio 1

c) Statistiche per la distribuzione dei dati

Come per la mediana campionaria, sia

$(x_{\sigma(i)})_{i\in\mathbb{N}}$ un riordinamento dei dati in ordine crescente.

Def.: Il percentile campionario K-esimo, con $K\in\{0,\dots,100\}$, del campione è dato dal valore

$$\bar{p}_K \doteq \begin{cases} x_{\sigma\left(\lceil \frac{n\cdot K}{100}\rceil\right)} & \text{se } n\cdot\frac{K}{100}\notin\mathbb{N}_0, \\ \frac{1}{2}\left(x_{\sigma(n\cdot\frac{K}{100})}+x_{\sigma(n\cdot\frac{K}{100}+1)}\right) & \text{se } n\cdot\frac{K}{100}\in\mathbb{N}_0. \end{cases}$$

[Per $x\in\mathbb{R}$ poniamo $\lceil x\rceil \doteq \min\{\ell\in\mathbb{Z}: \ell\geq x\}$.]

numeri interi

arrotondamento per eccesso

Osservazioni:

1) Il percentile 50-esimo coincide con la mediana.

Si dicono primo, secondo, terzo e ~~quartile~~ quarto quartile i percentili 25-esimo, 50-esimo, 75-esimo e 100-esimo.

2) La proporzione degli indici $i\in\{1,\dots,n\}$ tali che $x_i\leq\bar{p}_K$ è maggiore di o uguale a $\frac{K}{100}$:

$$\frac{\#\{i\in\{1,\dots,n\}: x_i\leq\bar{p}_K\}}{n}\geq\frac{K}{100}.$$ Più precisamente,

$$\#\{i\in\{1,\dots,n\}: x_i\leq\bar{p}_K\}\geq\lceil n\cdot\tfrac{K}{100}\rceil,\quad \#\{i\in\{1,\dots,n\}: x_i>\bar{p}_K\}\leq n-\lceil n\cdot\tfrac{K}{100}\rceil.$$

La varianza campionaria (equivalentemente, la deviazione standard) permette di stimare la proporzione dei dati che sono vicini alla (o lontani dalla) media campionaria, grazie alla disuguaglianza di Chebyshev. L'unità di misura spesso più conveniente è la deviazione standard

Disuguaglianza di Chebyshev (versione campionaria):

Siano $\bar{x}$ la media e sia s la deviazione standard campionaria del campione $(x_i)_{i\in\{1,\dots,n\}} \subset \mathbb{R}$.

Se $s > 0$, allora per ogni $\alpha > 0$:

a) $$\frac{\#\{i \in \{1,\dots,n\} : |x_i - \bar{x}| < \alpha \cdot s\}}{n} \geq 1 - \frac{(n-1)}{n\alpha^2} > 1 - \frac{1}{\alpha^2},$$

b) $$\frac{\#\{i \in \{1,\dots,n\} : |x_i - \bar{x}| \leq \alpha \cdot s\}}{n} > 1 - \frac{(n-1)}{n \cdot \alpha^2},$$

unilaterale

c) $$\frac{\#\{i \in \{1,\dots,n\} : x_i - \bar{x} \geq \alpha \cdot s\}}{n} < \frac{1}{1+\alpha^2}.$$

Se $s = 0$, allora $x_i = \bar{x}$ per ogni $i \in \{1,\dots,n\}$.

Parti a) e b) interessanti per $\alpha \geq 1$ (almeno $\alpha > \sqrt{\frac{n-1}{n}}$).

↓

Osservazioni:

1) La disuguaglianza di Chebyshev (nelle varie ~~ragioni~~ versioni) fornisce ~~una~~ stime per il caso peggiore, ma è valida per un qualsiasi campione di dati.

2) Le stime di sopra possono essere migliorate quando si hanno più informazioni sulla distribuzione dei dati (in particolare, per dati "approssimativamente normali").

3) Scegliendo $\alpha \in \{2, 3, 5\}$ nella 2) otteniamo:

- $$\frac{\#\{i \in \{1, \ldots, n\} : |x_i - \bar{x}| < 2s\}}{n} > \frac{3}{4} = 75\%$$
- $$\frac{\#\{i \in \{1, \ldots, n\} : |x_i - \bar{x}| < 3s\}}{n} > \frac{8}{9} \approx 89\%$$
- $$\frac{\#\{i \in \{1, \ldots, n\} : |x_i - \bar{x}| < 5s\}}{n} > \frac{24}{25} = 96\%$$

Deviazioni di più di cinque ~~quattro~~ volte la deviazione standard sono quindi abbastanza ~~molto~~ rare, persino nel caso peggiore.

Spesso i dati non sono univariati, ma bi- o multivariati, cioè elementi di $\mathbb{R}^d$ ($d=2$ nel caso bivariato), oppure con una struttura ancora più generale (a valori in uno spazio metrico).

In questo caso, ~~tra~~ è di interesse quantificare una possibile dipendenza tra le componenti (marginali) dei dati. Una statistica fondamentale per misurare la dipendenza tra due marginali di un campione di dati multivariati è la correlazione campionaria, a sua volta definita in termini delle deviazioni standard delle componenti e della covarianza campionaria.

Sia $(x^{(l)})_{l\in\{1,\dots,n\}} \subset \mathbb{R}^d$ un campione di numerosità n di dati d-variati. [Notazione: $x^{(l)} = (x_1^{(l)},\dots,x_d^{(l)})$]

Def.: La covarianza campionaria tra componente i-esima e j-esima del campione ($i,j\in\{1,\dots,d\}$) è data da
$$\mathrm{cov}_{i,j} \doteq \frac{1}{n-1}\left(\sum_{l=1}^{n}(x_i^{(l)}-\overline{x}_i)(x_j^{(l)}-\overline{x}_j)\right),$$
dove $\overline{x}_i$, $\overline{x}_j$ sono le relative medie campionarie marginali

Osservazione:

La matrice $d \times d$ data da

$$cov := (cov_{i,j})_{i,j \in \{1,\dots,d\}}$$ si dice

<u>matrice di covarianza</u>. Essa è

- simmetrica (cioè $cov_{i,j} = cov_{j,i}$ $\forall i,j$),
- e semidefinita positiva (cioè $x^T cov\, x \geq 0$ $\forall x \in \mathbb{R}^d$).

Le entrate sulla diagonale sono non-negative, infatti $\forall i \in \{1,\dots,d\}$: $cov_{i,i}$ è ~~una~~ la varianza della componente i-esima.

Def.: La correlazione campionaria tra le componenti i e j del campione è data da

$$\text{corr}_{i,j} \doteq \frac{\text{cov}_{i,j}}{s_i \cdot s_j},$$

dove s_i, s_j sono le deviazioni standard campionarie delle componenti i e j.

Osservazioni:

1) Per definizione,

$$\text{corr}_{i,j} = \frac{\sum_{l=1}^{d} (x_i^{(l)} - \bar{x}_i) \cdot (x_j^{(l)} - \bar{x}_j)}{\sqrt{\left(\sum_{l=1}^{d} (x_i^{(l)} - \bar{x}_i)^2\right) \cdot \left(\sum_{l=1}^{d} (x_j^{(l)} - \bar{x}_j)^2\right)}}$$

Nota: Stesso risultato se nella definizione di varianza e covarianza si usa il prefattore $\frac{1}{n}$ al posto di $\frac{1}{n-1}$.

2) $\text{corr}_{i,j} \in [-1, 1]$. Inoltre,

corr misura per il grado di dipendenza lineare-affine

$\text{corr}_{i,j} = 1$ se e solo se $\exists\, a, b \in \mathbb{R}$ con $b > 0$:
$x_j^{(l)} = a + b \cdot x_i^{(l)} \quad \forall\, l \in \{1, \ldots, n\}$.

$\text{Corr}_{i,j} = -1$ se e solo se $\exists\, a, b \in \mathbb{R}$ con $b < 0$:
$x_j^{(l)} = a + b \cdot x_i^{(l)} \quad \forall\, l \in \{1, \ldots, n$

Domanda: A cosa serve la probabilità nel contesto della statistica?

Spesso: Dati di interesse riguardano una "popolazione" (di individui, oggetti, "entità" generali) molto più numerosa della numerosità del campione a disposizione.

[Esempio: ci interessa l'età degli ~~individui~~ "italiani" ~~di una~~]

Per poter trarre conclusioni dai dati del campione sulle ~~distri~~ corrispondenti grandezze per la popolazione bisogna scegliere un campione rappresentativo.

Metodo più affidabile: scegliere "a caso" un sottoinsieme "abbastanza numeroso" della popolazione

⇝ dati del campione diventano risultato di un "esperimento aleatorio"

⇝ serve un <u>modello</u> <u>matematico</u> per descrivere esperimenti aleatori.

Altri motivi per introdurre un modello matematico per "il caso":

1) Dati "perturbati", in particolare errori di osservazione / misurazione:

$\cancel{Z}$ $\quad Y = X +$ "rumore"

valore rilevato $\nearrow$ $\quad$ $\nwarrow$ valore "vero"

2) Impossibilità di prevedere grandezze di interesse con certezza; per mancanza di informazioni o per la natura del fenomeno

(ad esempio, previsioni meteo, giochi d'azzardo, finanza)

Alcuni modelli semplici che incontreremo:

1) lancio di monete, dadi come ingrediente principale dell'esperimento ($\nearrow$ Kids University)

2) dinamiche aleatorie: passeggiata aleatoria

catene di Markov