Corso di Statistica
Facoltà di Economia

a.a. 2010-2011

francesco mola

Lezione n°4

Sommario

- Campo di variazione
- Varianza
- Scarto Quadratico medio
- Coefficiente di variazione
- Scostamenti dalla Media e dalla Mediana
- Mutua Variabilità
- Mutabilità

Generalità sulla variabilità

- Concetto di variabilità
- Importanza della variabilità
- Uso congiunto di indici di posizione ed indici di variabilità
- Variabilità assoluta e relativa

Variabilità e Dispersione

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

$$\begin{cases} A & 18 & 24 & 30 \\ B & 23 & 24 & 25 \\ C & 24 & 24 & 24 \end{cases}$$

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

Variabilità e Dispersione (cont.)

Possiamo dire che i tre studenti hanno un stesso comportamento agli esami?

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

Caratteristiche degli indici di variabilità

Un indice di variabilità gode delle seguenti caratteristiche:

$$IV(x_1, x_2, \cdots, x_n) \geq 0$$

$$IV(x_1+c, x_2+c, \cdots, x_n+c) = IV(x_1, x_2, \cdots, x_n)$$

$$IV(c, c, \cdots, c) = 0$$

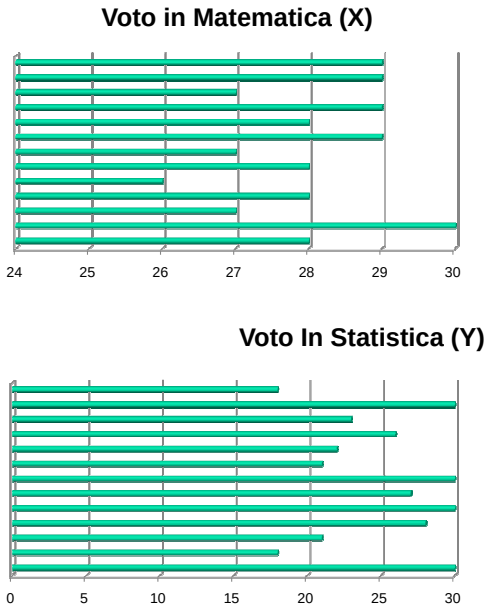
$$IV(x_1, x_2, \cdots, x_n) > IV(y_1, y_2, \cdots, y_n) \Rightarrow$$

X più variabile di Y

Un esempio su tredici studenti …

Studenti	Voto in Matematica (X)	Voto In Statistica (Y)	X-mx	(X-mx)^2	Y-my	(Y-my)^2
	28	30	-0,08	0,01	5,08	25,78
	30	18	1,92	3,70	-6,92	47,93
	27	21	-1,08	1,16	-3,92	15,39
	28	28	-0,08	0,01	3,08	9,47
	26	30	-2,08	4,31	5,08	25,78
	28	27	-0,08	0,01	2,08	4,31
	27	30	-1,08	1,16	5,08	25,78
	29	21	0,92	0,85	-3,92	15,39
	28	22	-0,08	0,01	-2,92	8,54
	29	26	0,92	0,85	1,08	1,16
	27	23	-1,08	1,16	-1,92	3,70
	29	30	0,92	0,85	5,08	25,78
	29	18	0,92	0,85	-6,92	47,93
	365	324	0,00	14,92	0,00	256,92

Le medie sono rispettivamente 28,08 e 24,92 e le mediane 28 e 26



Campo di variazione

$$V = \max(X) - \min(X)$$

E' un indice di variabilità assoluta

Per il nostro esempio abbiamo:

V(Matematica)=12 (teorica) =4(empirica)

V(Statistica)=12 (teorica) =12 (empirica)

Varianza

$$\sigma^2 = \frac{1}{n}\sum (x_i - \mu)^2$$

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

E' un indice di variabilità assoluta

Caratteristiche principali

- È una media
- Vale sempre che: $0 \le \sigma^2 \le \infty$

Per il nostro esempio abbiamo:

$$\sigma_X^2 = 1{,}15$$

$$\sigma_Y^2 = 19{,}15$$

Il massimo della varianza

Si dimostra che il massimo valore che la varianza può assumere (per quella particolare distribuzione empirica) è:

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

La varianza può essere anche vista come ….

$$\sigma^2 = \frac{1}{n}\sum (x_i - \mu)^2 = \frac{\sum x_i^2}{n} - \mu^2$$

dim.

$$\sigma^2 = \frac{1}{n}\sum (x_i - \mu)^2 = \frac{1}{n}\left[\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right] =$$

$$\frac{1}{n}\sum x_i^2 - 2\mu\frac{\sum x_i}{n} + \frac{n}{n}\mu^2 = \frac{1}{n}\sum x_i^2 - 2\mu^2 + \mu^2 =$$

$$\frac{1}{n}\sum x_i^2 - \mu^2 = \mu_2 - \mu^2$$

$$= \mu_2$$

Scarto Quadratico Medio

$$\sigma = \sqrt{\frac{1}{n}\sum (x_i - \mu)^2} = \sqrt{\sigma^2}$$

E' un indice di variabilità assoluta

Per il nostro esempio abbiamo:

$$\sigma_X = 1,07$$

$$\sigma_Y = 4,45$$

Caratteristiche principali

- È una media quadratica
- Vale sempre che:

$$0 \leq \sigma \leq \infty$$

Perché è utile lo sqm

Il problema principale della varianza è che è espressa nell'unità di misura del fenomeno al quadrato!!!!

Lo scarto quadratico medio risolve questo problema!!!!

Coefficiente di Variazione

$$CV = \frac{\sigma}{|\mu|} \quad \text{con } \mu \neq 0$$

E' un indice di variabilità relativa

Per il nostro esempio abbiamo:

$$CV_X = 0,04$$

$$CV_Y = 0,18$$

Determiniamo il massimo del coefficiente di variazione

Sappiamo che:

$$0 \leq \sigma^2 \leq \mu^2(n-1) \Rightarrow 0 \leq \sigma \leq \mu\sqrt{n-1}$$

$$0 \leq \frac{\sigma}{\mu} \leq \sqrt{n-1}$$

è il massimo

Coefficiente di Variazione normalizzato

$$CV_n = \frac{CV}{\sqrt{n-1}} \quad \text{con } 0 \leq CV_n \leq 1$$

E' un indice normalizzato

Per il nostro esempio abbiamo:

$$CVn_X = 0,01$$

$$CVn_Y = 0,05$$

Proprietà della varianza

Consideriamo una variabile X e consideriamo la seguente combinazione lineare: $Y = \beta X$

abbiamo che: $\sigma_y^2 = \beta^2 \sigma_x^2$

Consideriamo una variabile X e consideriamo la seguente combinazione lineare: $Y = \beta X + c$

abbiamo che: $\sigma_y^2 = \beta^2 \sigma_x^2$

Altri indici di variabilità

Scostamento Semplice Medio

$$S_\mu = \frac{1}{n}\sum |x_i - \mu|$$

Per il nostro esempio abbiamo:

$S\mu_X = 0,86$

$S\mu_Y = 4,08$

Scostamento Semplice Mediano

$$S_{Me} = \frac{1}{n}\sum |x_i - Me|$$

Per il nostro esempio abbiamo:

$S_{Me\ X} = 0,85$

$S_{Me\ Y} = 1,15$

Altri indici di variabilità

Differenza Interquartile

$$IQR = Q_3 - Q_1$$

con

$Q_1 = F(X) = 0,25$ $Q_3 = F(X) = 0,75$

Per il nostro esempio abbiamo:

$Q_{1X} = 27$ $Q_{3X} = 29$

$Q_3 - Q_1 = 2$

$Q_{1Y} = 21$ $Q_{3Y} = 30$

$Q_3 - Q_1 = 9$

Mutua Variabilità

- Cosa si intende per mutua variabilità?
- Cosa si intende per variabilità rispetto ad un centro
- Quali sono le differenze e le implicazioni?

Differenze Medie

Consideriamo una variabile $X=x_1,x_2, x_3$ con 3 modalità e proviamo ad individuare tutte le possibili differenze che possiamo costruire, cioè tutte le

$$d_{i,j} = |X_i - X_j| \qquad i, j = 1,2,3$$

Abbiamo:

$$\begin{array}{lll} |x_1 - x_1| & |x_1 - x_2| & |x_1 - x_3| \\ |x_2 - x_1| & |x_2 - x_2| & |x_2 - x_3| \\ |x_3 - x_1| & |x_3 - x_2| & |x_3 - x_3| \end{array}$$

Differenze Medie

È evidente che 3 di queste differenze sono pari a 0

$$\begin{matrix} 0 & |x_1 - x_2| & |x_1 - x_3| \\ |x_2 - x_1| & 0 & |x_2 - x_3| \\ |x_3 - x_1| & |x_3 - x_2| & 0 \end{matrix}$$

Abbiamo:

$$\Delta = \frac{\sum_{i \neq j=1}^{3} |x_i - x_j|}{3(3-1)}$$

Abbiamo un indice di variabilità che tiene conto della mutua variabilità

In generale abbiamo….

Differenze Medie Semplici

$$\Delta = \frac{\sum_{i \neq j=1}^{n} |x_i - x_j|}{n(n-1)}$$

n di queste differenze sono pari a zero e non vengono considerate al denominatore

Differenze Medie Semplici per distribuzioni di frequenza

$$\Delta = \frac{\sum_{i \neq j=1}^{k} |x_i - x_j| n_i n_j}{n(n-1)}$$

Differenze Medie con ripetizione

Differenze Medie con Ripetizione

$$\Delta' = \frac{\sum_{i \neq j=1}^{n} |x_i - x_j|}{n^2}$$

n di queste differenze sono pari a zero e non influenzano il numeratore ma solo il denominatore

Differenze Medie con Ripetizione

$$\Delta' = \frac{\sum_{i \neq j=1}^{k} |x_i - x_j| n_i n_j}{n^2}$$

Una relazione interessante

Si noti che vale la seguente relazione

$$\Delta = \Delta' \frac{n}{n-1}$$

e che…

$$Max(\Delta) = 2\mu \Rightarrow 0 \leq \frac{\Delta}{2\mu} \leq 1$$

Mutabilità per dati qualitativi

- Mutabilità
- Dispersione
- Eterogeneità
- Può anche essere applicato per dati quantitativi operando unicamente sulle frequenze!!

…

Consideriamo una mutabile $X=x_1,x_2,\ldots,x_k$ con k modalità e $n_1,n_2,\ldots,n_k$ frequenze associate a ciascuna modalità. Definiamo poi la seguente funzione:

$$d(x_i,x_j)=\begin{cases}1 & i\neq j\\ 0 & i=j\end{cases}$$

Ricorrendo alle differenze medie con ripetizione viste in precedenza possiamo definire il seguente indice:

L'indice del Gini

$$g=\frac{1}{n^2}\sum_i\sum_j d(x_i,x_j)n_in_j=$$

$$=\frac{1}{n^2}[(0)n_1n_1+(1)n_1n_2+\cdots+(1)n_1n_k+$$

$$+(1)n_2n_1+(0)n_2n_2+\cdots+(0)n_kn_k]=$$

$$=\frac{1}{n^2}[n_1(n-n_1)+n_2(n-n_2)+\cdots+n_k(n-n_k)]=$$

L'indice del Gini

$$=\frac{1}{n^2}\sum_{i=1}^{k}n_i(n-n_i)=\sum_{i=1}^{k}f_i(1-f_i)=$$

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

Indice di Eterogeneità del Gini

Il massimo di g

$$g_{\max} \equiv f_i = \frac{1}{k} \qquad \forall i = 1,2,\cdots,k$$

$$\Rightarrow g_{\max} = 1 - \sum_{i=1}^{k} \frac{1}{k^2} = 1 - \frac{k}{k^2} =$$

$$= 1 - \frac{1}{k} = \max$$

Normalizziamo g

$$g^* = \frac{g}{g_{\max}} = \frac{1 - \sum f_i^2}{1 - \frac{1}{k}} =$$

$$\Rightarrow g_{\max} = 1 - \sum_{i=1}^{k} \frac{1}{k^2} = 1 - \frac{k}{k^2} =$$

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

Indice di Entropia di Shannon

$$H = -\sum f_i \lg(f_i)$$