Modelli descrittivi, statistica e simulazione

Master per Smart Logistics specialist

Roberto Cordone
(roberto.cordone@unimi.it)

Statistica descrittiva — Cernusco S.N., martedì 28 febbraio 2017

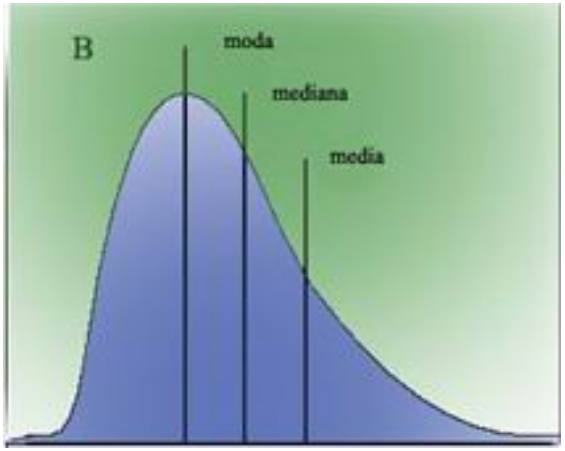
Indici di posizione

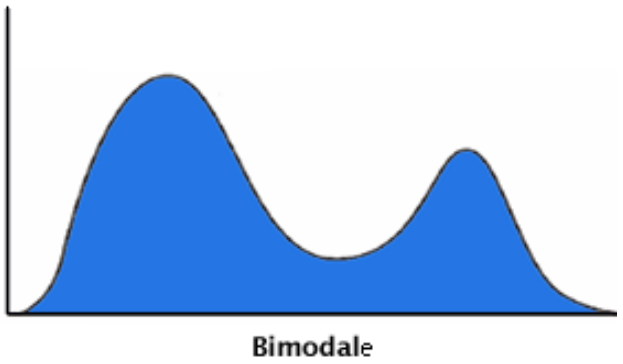
Un indice di posizione cerca di riassumere l'intera distribuzione di una popolazione in un solo valore

Idee possibili:

- per i caratteri nominali, l'unico indice possibile è la moda, cioè il valore più frequente
- per i caratteri quantitativi o qualitativi ordinali, si può definire anche la mediana, cioè il valore tale che metà della popolazione lo precede e metà lo segue
- per i caratteri quantitativi, si può definire anche la media, che combina i valori delle singole unità in modo che sommandoli si ottenga lo stesso valore

Moda





Vantaggi

- è l'unico indice possibile per i caratteri nominali
- è significativa se la sua frequenza assoluta prevale nettamente sugli altri valori

(*Vedi Esercizio 1-3*)

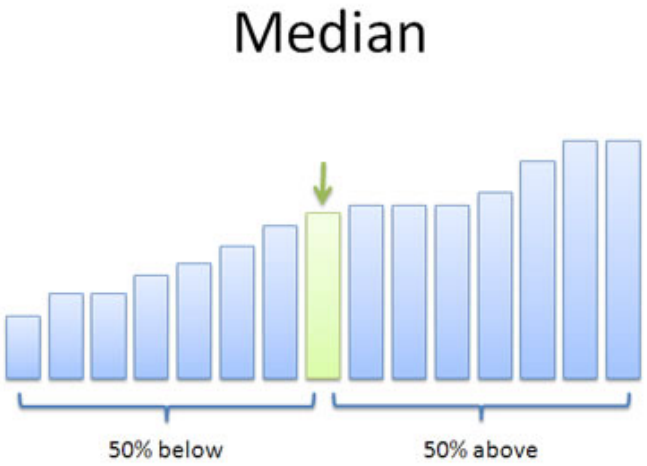
Svantaggi

- non è significativa se la frequenza è quasi uniforme
- può essere multipla
  (se il carattere è ordinale, anche con mode principali e secondarie)

Mediana

La mediana divide la popolazione in due popolazioni di ugual numero:

1. una con il 50% di unità precedenti la mediana
2. una con il 50% di unità successive alla mediana



Si calcola ordinando la popolazione e contando gli elementi

- se sono dispari, si prende l'elemento intermedio

$$10\ 11\ 13\ \mathbf{15}\ 16\ 23\ 26 \qquad \text{mediana} = 15$$

- se sono pari, si prende il valor medio dei due elementi intermedi:

$$10\ 11\ \mathbf{13}\ \mathbf{15}\ 16\ 23 \qquad \text{mediana} = \frac{13+15}{2} = 14$$

Mediana

Vantaggi

- richiede solo che i caratteri siano ordinati (*Vedi Esercizio 2-1*)
- è robusta rispetto a unità con valori estremi (*outliers*)
- per caratteri quantitativi, minimizza la somma degli scarti (differenze) assoluti fra i valori misurati e la mediana

$$\sum_{i \in M} |x_i - \text{mediana}|\, n_i \text{ è minima}$$

cioè è il valore mediamente più vicino a tutti gli altri

Nota: la somma è sulle singole unità, non sui valori dei caratteri

Quantili

Il concetto di mediana si può generalizzare considerando elementi che dividono la popolazione in proporzioni diverse dal $50\% - 50\%$

Per esempio, i cinque quartili

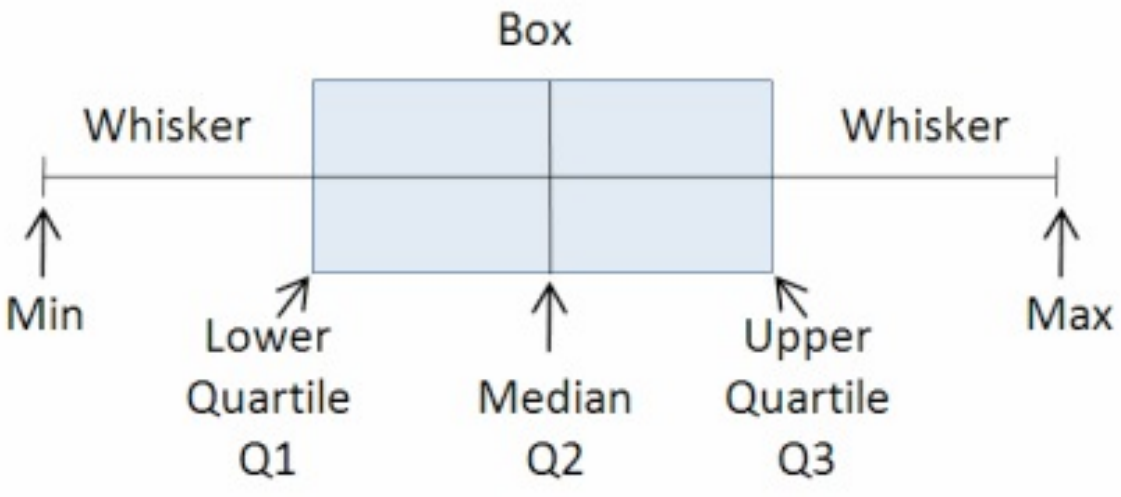
0. $Q0$ o minimo è il valore minimo della popolazione
1. $Q1$ o primo quartile è il valore che separa il 25% inferiore della popolazione dal 75% superiore
2. $Q2$ o secondo quartile è un altro nome della mediana
3. $Q3$ o terzo quartile è il valore che separa il 75% inferiore della popolazione dal 25% superiore
4. $Q4$ o massimo è il valore massimo della popolazione

Analogamente si definiscono quintili, decili, percentili

Diagrammi *box-and-whiskers*

I quartili si rappresentano graficamente con i diagrammi box-and-whiskers (o box-plot)

- un segmento che collega minimo e primo quartile
- una scatola rettangolare che collega primo e secondo quartile
- una scatola rettangolare che collega secondo e terzo quartile
- un segmento che collea terzo quartile e massimo



(*Vedi Esercizio 2-2*)

Media

La media aritmetica è il numero che, sostituito a ciascun carattere in una somma su tutta la popolazione, lascia la somma invariata

$$\mu = \frac{\sum_{i \in M} x_i n_i}{N}$$

dove la frequenza n_i raccoglie in un termine le unità di ugual valore x_i

È utile perché capita spesso di sommare i caratteri per calcolare un totale (per es., quando rappresentano costi, tempi, pesi, volumi. . .)

Se si fanno operazioni diverse dalla somma, sono opportune altre medie:

- la media geometrica lascia invariato il prodotto
- la media armonica lascia invariata la somma degli inversi

Media

Vantaggi:
Gode di molte proprietà matematiche utili nei calcoli

- rende nulla la somma degli scarti

$$\sum_{i \in M} (x_i - \mu) n_i = 0$$

- minimizza la somma dei quadrati degli scarti

$$\sum_{i \in M} (x_i - \mu)^2 n_i$$

Svantaggi

- è influenzata dai valori estremi, anche pochi, se molto diversi
  (*"se io mangio due polli e tu zero, è come mangiarne uno a testa"*)

Indici di variabilità

Non sempre gli indici di posizione descrivono in modo soddisfacente la distribuzione

Dati tre insiemi di tempi di carico merci o tempi di servizio alla cassa

Popolazione	Tempi					
1	10	12	24	24	37	40
2	18	22	24	24	29	30
3	21	24	24	24	26	28

qual è la situazione migliore (più efficiente, cioè con i tempi minori)?

Indici di variabilità

Non sempre gli indici di posizione descrivono in modo soddisfacente la distribuzione

Dati tre insiemi di tempi di carico merci o tempi di servizio alla cassa

Popolazione	Tempi					
1	10	12	24	24	37	40
2	18	22	24	24	29	30
3	21	24	24	24	26	28

qual è la situazione migliore (più efficiente, cioè con i tempi minori)?

Le tre popolazioni hanno

1. moda 24, mediana 24 e media 24.5
2. moda 24, mediana 24 e media 24.5
3. moda 24, mediana 24 e media 24.5

Ma la terza è preferibile, perché molto più concentrata:
gli indici di posizione sono significativi e si possono usare
per prevedere la qualità del servizio e decidere se investire per migliorarlo

Indici ordinali

Gli indici ordinali di variabilità sono analoghi alla mediana

Sono differenze fra quantili:

- campo di variazione (o range) è la differenza fra massimo e minimo

$$R = Q4 - Q0$$

- differenza interquartile è la differenza fra primo e terzo quartile

$$IQR = Q3 - Q1$$

La seconda é più robusta (la prima subisce molto gli outliers)

Questi indici

- sono nulli se e solo se i valori sono tutti uguali
- crescono quando la variabilità della popolazione cresce

Indici di dispersione

Gli indici di dispersione invece sono analoghi alla media e legati ad essa

- varianza è la media aritmetica dei quadrati degli scarti

$$\sigma^2 = \frac{\sum_{i \in M} n_i (x_i - \mu)^2}{N}$$

uguale alla differenza fra media dei quadrati e quadrato della media

$$\sigma^2 = \frac{\sum_{i \in M} n_i x_i^2}{N} - \mu^2$$

(*che è molto più veloce e semplice*)

- scarto quadratico o deviazione standard è la radice quadrata della varianza

$$\sigma = \sqrt{\frac{\sum_{i \in M} n_i (x_i - \mu)^2}{N}}$$

Anche questi indici

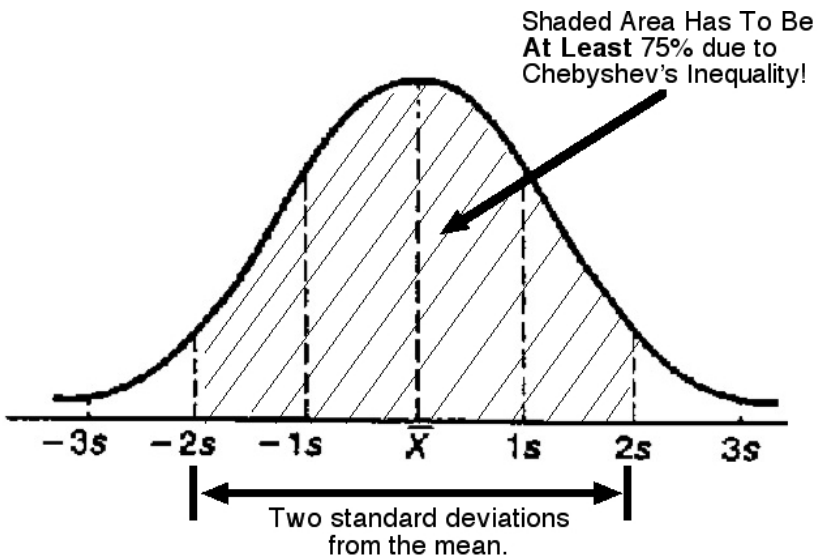
- sono nulli se e solo se i valori sono tutti uguali
- crescono quando la variabilità della popolazione cresce

Disuguaglianza di Čebyšëv

Si dimostra che una frazione pari ad almeno $\left(1 - \frac{1}{k^2}\right)$ della popolazione cade nell'intervallo $[\mu - k\sigma, \mu + k\sigma]$

In termini più pratici

- almeno il 75% dei valori sono compresi tra $\mu - 2\sigma$ e $\mu + 2\sigma$
- almeno il 96% dei valori sono compresi tra $\mu - 5\sigma$ e $\mu + 5\sigma$
- almeno il 99% dei valori sono compresi tra $\mu - 10\sigma$ e $\mu + 10\sigma$



Nota bene: l'intervallo è stimato per eccesso!

I percentili effettivi, una volta calcolati, di solito sono più stretti

(*Vedi Esercizio 2-2*)

Indici di forma: asimmetria

Con uguali indici di posizione e variabilità, si può avere forma diversa

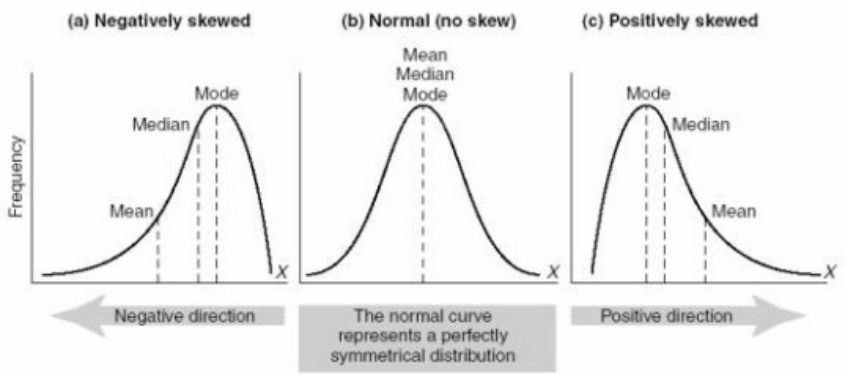
L'asimmetria (o skewness) indica se la distribuzione pende da un lato

$$CS = \frac{1}{N} \cdot \frac{\sum_{i \in M} n_i (x_i - \mu)^3}{\sigma^3}$$

Il segno indica il verso della pendenza

- se $CS < 0$, pende a sinistra;
- se $CS = 0$, è (complessivamente) simmetrica;
- se $CS > 0$, pende a destra.

Il valore assoluto misura l'intensità: per $|CS| > 1$, l'asimmetria è forte



Nota: la definizione di Excel è leggermente diversa (più grande in valore assoluto)

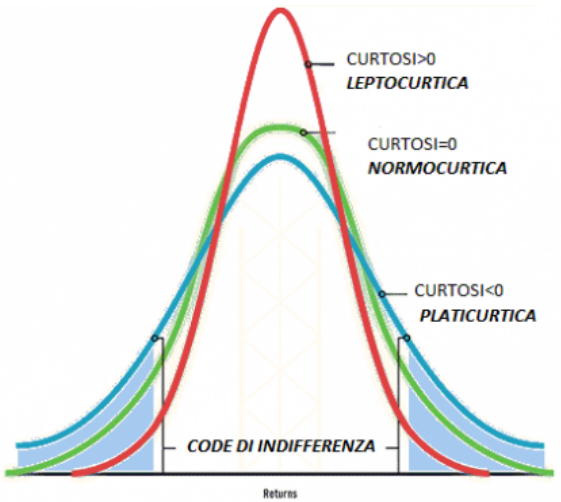
Indici di forma: curtosi

La curtosi misura la curvatura della distribuzione

$$CK = \frac{1}{N} \cdot \frac{\sum_{i \in M} n_i (x_i - \mu)^4}{\sigma^4}$$

Il segno della differenza $CK - 3$ indica l'intensità della curvatura

- se $CK - 3 < 0$, la distribuzione è piatta (platicurtica);
- se $CK - 3 = 0$, la distribuzione è moderatamente curva;
- se $CK - 3 > 0$, la distribuzione è appuntita (leptocurtica).



(*Vedi Esercizio 2-3 ed Esercizio 2-4*)

Nota: Excel calcola $CK - 3$ e ha una definizione leggermente diversa (maggiore)