Lezione 5: Indici statistici di variabilità

Corso di Statistica
Facoltà di Economia
Università della Basilicata

Prof. Massimo Aria

- aria@unina.it

Variabilità

La **variabilità** di una distribuzione esprime la tendenza delle unità statistiche di un collettivo ad assumere diverse modalità del carattere.

In un processo mirato alla descrizione di un fenomeno, l'individuazione di un indice di posizione non può ritenersi esaustivo.

Infatti alla conoscenza della tendenza centrale si accompagna l'esigenza di descrivere quanto l'indice di posizione considerato possa ritenersi realmente rappresentativo dei valori assunti dalle unità del collettivo.

In altre parole si vuole capire quanto le modalità osservate sulla popolazione siano vicine o lontane dal "centro" della distribuzione.

In una disamina delle misure di variabilità di una distribuzione si distingue usualmente tra:

- *dispersione rispetto ad un centro*
- *mutevolezza delle frequenze*
- *dispersione reciproca*

Un esempio esplicativo

Consideriamo il seguente esempio di tre studenti che hanno superato ciascuno tre esami:

A	18	24	30
B	23	24	25
C	24	24	24

È facile verificare che se calcoliamo il voto medio e quello mediano per ciascun studente esso è pari a 24

Si può affermare che i tre studenti hanno uno stesso comportamento agli esami?

Dall'esempio risulta evidente che da soli gli indici di posizione non riescono a svelare esaustivamente il "segreto" delle distribuzioni!!

Indice di variabilità

Una **misura di variabilità** *V* definita sulle osservazioni (x_1, x_2,…, x_N) è tale se soddisfa i seguenti *assiomi*:

1) L'indice *V* è non negativo

2) L'indice *V* è nullo quando le unità assumono tutte la stessa modalità

3) L'indice *V* non muta quando a tutte le modalità è aggiunta (o sottratta) una costante

4) Se *V(X)>V(Y)* allora X è più variabile di Y

5) L'indice *V* aumenta al crescere della variabilità

Assiomi di un indice di variabilità

1) $V(x_1, x_2, \ldots, x_N) \geq 0$
2) $V(c, c, \ldots, c) = 0$
3) $V(x_1 + c, x_2 + c, \ldots, x_N + c) = V(x_1, x_2, \ldots, x_N)$
4) $V(x_1, x_2, \ldots, x_N) \geq V(y_1, y_2, \ldots, y_M)$ X più variabile di Y

Assiomi per la definizione di un indice di variabilità.

Tipologia di indici di variabilità

Gli indici di variabilità si distinguono in tre categorie:

a) **Indici che misurano la variabilità rispetto ad una misura di posizione**.
Questi si basano su una sintesi degli scarti delle modalità rispetto al valore centrale di riferimento (i.e, la media)

b) **Indici che misurano la variabilità rispetto all'ordinamento delle modalità**.
Questi si basano sulla funzione di ripartizione empirica e quindi all'ordine che assumono le modalità nella distribuzione considerata

c) **Indici che misurano la variabilità reciproca tra tutte le modalità considerate due a due**.
Effettuano una sintesi dell'insieme degli scostamenti tra i valori della distribuzioni considerati due per volta.

Variabilità rispetto a un centro

Un indice di variabilità rispetto ad un centro misura la presenza o meno di una certa stabilità dei valori assunti dalle unità rispetto alla misura di tendenza centrale.

Gli indici maggiormente diffusi si basano sul concetto di "scarto" (o scostamento) delle modalità rispetto alla media (intesa come media aritmetica).

Tra questi, si ricordano:
- la **varianza**
- la **devianza**
- lo **scarto quadratico medio**
- lo **scostamento semplice dalla media**

Scarti dalla media

$$(x_1 - \mu), (x_2 - \mu), \ldots, (x_N - \mu)$$

….o semplicemente **scarti**

Varianza

L'indice più importante per esprimere la variabilità di una distribuzione rispetto a un centro è la **varianza**.

Essa si definisce come la *media degli scarti al quadrato.*

Come è facile verificare gode di tutte le caratteristiche necessarie agli indici di variabilità:

1) È una misura non negativa
2) cresce al crescere della misura degli scarti e quindi della variabilità della distribuzione
3) È nulla se le unità assumono tutte lo stesso valore (*variabile degenere*).
4) Se si aggiunge una costante a tutte le osservazione, la misura degli scarti non cambia e quindi la varianza resta immutata.

Varianza

Distribuzione unitaria $$\sigma^2 = \frac{1}{N}\sum (x_l - \mu)^2$$

Distribuzione di frequenza $$\sigma^2 = \frac{1}{\sum n_i}\sum (x_i - \mu)^2 n_i$$

Formulazione della varianza

Devianza e Scarto quadratico medio

Si definisce **devianza** la *somma degli scarti al quadrato*.

Essa è pari al numeratore della varianza.

Si definisce **scarto quadratico medio** (sqm) la *radice della media degli scarti al quadrato*.

Esso è pari alla radice quadrata della varianza.

Tra i tra indici, lo SQM è quello che si presta a più facile interpretazione in quanto espresso nella stessa unità di misura della variabile X.

Lo SQM può leggersi come "*lo scostamento medio delle modalità della distribuzione rispetto alla media*".

Devianza

Distribuzione unitaria $$SS = \sum (x_l - \mu)^2$$

Distribuzione di frequenza $$SS = \sum (x_i - \mu)^2 \cdot n_i$$

Formulazione della devianza

Scarto Quadratico Medio

Distribuzione unitaria $$\sigma = \sqrt{\frac{1}{N}\sum (x_l - \mu)^2} = \sqrt{\sigma^2}$$

Distribuzione di frequenza $$\sigma = \sqrt{\frac{1}{\sum n_i}\sum (x_i - \mu)^2 \cdot n_i} = \sqrt{\sigma^2}$$

Formulazione dello Scarto Quadratico Medio

Scostamento semplice

Si definisce **scostamento semplice dalla media** S(M) la media degli scarti in valore assoluto dalla media.

Si definisce **scostamento semplice dalla mediana** S(Me) la media degli scarti in valore assoluto dalla mediana.

Scostamento semplice

Dalla media $$S(\mu) = \frac{1}{N}\sum|x_i - \mu|$$

Dalla mediana $$S(Me) = \frac{1}{N}\sum|x_i - Me|$$

Formulazione dello scostamento semplice rispetto alla media e alla mediana

Coefficiente di variazione

Le misure presentate sino ad ora rappresentano degli **indici di variabilità assoluta**.

Essi infatti assumono valori in una scala di variazione che dipende strettamente dall'unità di misura e dall'intervallo in cui la variabile assume valori.

Ciò rende difficile il confronto tra distribuzioni diverse (si pensi alla comparazione tra la variabilità del peso dei neonati e delle mamme!!)

Per ovviare a questo problema si ricorre alla costruzione di **indici di variabilità relativa**.

Il più diffuso è il **coefficiente di variazione** (CV) che si ottiene rapportando lo SQM alla media in valore assoluto.
Il risultato è una misura proporzionale della variabilità rispetto alla media.

Il CV è indipendente dall'unità di misura, cioè *è un numero puro.*

Coefficiente di Variazione

$$CV = \frac{\sigma}{|\mu|} \cdot 100$$

$$\text{con } \mu \neq 0$$

Indici di variabilità delle modalità ordinate

Sono misure di variabilità derivate dalla funzione di ripartizione empirica, attraverso l'uso dei concetti di percentili e quartili di una distribuzione.

Gli indici più utilizzati sono:

- il **campo di variazione** (*range*) R(X)

È definito come differenza tra il valore massimo (100° percentile) e minimo (1° percentile) della distribuzione.

- la **differenza inter-quartile** IQR(X)

È definita come differenza tra il terzo e il primo quartile della distribuzione.

Le due misure si differenziano per il **grado di robustezza**.

Il range risente anche di un solo valore anomalo mentre la differenza interquartile, escludendo le code della distribuzione, è meno influenzata da valori estremi della stessa.

Campo di variazione

$$R(X) = x_{\max} - x_{\min}$$

Differenza Inter-Quartile

$$IQR(X) = Q_3 - Q_1$$

Un esempio di calcolo degli indici di variabilità

x_i	n_i	$x_i \cdot n_i$	$(x_i - \mu)$	$(x_i - \mu)^2 \cdot n_i$	$\lvert x_i - \mu \rvert \cdot n_i$	$\lvert x_i - Me \rvert \cdot n_i$
10	2	20	-4,48	40,14	8,96	10
12	3	36	-2,48	18,45	7,44	9
13	4	52	-1,48	8,76	5,92	8
15	7	105	0,52	1,89	3,64	0
16	5	80	1,52	11,55	7,6	5
17	3	51	2,25	19,05	7,56	6
18	1	18	3,52	12,39	3,52	3
	25	**362**		**112,24**	**44,64**	**41**

$Media\ \mu = \frac{362}{25} = 14{,}48$

$Mediana\ Me = 15$

$Devianza\ SS = \sum (x_i - \mu)^2 \cdot n_i = 112{,}24$

$Varianza\ \sigma^2 = \frac{\sum (x_i - \mu)^2 \cdot n_i}{N} = \frac{112{,}24}{25} = 4{,}49$

$SQM\ \sigma = \sqrt{4{,}49} = 2{,}12$

$S(\mu) = \frac{\sum |x_i - \mu| \cdot n_i}{N} = \frac{44{,}64}{25} = 1{,}79$

$S(Me) = \frac{\sum |x_i - Me| \cdot n_i}{N} = \frac{41}{25} = 1{,}64$

$CV = \frac{\sigma}{|\mu|} \cdot 100 = \frac{2{,}12}{14{,}48} \cdot 100 = 14{,}63\%$

Mutabilità

La **mutabilità** è l'espressione della variabilità nel contesto di caratteri qualitativi.

Si parla di **massima eterogenietà** di una mutabile quanto tutte le modalità assumono pari frequenza assoluta o relativa (es. variabile genere: Maschi 50%, Femmine 50%).

Al contrario si parla di **massima omogeneità** di una mutabile quando le unità assumono tutte lo stesso attributo (es. variabile genere: Maschi 100%, Femmine 0%).

Un indice di mutabilità è l'**indice di eterogenietà** H di Gini.

L'indice di eterogeneità di Gini

$$H = 1 - \sum_{i=1}^{k} f_i^2$$

- In presenza di massima omogeneità

$$H_{min} = 1 - \sum f_i^2 = 1 - (0 + 0 + \ldots 1 + 0) = 0$$

- In presenza di massima eterogeneità

$$H_{max} = 1 - \sum f_i^2 = 1 - \sum \left(\frac{1}{k}\right)^2 = 1 - k\left(\frac{1}{k^2}\right) = \frac{k-1}{k}$$

Formulazione dell'indice di eterogeneità

Concentrazione

La **concentrazione** di una variabile X deriva dalla possibilità di trasferire l'ammontare del fenomeno da un'unità statistica ad un'altra, avvicinandosi o allontanandosi dalla situazione di equidistribuzione dell'ammontare complessivo della variabile.

Si parla di **concentrazione minima** (equidistribuzione) quando l'ammontare complessivo della variabile è ripartito in misura uguale tra tutte le unità statistiche.

Si parla di **concentrazione massima** quando l'ammontare complessivo della variabile è posseduto da un'unica unità statistica mentre le rimanti posseggono 0.

Per misurare la concentrazione si costruisce un indice che confronta la frazione cumulata di unità statistiche (p_l) con la frazione cumulata di ammontare del fenomeno (q_l).

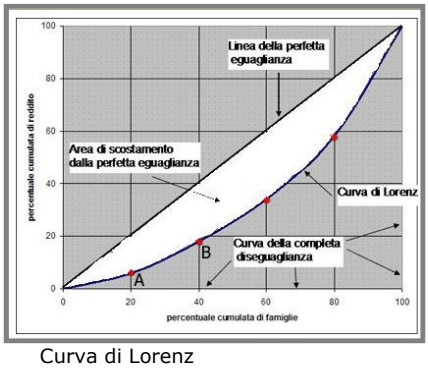
La concentrazione è rappresentata attraverso un grafico, **curva di Lorenz**, dove la bisettrice è pari alla situazione di equidistribuzione e l'area compresa tra questa e la curva misura invece l'indice R.

Rapporto di concentrazione di Gini

$$R = \frac{\sum_{l=1}^{N-1} (p_l - q_l)}{\sum_{l=1}^{N-1} p_l}$$

$$R = \begin{cases} 0 & \text{equidistribuzione} \\ 1 & \text{massima concentrazione} \end{cases}$$

Formulazione del rapporto di concentrazione



Curva di Lorenz

Nella prossima lezione

Nella prossima lezione si affronteranno i seguenti argomenti:

- standardizzazione di una variabile
- indici di curtosi e asimmetria