## Esempio

| cap. investito | prezzo titoli | quantità |
|---|---|---|
| $L_i$ | $p_i$ | $q_i$ |

Obiettivo: determinare prezzo medio che lascia invariata la disponibilità monetaria globale.

$\Rightarrow \bar{\bar{p}} = ?$

$$\sum_{i=1}^{r} L_i = \sum_{i=1}^{r} p_i \cdot q_i \Rightarrow \sum_{i=1}^{r} p_i \cdot q_i = \sum_{i=1}^{r} \bar{\bar{p}} \cdot q_i \Rightarrow \bar{\bar{p}} = \frac{\sum_{i=1}^{r} p_i \cdot q_i}{\sum_{i=1}^{r} q_i}$$

$q_i$ è il sistema dei pesi

risolvo per $\bar{\bar{p}}$

NB: $\bar{\bar{p}}$ non è una media in senso stretto.

## Esempio

Si considerino $n$ unità statistiche variabili. $D$ = disp. finanziaria; $S$ = spese fisse t.c. $D > S$. $X$ = propensione all'investimento $\in (0,1)$.
$k$ = ammontare investimenti di ciascun soggetto.
Relazione tra le variabili: $k_i = (d_i - s_i) \cdot x_i$
Si determini la propensione media all'investimento che lascia invariato il tot. spese fisse.

$\bar{\bar{x}} = ?$ $\quad \sum_{i=1}^{r} s_i$ ← tot spese $= \text{costante} = \varphi(\cdot)$

Ma devo vedere la $x$ nella funzione invariante $\varphi$

$$\Rightarrow k_i = (d_i - s_i) x_i \Rightarrow s_i = d_i - \frac{k_i}{x_i} \Rightarrow \sum_{i=1}^{r} s_i = \sum_{i=1}^{r} \left(d_i - \frac{k_i}{x_i}\right)$$

$$\Rightarrow \sum_{i=1}^{r} \left(d_i - \frac{k_i}{x_i}\right) = \sum_{i=1}^{r} \left(d_i - \frac{k_i}{\bar{\bar{x}}}\right) \Rightarrow \sum_{i=1}^{r} d_i - \sum_{i=1}^{r} \frac{k_i}{x_i} = \sum_{i=1}^{r} d_i - \sum_{i=1}^{r} \frac{k_i}{\bar{\bar{x}}}$$

$$\Rightarrow \frac{1}{\bar{\bar{x}}} \sum_{i=1}^{r} k_i - \sum_{i=1}^{r} \frac{k_i}{x_i} = 0 \Rightarrow \frac{1}{\bar{\bar{x}}} \sum_{i=1}^{r} k_i = \sum_{i=1}^{r} \frac{k_i}{x_i} \Rightarrow \frac{1}{\bar{\bar{x}}} = \frac{\sum_{i=1}^{r} \frac{k_i}{x_i}}{\sum_{i=1}^{r} k_i}$$

$$\Rightarrow \bar{\bar{x}} = \frac{\sum_{i=1}^{r} k_i}{\sum_{i=1}^{r} \frac{k_i}{x_i}}$$

→ non è una media in senso stretto, tuttavia è riconducibile come media armonica con sistema di pesi $k_i$.

# Lezione 9 - 21/10/21

## ② Metodo della minimizzazione del danno

Danno = perdita di informazione

$X = x_1, x_2, \dots, x_n$ : osservazioni

Sintesi : $\alpha = \alpha(X) = \alpha(x_1, x_2, \dots, x_r ; n_1, n_2, \dots, n_r)$

$\alpha$ sintetizza la distribuzione di frequenza di $X$

Obiettivo : minimizzare il danno globale

ERRORE = $e_i = x_i - \alpha$ dove $\alpha$ è una media, ma bisogna definire quale.

es. Due studenti : A e B, entrambi media voti = 8

ma : - A studia tutti i giorni e prende sempre (e solo) 8

- B studia poco : prende 6 ma poi recupera con un 10

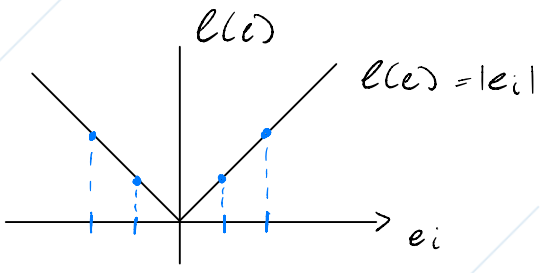
Media = 8 rappresenta (sintetizza) A con nessuna perdita di informazione, cosa che avviene invece con B.

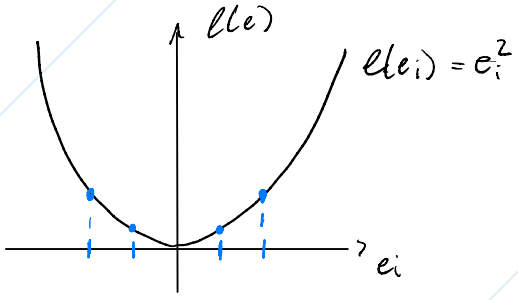
Per quantificare l'errore globale : somma di tutti gli $e_i$

Ma gli errori $e_i$ possono compensarsi

⇒ annulliamo la compensazione applicando : - potenze pari
- valore assoluto

$\ell(e)$ = funzione $\ell$ degli errori $e_i$ = ?




$\ell(e)$ deve essere :

- simmetrica
- continua
- minima (o meglio ancora nulla) nell'origine
- monotona decrescente per $e_i < 0$
- monotona crescente per $e_i > 0$

$\ell(e) = |e|^s$ $\quad s = 1, 2, \dots$ (vanno già bene $s=1$ e $s=2$)

**DANNO** = Perdita globale = $D$

$D = \sum_{i=1}^{r} \ell(e_i)\, n_i$ dove $e_i = x_i - a$

$a$ è una media, ancora non precisata

cerchiamo $a$ che minimizzi $D$

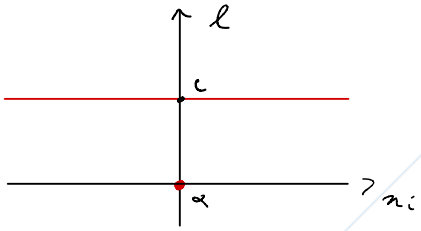
$\ell$ = funzione applicata agli errori

$\Rightarrow D = \sum_{i=1}^{r} |x_i - a|^s \cdot n_i$ $\quad$ **NB:** $s=0$ può servire per studiare $D$ al limite $s \to 0$

**$s = 0$** $\quad D = \sum_{i=1}^{r} |x_i - a|^0 \cdot n_i$ (messa così è inutile, è $= n$)

$\ell(e) = \ell(x_i - a)$ $\quad = 0 \iff x_i = a$

costante $\iff x_i \neq a$




Passaggio al limite:

caso ① : $x_i \neq a$

$\Rightarrow |x_i - a| > 0 \ \forall i$ $\quad \lim_{s \to 0} |x_i - a|^s \to 1$

$\Rightarrow D = \lim_{s \to 0} \sum_{i=1}^{r} |x_i - a|^s \cdot n_i = \sum_{i=1}^{r} n_i = n$

caso ② : $x_j = a$ , $x_i \neq a$ $\quad i = 1, 2, \dots, j-1, j+1, \dots, r$

$\Rightarrow \begin{cases} |x_i - a| > 0 & i \neq j \\ |x_i - a| = 0 & i = j \end{cases}$

Quindi $\sum_{i=1}^{r} |x_i - a|^s \cdot n_i = \sum_{i \neq j} |x_i - a|^s n_i + 0$ $\quad \sum_{i \neq j}$

$D = \sum_{i=j} \lim_{s \to 0} |x_i - a|^s \cdot n_i = n_1 + n_2 + \dots + n_{j-1} + n_{j+1} + \dots + n_r$

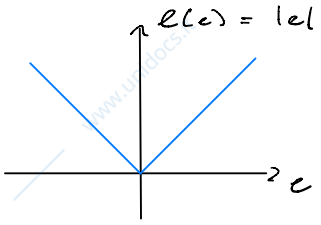
$= n - n_j < n$

$\Rightarrow D = n - n_j < n$ $\quad \leftarrow$ vogliamo minimizzarla

Per minimizzare $D$ bisogna massimizzare $n_j$

$\Rightarrow a = x_i =$ **MODA**

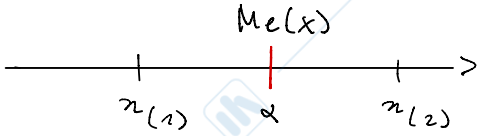
**s = 1** $\quad \ell(e) = |e|$




$$D = \sum_{i=1}^{r} |x_i - \alpha| n_i$$

$D_{min} \Rightarrow \alpha =$ MEDIANA

Ipotizziamo di avere solo 2 $x_i$ $(n = 2)$ :

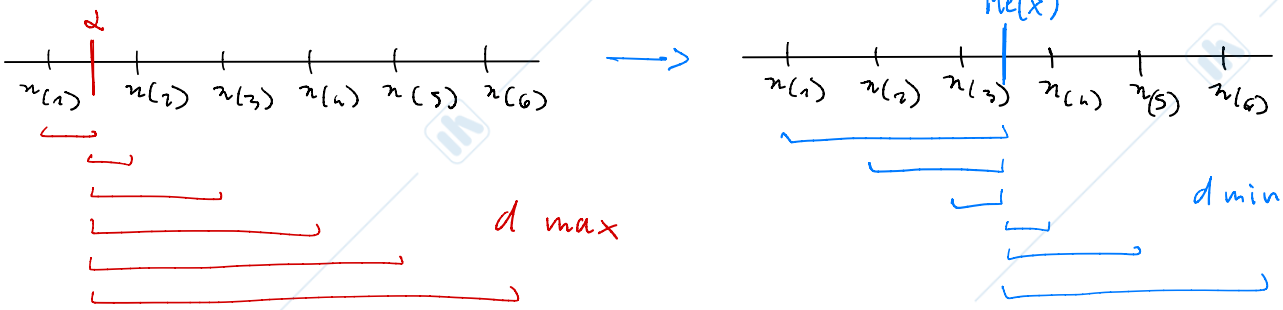



$\Rightarrow n = 2 \quad \{ x_{(1)} ; x_{(2)} \{ \quad x_1 < x_2 \quad d = x_{(2)} - x_{(1)}$

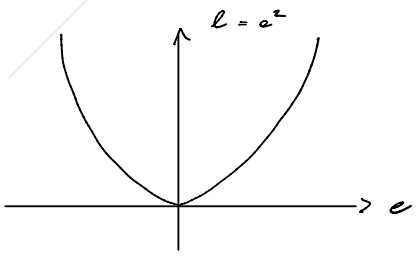
$$\sum_k |x_k - \alpha| = \begin{cases} 2|x_{(1)} - \alpha| + d & \alpha < x_{(1)} \\ 2|x_{(2)} - \alpha| + d & \alpha > x_{(2)} \\ d & x_{(1)} \le \alpha \le x_{(2)} \end{cases}$$

: condizione di minimo

Ipotizziamo $n = 6$




**s = 2** $\quad \ell(e) = e^2$




$$D = \sum_{i=1}^{r} (x_i - \alpha)^2 n_i$$

$$x_i^2 + \alpha^2 - 2\alpha x_i$$ : funzione di II grado di argomento $\alpha$

$$\frac{dD}{d\alpha} = -2 \sum_{i=1}^{r} (x_i - \alpha) n_i = 0 \qquad \sum_{i=1}^{r} (x_i - \alpha) n_i = 0$$

$\alpha$ = Media aritmetica $\mu$ (in quanto $\mu$ annulla la somma degli scarti)

$$\frac{d^2 D}{d\alpha^2} = 2 \sum_{i=1}^{r} n_i = 2n > 0$$

# VARIABILITÁ

: oscillazione intorno a un indice di posizione

Mediana : non risente di valori estremi (non considera il valore delle osservazioni bensí il loro ordine)

Media : risente della presenza di valori estremi

## Indici di variabilitá (o di dispersione)

Variabilitá : attitudine di un fenomeno ad assumere intensitá di valore differenti.

Nel caso di fenomeni qualitativi si parla di **mutabilitá** o eterogeneitá.

[ NOTA : non tratteremo indici di variabilitá globale ]

## Proprietá (sia per variabili quantitative che qualitative)

① $V(X) \geq 0$ [$V(X)$ = variabilitá di $X$ (per ora)]

② $V(X) = 0$ se $x_i$ sono costanti

③ $V(X + d) = V(X)$ : invarianza per traslazione

=> se aggiungo o sottraggo una stessa costante ad ogni $x_i$ non cambia la sua variabilitá.

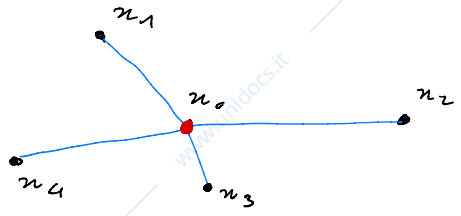
④ $|x_i - x_j| \leq |y_i - y_j| \quad \forall i,j \Rightarrow V(X) \leq V(Y)$

=> monotona crescente rispetto agli scarti in modulo

## Lezione 10 - 25/10/21

$d = |x_i - x_0| \quad i = 1, 2, \dots, n$

$x_0$ = CENTRO (media, moda, mediana) o POLO




Il modulo (o la potenza pari) serve per evitare la compensazione degli scarti positivi e negativi dal centro $x_0$.

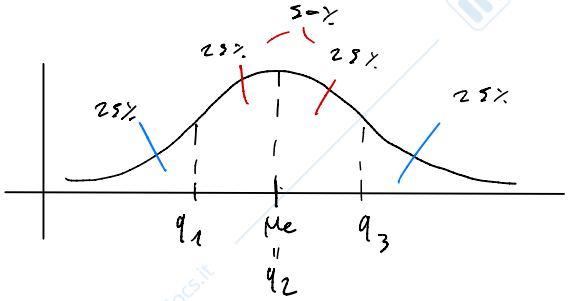
## Campo di variazione o Range

$$\text{Range} = x_{max} - x_{min}$$

È una misura "embrionale" dello scostamento da un centro (che infatti qui non è espresso)

## Differenza interquartile

$DQ = q_3 - q_1$ => considero il 50% centrale della distribuzione




Anche con la DQ non si considera la dispersione da un centro

Vediamo ora gli indici di variabilità rispetto a un centro.

$$d_i = |x_i - x_0| \quad i = 1, 2, \dots, n$$

## Scostamenti medi assoluti

$$x_0 S_r = \left( \frac{\sum_{i=1}^{n} |x_i - x_0|^r n_i}{n} \right)^{1/r} \qquad r > 0$$

$x_0 S_r$ è una media potenziata di ordine $r$ di $d_i$.

Come scelgo la media potenziata e l'indice di posizione $x_0$?

NB: oltre a rispettare le proprietà degli indici di variabilità, $x_0 S_r$ eredita anche le proprietà della media potenziata e dell'indice di posizione $x_0$ scelto.

- $x_0$ = Mediana $r = 1$ => SMA dalla Mediana

$$Me\,S_1 = \frac{1}{n} \sum_{i=1}^{n} |x_i - Me|\, n_i$$ (che è minima per proprietà della Mediana)

- $x_0$ = Media aritmetica $r = 1$ => SMA dalla Media

$$\mu S_1 = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|\, n_i$$

- $x_0$ = Media aritmetica $r = 2$ => SCARTO QUADRATICO MEDIO

$$\mu S_2 = \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 n_i}$$

$(\mu S_2)^2 = \sigma^2 =$ VARIANZA

## Proprietà

① $Me\,S_1 \leq n_0\,S_1$

⇒ la mediana minimizza lo scostamento rispetto a qualsiasi altro indice di posizione per l'ordine 1 (per proprietà della Mediana).

② $\mu S_2 \leq n_0 S_2$

└ Media aritmetica

③ $Me\,S_1 \leq Me\,S_2$ , $\mu S_1 \leq \mu S_2$ , ..

⇒ a parità di indice di posizione, ad un ordine maggiore di media potenziata impiegata corrisponde uno SMA maggiore.

① + ③ $Me\,S_1 \leq \mu S_2$

Particolare rilevanza ha la VARIANZA

operatore media

$$\sigma^2 = (SQM)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 n_i = M\left[(x-\mu)^2\right]$$

Formula Operativa per il calcolo della varianza

$$\sigma^2 = M(x^2) - \mu^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{n} - \mu^2$$

Dim.

$$\sigma^2 = M\left[(x-\mu)^2\right] = M\left[x^2 + \mu^2 - 2x\mu\right] = M(x^2) + \mu^2 - 2\mu M(x) =$$

$$= M(x^2) + \mu^2 - 2\mu^2 = M(x^2) - \mu^2 \quad \text{CVD}$$

Esempio : Formula operativa

| $x_i$ | $n_i$ | $x_i n_i$ | $x_i^2 n_i$ |
|---|---|---|---|
| 5 | 5 | 25 | 125 |
| 6 | 10 | 60 | 360 |
| 7 | 20 | 140 | 980 |
| 8 | 10 | 80 | 640 |
| 9 | 5 | 45 | 405 |
| | 50 | 350 | 2510 |

$\mu = 7$ $\quad M(x) = \frac{350}{50} = 7$ $\quad M(x^2) = \frac{2510}{50} = 50,2$

$\sigma^2 = 50,2 - 7^2 = 1,2$

$\sigma = \sqrt{1,2} = 1,095$

Metodo "classico" (con $\sigma^2 = \frac{1}{n}\sum_{i=1}^{m}(x_i-\mu)^2 n_i$)

| $x_i$ | $n_i$ | $x_i - \mu$ | $(x_i-\mu)^2 n_i$ |
|---|---|---|---|
| 5 | 5 | -2 | 20 |
| 6 | 10 | -1 | 10 |
| 7 | 20 | 0 | 0 |
| 8 | 10 | 1 | 10 |
| 9 | 5 | 2 | 20 |
| | 50 | 0 | 60 |

$$\sigma^2 = \frac{60}{50} = 1,2$$

NB: La varianza $\sigma^2$ restituisce una quantità espressa nell'unità di misura di partenza al quadrato, non utile nella pratica; per questo si usa lo SQM.

## Proprietà della varianza

Consideriamo l'operatore $Var(x)$ (o $V(x)$) e la costante $\lambda$.

① $V(\lambda) = 0$ $\Rightarrow$ $\sigma^2$ di una costante è ovviamente nulla

DIM. $M(\lambda^2) - [M(\lambda)]^2 = \lambda^2 - \lambda^2 = 0$

② $V(\lambda X) = \lambda^2 V(X)$ NB: $V(\lambda X)$ non è una traslazione, bensì un cambiamento di scala.

DIM. $V(\lambda X) \overset{DEF.}{=} M[\lambda X - M(\lambda X)]^2 = M[\lambda X - \lambda M(X)]^2 =$

$= \lambda^2 M[X - M(X)]^2 \overset{DEF.}{=} \lambda^2 V(X)$

③ $V(\lambda + X) = V(X)$ $\Rightarrow$ invariante per traslazione

DIM. $V(\lambda + X) = M[(\lambda+X) - M(\lambda+X)]^2 = M[(\lambda+X) - M(\lambda) - M(X)]^2 =$

$= M[\lambda + X - \lambda - M(X)]^2 = M[X - M(X)]^2 = V(X)$

④ $V(X+Y) = VAR(X) + VAR(Y) + 2M[(X-M(X))(Y-M(Y))]$

$\Rightarrow VAR(X+Y) = M[(X+Y) - M(X+Y)]^2 = M[(X+Y) - M(X) - M(Y)]^2 =$

$= M[(X-M(X)) + (Y-M(Y))]^2 = M[(X-M(X))^2] + M[Y-M(Y)]^2 +$

$+ 2M[(X-M(X))(Y-M(Y))] = VAR(X) + VAR(Y) + 2\,COV(XY)$

COVARIANZA

④ bis: se $Cov(X,Y) = 0$ $\Rightarrow$ $VAR(X+Y) = VAR(X) + VAR(Y)$

# Indici di variabilità relativa

## COEFFICIENTE DI VARIAZIONE

$CV = \frac{\sigma}{\mu}$ => Numero puro => adimensionale, indipendente da unità di misura

└ rapporto tra indice di dispersione e indice di posizione

NB: $CV = \frac{\sigma}{\mu}$ deve essere $\geq 0$ ; se $\mu < 0$ => $CV = \frac{\sigma}{|\mu|}$

# Indici di variabilità normalizzati

Indici normalizzati:

- numeri puri => privi di unità di misura
- campo di variazione $[0, 1]$

NB: È buona norma normalizzare solo con caratteri trasferibili.

Normalizzare un indice:

$$I_{min} \leq I \leq I_{max}$$

$$I_N = \frac{I - I_{min}}{I_{max} - I_{min}} \Rightarrow 0 \leq I_N \leq 1$$

$$\sigma^2_N = \frac{\sigma^2 - 0}{\sigma^2_{MAX} - 0} = \frac{\sigma^2}{\sigma^2_{max}} \Rightarrow 0 \leq \sigma^2_N \leq 1$$

$\sigma^2_{min} = 0$

Ma $\sigma^2_{max} = ?$

Intervallo di dispersione $[A, B]$

| $x_i$ | $n_i$ |
|---|---|
| A | $n_A$ |
| B | $n_B$ |
| | $n$ |

① $n = n_A + n_B$ => invarianza della popolazione

② $\mu = \frac{A n_A + B n_B}{n}$ => invarianza della media

$$\Rightarrow \begin{cases} n_A + n_B = n \\ \frac{A n_A + B n_B}{n} = \mu \end{cases} \quad \begin{matrix} n_A = ? \\ n_B = ? \end{matrix} \quad \begin{cases} n_A = n - n_B \\ n\mu = A(n - n_B) + B n_B \end{cases}$$

$$\begin{cases} n_A = n - n_B \\ n\mu = n_A - A n_B + B n_B \end{cases} \quad \begin{cases} - \\ n\mu = n_B(B - A) + n_A \end{cases}$$

$$\begin{cases} n_A = n - n_B \\ n_B = \dfrac{n\mu - nA}{B - A} \end{cases} \qquad n_A = n - \frac{n\mu - nA}{B-A} = \frac{n(B-A) - n(\mu - A)}{B-A} =$$

$$= \frac{n[B - A - \mu + A]}{B-A} = n\,\frac{B-\mu}{B-A}$$

$$n_A = n\,\frac{B-\mu}{B-A} \Rightarrow \quad \wp_A = \frac{B-\mu}{B-A}$$

$$\Rightarrow$$

$$n_B = n\,\frac{\mu - A}{B-A} \Rightarrow \quad \wp_B = \frac{\mu - A}{B-A} = 1 - \wp_A$$

Quindi:

| $x_i$ | $n_i$ |
|---|---|
| $A$ | $n\,\frac{B-\mu}{B-A}$ |
| $B$ | $\frac{n - n_A}{n}$ |

$$\sigma^2_{max} = M(x^2) - \mu^2 = \frac{A^2}{n}\, n\, \frac{B-\mu}{B-A} + \frac{B^2}{n}(n - n_A) - \mu^2 =$$

$$= A^2\,\frac{B-\mu}{B-A} + \frac{B^2}{n}\left(n - n\,\frac{B-\mu}{B-A}\right) - \mu^2 =$$

$$= A^2\,\frac{B-\mu}{B-A} + B^2\left(1 - \frac{B-\mu}{B-A}\right) - \mu^2 =$$

$$= A^2\,\frac{B-\mu}{(B-A)} + B^2 - B^2\,\frac{B-\mu}{B-A} - \mu^2 = \frac{B-\mu}{B-A}\left(A^2 - B^2\right) + B^2 - \mu^2$$

$$= \frac{B-\mu}{B-A}\,(A+B)(A-B) + \left(B^2 - \mu^2\right) = -(B-\mu)(A+B) + (B+\mu)(B-\mu)$$

$$= (B-\mu)(-A - B + B + \mu) = (B-\mu)(\mu - A) \quad \text{CVD}$$

Consideriamo ora l'insieme classico $[0, T]$ $\quad T = \sum_{i=1}^{r} x_i\, n_i$

| $x_i$ | $n_i$ |
|---|---|
| $0$ | $n-1$ |
| $T$ | $1$ |
| | $n$ |

Infatti:

| $x_i$ | $n_i$ |
|---|---|
| $0$ | $n\,\frac{B-\mu}{B-A} = n\,\frac{(n\mu - \mu)}{n\mu - 0} = n\,\frac{\mu(n-1)}{n\mu}$ |
| $T$ | $n - n_A = 1 \qquad = n - 1$ |
| | $n$ |

$$\sigma^2_{max} = (B-\mu)(\mu - A) = (n\mu - \mu)(\mu - 0) = \mu(n-1)\mu$$

$$= \mu^2(n-1)$$

Esempio

| $x_i$ | $n_i$ |
|---|---|
| 6 | 3 |
| 7 | 8 |
| 8 | 6 |
| 9 | 4 |
| 10 | 19 |
| | 40 |

$[A, B] = [6, 10]$

$n_A = n \frac{B - \mu}{B - A} = 13$

$n_B = 40 - 13 = 27$

$\mu = 8,7 \qquad \sigma^2 = 2,01$

$\sigma^2_{max} = (B - \mu)(\mu - A) = (10 - 8,7)(8,7 - 6) = 3,51$

$\sigma^2_p = \frac{\sigma^2}{\sigma^2_{max}} = \frac{2,01}{3,51} = 0,5726$

## Lezione 12 - 28/10/21

**Mutabilità** = variabilità di variabili statistiche qualitative (= mutabili)

### Indici di variabilità per fenomeni qualitativi

① Mutabilità nulla = una sola modalità assume la totalità delle frequenze

$\Rightarrow \quad \exists\, n_i = n, \; n_j = 0 \; \forall j \neq i$

② Mutabilità massima = frequenze equidistribuite tra le modalità

$k$ = numero delle modalità $\qquad n_1 = n_2 = \dots = n_k = \frac{n}{k}$

NB: è importante definire i due "paletti" di min e max di un indice in quanto sono essenziale per normalizzarlo.

### Indice di Gini $E(X)$ (o di eterogeneità)

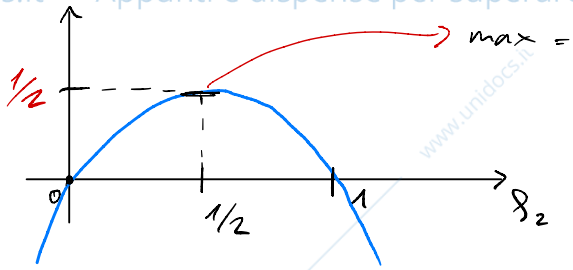
Ipotizziamo $X$ dicotomica $[0; 1]$

| $x_i$ | $n_i$ | $g_i$ |
|---|---|---|
| 0 | $n_1$ | $g_1$ |
| 1 | $n_2$ | $g_2 = 1 - g_1$ |
| | $n$ | 1 |

$\sigma^2_x = M(x^2) - \mu^2 = (0^2 \cdot g_1 + 1^2 \cdot g_2) - (0 \cdot g_1 + 1 \cdot g_2)^2$

$= \underbrace{g_2 - g_2^2}_{\text{parabola}} = g_2(1 - g_2)$

$y = p_2 - p_2^2$




Ipotizziamo k modalità

$$E(x) = \sum_{i=1}^{k} p_i (1 - p_i) = 1 - \sum_{i=1}^{k} p_i^2$$

dim. $E(x) = \sum_{i=1}^{k} p_i (1 - p_i) = \sum_{i=1}^{k} p_i - p_i^2 = \sum_{i=1}^{k} p_i - \sum_{i=1}^{k} p_i^2 =$

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

## Proprietà del Gini

① $\min E(x) = 0$

es. $p_1 = 1 \quad p_2 = p_3 = \dots = p_k = 0$ (minima eterogeneità)

$E(x) = 1 - \left( p_1^2 - \sum_{i=2}^{k} p_i^2 \right) = 0$ ✓ rispetta la condizione ①

($p_1^2 = 1$, $\sum_{i=2}^{k} p_i^2 = 0$)

② $\max E(x)$

$p_1 = p_2 = \dots = p_k = \frac{1}{k}$ $\qquad E(x) = 1 - \sum_{i=1}^{k} \left(\frac{1}{k}\right)^2 = 1 - k\left(\frac{1}{k}\right)^2 = 1 - \frac{1}{k}$

## Indice di Gini normalizzato

$$E_N(x) = \frac{E(x) - E_{min}(x)}{E_{max}(x) - E_{min}(x)} = \frac{E(x)}{E_{max}(x)}$$

($E_{min}(x) = 0$, $E_{min}(x) = 0$)

Ovviamente: $0 \leq E_N(x) \leq 1$

$$\Rightarrow E_N(x) = \frac{1 - \sum_{i=1}^{k} p_i^2}{1 - 1/k}$$

Esempio n° addetti x settore

| $x_i$ | $h_i$ | $f_i$ | $f_i^2$ |
|---|---|---|---|
| AGR | 120 | 0,0311 | 0,0010 |
| IND | 1673 | 0,4334 | 0,1878 |
| TER | 2067 | 0,5355 | 0,2868 |
| | 3860 | 1 | 0,4756 |

$E(x) = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0,4756 = 0,5244$

$E_{max}(x) = 1 - \frac{1}{k} = 1 - \frac{1}{3} = 0,\overline{6}$

$E_N(x) = \frac{E(x)}{E_{max}(x)} = \frac{0,5244}{0,\overline{6}} = 0,7866$

=> il Gini è vicino al suo max
=> forte eterogeneità

=> forte disomogeneità tra i settori

## Momenti di una distribuzione

Momento r-esimo di X da $\alpha$ = media aritmetica delle potenze r-esime degli scarti da $\alpha$

$$_{\alpha}\mu_r = M\left[(x_i - \alpha)^r\right] \quad h_i = 1 \text{ oppure } \neq 1$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^r$$ se le frequenze sono unitarie

$$= \frac{1}{n}\sum_{i=1}^{k}(x_i - \alpha)^r h_i$$ se le frequenze non sono unitarie ma si ripropongono su più unità statistiche

**$\alpha = 0$** Momenti dall'origine

$$_{0}\mu_r \equiv \mu_r = M\left[(x-0)^r\right] = M\left[x^r\right] = \frac{1}{n}\sum_{i=1}^{r} x_i^r n_i$$

Notare che:
- $r = 1 \Rightarrow \mu_1$ = Media aritmetica
- $r = 2 \Rightarrow \mu_2 = M(x^2)$

**$\alpha = \mu$** Momenti dalla media o Momenti CENTRALI

$$\bar{\mu}_r = \frac{1}{n}\sum_{i=1}^{k}(x_i - \mu)^r n_i$$

Notare che:
- $r = 1 \Rightarrow \bar{\mu}_1 = 0$ (Proprietà media aritmetica)
- $r = 2 \Rightarrow \bar{\mu}_2$ = Varianza

Quindi: $\mu_1 = M(x) = \mu$

$\mu_2 = M(x^2) = \frac{1}{n}\sum_{i=n}^{k} x_i^2 n_i$

$\bar{\mu}_1 = M[(x-\mu)] = 0$ ($\mu$ è $\mu_1$)

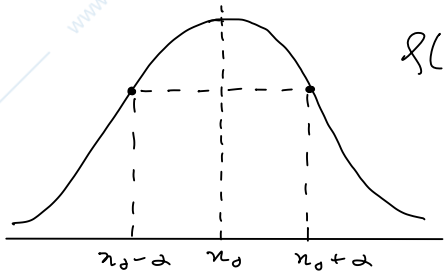
$\bar{\mu}_2 = M[(x_i-\mu)^2] = \sigma^2 = M(x^2) - [M(x)]^2$ ($M(x^2)$ è $\mu_2$)

NB: I momenti sono tra loro legati (infatti, ad es., $\bar{\mu}_1$ è funzione di $\mu_1$)

## Indicatori di forma

### Asimmetria

$x = x_0$ , $a > 0$



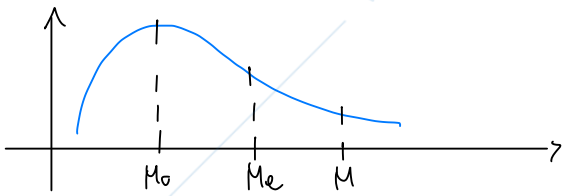

$f(x_0 - a) = f(x_0 + a)$ => $f$ simmetrica rispetto a $x_0$

Vogliamo: - Funzione continua
- Distr. unimodale (un solo punto di "picco")

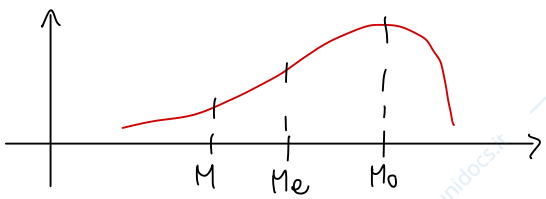
Se la distribuzione è simmetrica => $M(x) = Me(x) = Mo(x)$

Asimmetria

Positiva => $Mo(x) < Me(x) < M(x)$




Negativa => $Mo(x) > Me(x) > M(x)$



# Indice di (a)simmetria di Fisher $\gamma_1$

In una distribuzione simmetrica, i momenti centrali di ordine dispari sono nulli.

$\bar{\mu}_{2r+1} = 0 \qquad r = 0, 1, 2, \dots$

↳ $\bar{\mu}_3 = \bar{\mu}_5 = \dots = 0$

Per verificarlo, dovremmo verificare <u>tutti</u> i momenti centrali di ordine dispari, che sono $\infty$.

⇒ $\gamma_1$ dice solo se una distribuzione è asimmetrica (se trovo un m.c. di ordine dispari $\neq 0$)

NB: Simmetria ⇒ $Me = Mo = M$

<u>MA</u> $Me = Mo = M$ ⇏ Simmetria

L'<u>unico</u> modo per verificare la simmetria è per via grafica.

Tornando all'asimmetria:

- Asimmetria positiva: prevalgono gli scarti positivi ⇒ $\bar{\mu}_{2r+1} > 0$
- Asimmetria negativa: prevalgono gli scarti negativi ⇒ $\bar{\mu}_{2r+1} < 0$

$$\gamma_1 = \frac{\bar{\mu}_3}{\sigma^3}$$

$\gamma_1$ è adimensionale

$\gamma_1 > 0 \Rightarrow \bar{\mu}_3 > 0 \Rightarrow$ asimmetria positiva

$\gamma_1 < 0 \Rightarrow \bar{\mu}_3 < 0 \Rightarrow$ asimmetria negativa

$\gamma_1 = 0$ condizione necessaria ma non sufficiente per la simmetria

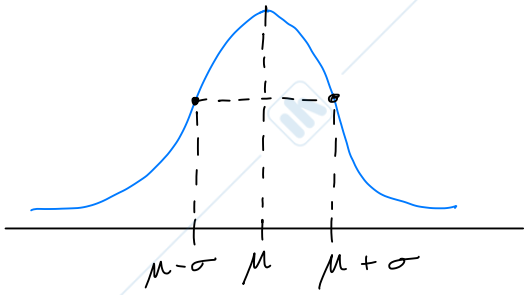
Quindi $\gamma_1$ è un <u>SINTOMO</u> di simmetria.

Lezione 13 - 2/11/21

## Indice di Curtosi

CURTOSI = grado di "appiattimento" di una curva (rispetto alla gaussiana)

Gaussiana:




- variabile statistica continua
- unimodale (un solo punto di massimo, in $\mu = Me = Mo$)
- Area sottesa dalla curva:

$$\varphi(x) = \frac{1}{\sqrt{2\pi b}} \exp\left\{-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right\}$$

dove $a = M(x) = \mu$, $b = V(x) = \sigma^2$ e $\int_{-\infty}^{+\infty} \varphi(x)\,dx = 1$

- Simmetrica, con asse di simmetria $x = \mu$
- $\bar{\mu}_3 = \bar{\mu}_5 = \dots = 0$ e $\bar{\mu}_2 = \sigma^2$, $\bar{\mu}_4 = 3\sigma^4$
- $\gamma_1 = 0$
- $\gamma_2 = \frac{\bar{\mu}_4}{\sigma^4} = \frac{3\sigma^4}{\sigma^4} = 3$

=> Se $\gamma_2 = 3$, la curva in questione ha la forma di una Normale.

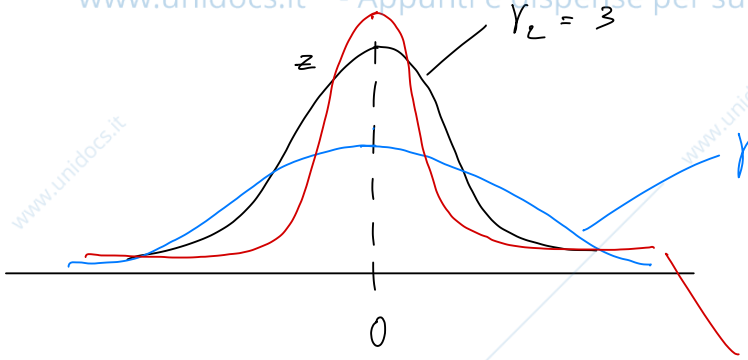
Ci interessa studiare $\gamma_2$:

- $< 3$ Platicurtica
- $= 3$ Normocurtica
- $> 3$ Leptocurtica

Standardizzare una variabile statistica:

$$z = \frac{x-\mu}{\sigma}$$

$z$ ha $\mu = 0$, $\sigma^2 = 1$




# STATISTICA DESCRITTIVA BIVARIATA

: rilevazione di due caratteri per ogni unità statistica (tipicamente mettendoli in relazione)

Distribuzione CONGIUNTA di frequenze assolute dei caratteri X, Y in una tabella a doppia entrata:

| $y_j \backslash x_i$ | $x_1$ | $x_2$ | ... | $x_i$ | ... | $x_k$ | $n_{\cdot j}$ |
|---|---|---|---|---|---|---|---|
| $y_1$ | $n_{11}$ | $n_{21}$ | | $n_{i1}$ | | $n_{k1}$ | $n_{\cdot 1}$ |
| $y_2$ | $n_{12}$ | $n_{22}$ | | $n_{i2}$ | | $n_{k2}$ | $n_{\cdot 2}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ | ⋮ |
| $y_j$ | $n_{1j}$ | $n_{2j}$ | | $n_{ij}$ | | $n_{kj}$ | $n_{\cdot j}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ | ⋮ |
| $y_m$ | $n_{1m}$ | $n_{2m}$ | | $n_{im}$ | | $n_{km}$ | $n_{\cdot m}$ |
| $n_{i\cdot}$ | $n_{1\cdot}$ | $n_{2\cdot}$ | | $n_{i\cdot}$ | | $n_{k\cdot}$ | $n$ |

$n$: numerosità totale

Le frequenze relative si ottengono: $f_{ij} = \frac{n_{ij}}{n}$

(al posto di $n$ c'è 1)

Dalla tabella si possono ricavare diverse distribuzioni univariate:

- Distribuzioni marginali :
  - rispetto a Y
  - rispetto a X
- Distribuzioni condizionale :
  - Y condizionato $x_1$
  - X condizionato $y_2$

oss: le condizionate permettono di evidenziare come varia X dato un certo valore di Y.

Distribuzioni marginali

X

| $x_i$ | $n_{i\bullet}$ |
|---|---|
| $x_1$ | $n_{1\bullet}$ |
| $x_2$ | $n_{2\bullet}$ |
| ⋮ | ⋮ |
| $x_r$ | $n_{r\bullet}$ |
| | $n$ |

Y

| $y_1$ | $n_{\bullet j}$ |
|---|---|
| $y_1$ | $n_{\bullet 1}$ |
| $y_2$ | $n_{\bullet 2}$ |
| ⋮ | ⋮ |
| $y_m$ | $n_{\bullet m}$ |
| | $n$ |

Quante unità statistiche presentano una certa modalità di Y indipendentemente da come si comporta X

## Variabili statistiche condizionate

$X|y_j$ o $X|Y=y_j$ in matematica: $x = g(y)$

=> studio come si comporta X (= come varia X) data una certa modalità di Y fissata.

| $x_i \mid y_J$ | $n_{iJ}$ |
|---|---|
| $x_1$ | $n_{1J}$ |
| $x_2$ | $n_{2J}$ |
| ⋮ | ⋮ |
| $x_r$ | $n_{rJ}$ |
| | $n_{\bullet J}$ → totale di riga |

Fisso la $y_j$ e scorro tra le colonne $x_i$

$J = 1, 2, \dots, m$ => Ci sono m possibili distribuzioni univariate di questo tipo $(X|Y=y_j)$

$Y|x_i$

| $Y \mid x_i$ | $n_{ij}$ |
|---|---|
| $y_1$ | $n_{i1}$ |
| $y_2$ | $n_{i2}$ |
| ⋮ | ⋮ |
| $y_m$ | $n_{i\bullet}$ → totale di colonna |

Fisso la colonna $x_i$ e scorro tra le righe $y_j$

$i = 1, 2, \dots, r$

Quindi:

① Distribuzione marginale di $X$

$$n_{i\cdot} = \sum_{j=1}^{n} n_{ij} \qquad \sum_{i=1}^{r} n_{i\cdot} = n$$

└ somma di tutte le r freq. assolute marginali

② Distribuzione marginale di $Y$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \qquad \sum_{j=1}^{m} n_{\cdot j} = n$$

③ Distribuzione condizionata $X|Y=y_j$

$(x_i \,;\, n_{ij})$ con $j$ fissato

④ Distribuzione condizionata $Y|X=x_i$

$(y_j \,;\, n_{ij})$ con $i$ fissato

In termini di Frequenze relative:

$$f_{ij} = \frac{n_{ij}}{n} \qquad f_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad f_{\cdot j} = \frac{n_{\cdot j}}{n}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{r} f_{i\cdot} = \sum_{j=1}^{m} f_{\cdot j} = 1$$

$$f_{y_j|x_i} = \frac{n_{ij}}{n_{i\cdot}} \qquad f_{x_i|y_j} = \frac{n_{ij}}{n_{\cdot j}}$$

$$\sum_{j=1}^{m} f_{y_j|x_i} = \sum_{j=1}^{m} \frac{n_{ij}}{n_{i\cdot}} = 1 \qquad \sum_{i=1}^{r} f_{x_i|y_j} = \sum_{i=1}^{r} \frac{n_{ij}}{n_{\cdot j}} = 1$$

Esempio

| FATT. \ CAT. | DET | G.D. | GROSS. | |
|---|---|---|---|---|
| 0 — 10 | 3 | 0 | 1 | 4 |
| 10 — 20 | 3 | 0 | 5 | 8 |
| 20 — 30 | 0 | 2 | 2 | 4 |
| 30 — 40 | 0 | 2 | 1 | 3 |
| 40 — 50 | 0 | 0 | 1 | 1 |
| | 6 | 4 | 10 | 20 |

| $x_i$ | $n_{i\cdot}$ | $f_{i\cdot}$ | $n_{i3}$ | $f_{i3} = n_{i3}/n_{\cdot 3}$ |
|---|---|---|---|---|
| DET | 6 | $6/20 = 0{,}3$ | 0 | 0 |
| G.D. | 4 | $4/20 = 0{,}2$ | 2 | $2/4 = 0{,}5$ |
| GR. | 10 | $10/20 = 0{,}5$ | 2 | $2/4 = 0{,}5$ |
| | 20 | 1 | 4 | 1 |

# Connessione e Indipendenza stocastica

es. $Y$ = professione paterna  $X$ = scolarità dei figli

| Y\X | L.M. | DIP. | LAUREA | $n_{\cdot j}$ |
|---|---|---|---|---|
| OP. | 3 | 2 | 5 | 10 |
| IMP. | 9 | 6 | 15 | 30 |
| DIR. | 18 | 12 | 30 | 60 |
| $n_{i\cdot}$ | 30 | 20 | 50 | 100 |

$Y|X = x_i$ (professione paterna condizionata da istruzione dei figli?)

| $y_j$ | $n_{ij}/n_{1\cdot}$ |
|---|---|
| OP. | 0,1 |
| IMP. | 0,3 |
| DIR. | 0,6 |
| | 1 |

=

| $y_j$ | $n_{ij}/n_{2\cdot}$ |
|---|---|
| OP. | 0,1 |
| IMP. | 0,3 |
| DIR. | 0,6 |
| | 1 |

=

| $y_j$ | $n_{ij}/n_{3\cdot}$ |
|---|---|
| OP. | 0,1 |
| IMP. | 0,3 |
| DIR. | 0,6 |
| | 1 |

le f. relative condizionate sono uguali indipendentemente dalla $x_i$ che scelgo come condizionante.

$X|Y = y_j$

| $x_i$ | $n_{ij}/n_{\cdot 1}$ |
|---|---|
| L.M. | 0,3 |
| DIP. | 0,2 |
| LAUREA | 0,5 |
| | 1 |

=

| $x_i$ | $n_{ij}/n_{\cdot 2}$ |
|---|---|
| L.M. | 0,3 |
| DIP. | 0,2 |
| LAUREA | 0,5 |
| | 1 |

=

| $x_i$ | $n_{ij}/n_{\cdot 3}$ |
|---|---|
| L.M. | 0,3 |
| DIP. | 0,2 |
| LAUREA | 0,5 |
| | 1 |

Idem come prima.

**TEO.** (Indipendenza stocastica)

Sia $(X, Y)$ una variabile (o mutabile) statistica doppia.[1]

Si dice che le componenti $X$ e $Y$ sono stocasticamente indipendenti se le variabili condizionate hanno la stessa distribuzione percentuale.

1 Lavorando con le frequenze e non con le modalità, l'indipendenza stocastica vale sia per le variabili che per le mutabili.

UB: Indipendenza stocastica $\Rightarrow$ assenza di legame funzionale.

Quindi:

ⓐ $X|Y=y_j \Rightarrow \begin{matrix} x_1 \\ \vdots \\ x_r \end{matrix} \quad \begin{matrix} n_{1j}/n_{\cdot j} \\ \vdots \\ n_{rj}/n_{\cdot j} \end{matrix} \quad \begin{matrix} = H_1 \\ \vdots \\ = H_r \end{matrix}$ $\forall j$, fissandolo ogni volta

$H$ è indipendente da $j$, $\forall j$

ⓑ $Y|X=x_i \Rightarrow \begin{matrix} y_1 \\ \vdots \\ y_m \end{matrix} \quad \begin{matrix} n_{i1}/n_{i\cdot} \\ \vdots \\ n_{im}/n_{i\cdot} \end{matrix} \quad \begin{matrix} = k_1 \\ \\ = k_m \end{matrix}$ $\forall i$ fissato

$k$ è indipendente da $i$

**Teo.**

(i) Condizione necessaria e sufficiente per l'indipendenza stocastica è la fattorizzazione delle frequenze congiunte.

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \qquad \forall i,j$$

(⇒ il prodotto delle marginali diviso per la numerosità totale deve dare la frequenza congiunta)

(ii) L'indipendenza stocastica è <u>simmetrica</u>.

$X$ è indip. stoc. da $Y$ $\iff$ $Y$ è indip. stoc. da $X$

(iii) Le distribuzioni condizionate sono uguali alle marginali percentuali in caso di indipendenza stocastica.

$$H_i = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{n} \leftarrow ⓐ \qquad k_j = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n} \leftarrow ⓑ$$

**DIM.** $(Y|X)$

$$\frac{n_{ij}}{n_{i\cdot}} = k_j \ \forall i \Rightarrow n_{ij} = k_j \cdot n_{i\cdot} \Rightarrow \sum_{i=1}^{r} n_{ij} = \sum_{i=1}^{r} k_j \cdot n_{i\cdot}$$

$$\Rightarrow n_{\cdot j} = \sum_{i=1}^{r} k_j \cdot n_{i\cdot} \Rightarrow n_{\cdot j} = k_j \sum_{i=1}^{r} n_{i\cdot} \Rightarrow n_{\cdot j} = k_j \cdot n$$

$$\Rightarrow k_j = \frac{n_{\cdot j}}{n} \quad \text{(DIM iii)}$$

$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n} \Rightarrow n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \quad \text{(DIM i)}$$