VARIABILITÀ

(INDICI di VARIABILITÀ)
perché l'indice di posizione

⇒ attitudine del fenomeno a manifestarsi in modi diversi
- un fenomeno si manifesta diversamente in modi diversi
- osservazione ripetuta di una stessa grandezza

⇒ PROPRIETÀ di un indice di variabilità:

1. vale 0 se tutte le modalità sono uguali (= anche a μ)
2. aumenta al crescere della variabilità
3. NON varia per trasformazioni di posizione (traslazioni)

NB Noi vedremo:
- Scostamenti medi
- Varianza
- Intervalli di variazione
- Indici relativi di variabilità
- Variabilità per le variabili qualitative

SCOSTAMENTI MEDI ASSOLUTI di ORDINE S → S_μ^s

indici di variabilità

Distribuzione disaggregata di X:

$$S_\mu^{s} = \left\{ \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|^s \right\}^{\frac{1}{s}} \quad s \neq 0$$

μ: indice di posizione

- Media potenziata delle differenze tra le modalità con cui si è manifestato il fenomeno X e μ
- $\forall s$ e $\forall \mu$, si ha che $S_\mu^s = 0$ SE E SOLO SE X ha variabilità nulla (1 sola modalità che coincide con μ)

⇓

DISTRIBUZIONE DEGENERE

- Distribuzione di freq. ass. per X con K modalità:

$$S_\mu^s = \left\{ \frac{1}{N} \sum_{i=1}^{K} |x_i - \mu|^s n_i \right\}^{1/s}$$

- Distribuzione di freq. relativa

$$S_\mu^s = \left\{ \sum_{i=1}^{K} |x_i - \mu|^s f_i \right\}^{1/s}$$

VARIANZA → $Var(X)$ → σ^2

$$\Rightarrow Var(X) = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N} = \sum_{i=1}^{K} \frac{(x_i-\mu)^2 n_i}{N} = \sum_{i=1}^{K} (x_i-\mu)^2 f_i \geq 0$$

Distr. DISAGG. — Distr. freq. ass. — Distr. freq. rel.

Scostamento medio assoluto di ordine 2 dalla media aritmetica, elevato al quadrato

- Più noto indice di variabilità
- $Var(X) \geq 0$ — misura dell'attitudine di X ad assumere valori diversi
- Si indica spesso con σ^2_x (sigma quadro)
- Quando X è degenere (tutti i valori osservati sono = tra loro) la varianza assume valore 0
- Scarto quadratico medio = $\sqrt{Var(X)} = SD = \sigma_x$ = DEVIAZIONE STANDARD = scostamento medio assoluto di ordine 2

NB: NON esiste un limite superiore alla variabilità
⇓
non si può stabilire se un fenomeno sia tanto o poco variabile in assoluto
⇓
- Si possono però fare confronti di variabilità tra fenomeni

FORMULA OPERATIVA DI CALCOLO per ottenere la varianza x distr. disagg.

$$Var(X) = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N} = \sum_{i=1}^{N} \frac{(x_i^2+\mu^2-2x_i\mu)}{N} = \sum_{i=1}^{N} \frac{x_i^2}{N} + \sum_{i=1}^{N} \frac{\mu^2}{N} - \sum_{i=1}^{N} \frac{2x_i\mu}{N} =$$

$$= \sum_{i=1}^{N} \frac{x_i^2}{N} + \mu^2 - 2\mu \sum_{i=1}^{N} \frac{x_i}{N} = \boxed{\sum_{i=1}^{N} \frac{x_i^2}{N} - \mu^2}$$ = MEDIA DEI QUADRATI − QUADRATO DELLA MEDIA

VARIANZA di UNA TRASFORMAZIONE LINEARE: (IMPORTANTE si trova spesso)

Si consideri $y = a + bx$, $a, b \in \mathbb{R}$. Ricordando la media di una trasformazione lineare

$$Var(y) = \sum_{i=1}^{N} \frac{(y_i-\mu)^2}{N} = \sum_{i=1}^{N} \frac{(a+bx_i-a-b\mu)^2}{N} = \sum_{i=1}^{N} \frac{(bx_i-b\mu)^2}{N} = b^2 \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N} =$$

$$= b^2 Var(X)$$

NB: LA VARIANZA NON È UN OPERATORE LINEARE

$y = a + bX \rightarrow Var(y) = b^2 Var(x)$

$\mu_y = a + b\mu_x$

COEFFICIENTE di VARIAZIONE → CV(X)

- è un indice relativo di VARIABILITÀ ⇒ elimina l'effetto dell'ordine di grandezza di un fenomeno e è ADIMENSIONALE
- Strumento per il confronto di variabilità tra due o più variabili
- Strumento per il confronto di variabilità tra fenomeni analoghi rilevati su gruppi distinti di unità statistiche
- Indice che ELIMINA l'effetto dell'ordine di grandezza del fenomeno
- Il coefficiente di variazione è ADIMENSIONALE

$$CV(X) = \frac{\sqrt{Var(X)}}{|\mu|} = \frac{\text{Deviazione standard di } X}{|\mu|}$$

⇓ EVITA che vi siano EFFETTO di COMPENSAZIONE per i quali la media assume valori prossimi allo 0 rendendo falsamente ampio il coefficiente di variazione

INTERVALLI DI VARIAZIONE → Δ_c e Δ_q

- Indicatore ottenuto come DIFFERENZA tra valori che occupano particolari posizioni nella distribuzione

- Campo di variazione

$$\boxed{\Delta_c = x_{(N)} - x_{(1)}} = \text{valore massimo} - \text{valore minimo}$$

(*) Validità limitata poichè non considera gli altri N-2 termini della distribuzione

- Scarto interquartilico

$$\boxed{\Delta_q = q_3 - q_1}$$

(*) Validità limitata poichè può essere NULLO anche se non è nulla la variabilità

MUTABILITÀ ⇒ Variabilità per variabili QUALITATIVE

⇓ non è possibile misurare la distanza tra le modalità (xk non sono numeriche)

⇓ si valuta la distribuzione delle frequenze sul supporto della variabile

MINIMA ⇓ si ha quando una sola modalità presenta la max frequenza (pari al tot. delle osservazioni)

MASSIMA ⇓ si ha quando tutte le modalità presentano le stesse frequenze (= numero di u.s per ciascuna modalità)

- l'indice di MUTABILITÀ deve rappresentare alcune caratteristiche
  - * valore minimo in caso di minima mutabilità
  - * valori crescenti al crescere della mutabilità
  - * valore massimo quando la mutabilità è massima

- Distribuzione di frequenze relative $(x_1, f_1) \dots (x_k, f_k)$

2 indici di mutabilità

INDICE DI GINI

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

MIN MUTUAB

$$G = 1 - (0^2 + \dots + 0^2 + 1^2 + 0^2 \dots + 0^2) = 0$$

Solo un valore $f_k \neq 0$

MAX MUTUAB

$$G = 1 - \left(\frac{1}{k^2} + \dots + \frac{1}{k^2}\right) = 1 - k \cdot \frac{1}{k^2} = \frac{k-1}{k}$$

$f_k = \frac{1}{k} \ \forall k$

INDICE di ENTROPIA di SHANNON

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

NB1 c'è il − perché il ln di f_i è un valore compreso tra 0 e 1, quindi è un valore negativo. Così cambia il segno

MIN MUTUAB

$$H = -(0 \ln 0 + \dots + 0 \ln 0 + 1 \ln 1 + 0 \ln 0 + \dots + 0 \ln 0) = 0$$

Solo un valore $f_n \neq 0$

MAX MUTUAB

$f_n = \frac{1}{k} \ \forall k$

$$H = -\left(\frac{1}{k} \ln \frac{1}{k} + \dots + \frac{1}{k} \ln \frac{1}{k}\right) = -k \cdot \frac{1}{k} \ln \frac{1}{k} = -\ln \frac{1}{k} = \ln k$$

Quindi

$$0 \leq G \leq \frac{k-1}{k} \qquad 0 \leq H \leq \ln k$$

NB2 INDICI di MUTUAB. NORMALIZZATI

Normalizzazione: trasformazione di un indice I con un proprio campo di variazione $[I_{min}, I_{max}]$ in un indice definito sull'intervallo $[0, 1]$

$$\boxed{\tilde{I} = \frac{I - I_{min}}{I_{max} - I_{min}}}$$

IND. di GINI NORM: $\boxed{\tilde{G} = \frac{G}{\frac{k-1}{k}} = \frac{k}{k-1} G}$

ENTROPIA di SHANNON NORM: $\boxed{\tilde{H} = \frac{H}{\ln k}}$

NB3 $\tilde{G} = 0$ o $\tilde{H} = 0 \Rightarrow$ MIN MUTUABILITÀ

$\tilde{G} = 1$ o $\tilde{H} = 1 \Rightarrow$ MAX MUTUABILITÀ

STANDARDIZZAZIONE

(usata molto)

- strumento per rendere i dati confrontabili tra loro

$\Downarrow$

Sia X variabile di media μ e varianza σ^2:

- VARIABILE STANDARDIZZATA $\boxed{Z = \frac{X-\mu}{\sigma}}$ da cui $\boxed{X = \mu + Z\sigma}$

  $Z = -\frac{\mu}{\sigma} + \frac{1}{\sigma}X$

- MEDIA di Z $\boxed{\mu_Z = \frac{\mu-\mu}{\sigma} = 0}$ $\rightarrow$ $\mu_Z = -\frac{\mu}{\sigma} + \frac{1}{\sigma}X \cdot \mu = -1+1=0$

- VARIANZA di Z $\boxed{Var(Z) = Var\left(\frac{X-\mu}{\sigma}\right) = \frac{Var(X)}{\sigma^2} = 1}$

  $= Var\left(\frac{X}{\sigma}\right) = \frac{1}{\sigma^2} Var(X) = \frac{\sigma^2}{\sigma^2} = 1$

SCORE

- la quantità $\boxed{z_i = \frac{x_i-\mu}{\sigma}}$ detta score misura il NUMERO di SCARTI quadratici medi che separano una osservazione x_i dal valore medio μ

$\Downarrow$

- $z_i = k$, $k>0$: l'osservazione è sopra la media μ di k volte lo scarto quadratico medio $\Rightarrow \frac{x_i-\mu}{\sigma} = k \Rightarrow \boxed{x_i = \mu + k\sigma}$
- $z_i = k$, $k<0$: l'osservazione è sotto la media μ di k volte lo scarto quadratico medio $\Rightarrow \frac{x_i-\mu}{\sigma} = -k \Rightarrow \boxed{x_i = \mu - k\sigma}$
- $z_i = k$, $k=0$: l'osservazione coincide con la media