STATISTICA II

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

TEMA DEL CORSO: utilizzare un software per:
- analizzare Tabelle
- analizzare serie di dati (stringhe)

L'obiettivo è quello di estrarre informazioni visive e numeriche da strutture di numeri.

Per arrivare alla CORRELAZIONE ...

- X VARIABILE ALEATORIA: grandezza che non si conosce in anticipo
- $E[X] = \mu_X$ VALORE ATTESO
- $Var[X] = E[(X - E[X])^2] = E[(X-\mu_X)^2] =$

$= E[X^2] - \mu_X^2 = \sigma_X^2$ dove σ_X^2 è la DEVIAZIONE STANDARD

DEFINIZIONE dato X e Y chiamiamo **COVARIANZA** tra X e Y il numero:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

Se $X = Y \Rightarrow Cov(X,X) = Var(X)$

DIMOSTRAZIONE

$$Cov(X,Y) = E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] =$$

$$= E[XY] - \mu_Y E[X] - \mu_X \underbrace{E[Y]}_{\mu_Y} + \mu_X\mu_Y =$$

$$= E[XY] - \mu_X\mu_Y$$

A livello pratico la covarianza non dà un messaggio intuitivo a causa dell'unità di misura, così come la varianza.
Nel caso della varianza si considera la deviazione standard mentre per la covarianza si introduce la correlazione.

DEFINIZIONE chiamiamo **COEFFICIENTE DI CORRELAZIONE** tra X e Y il numero definito da:

$$\text{Corr}(X,Y) = \rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Il difetto è che la covarianza dipende dall'unità di misura perché il suo scopo è diverso da quello di σ (che misura la dispersione) in quanto deve catturare il legame tra due stringhe.

TEOREMA ① $\text{Cov}(\alpha X, \beta Y) = \alpha\beta\, \text{Cov}(X,Y) \quad \forall \alpha,\beta \in \mathbb{R}$

e

$\text{Corr}(\alpha X, \beta Y) = \text{Corr}(X,Y) \quad \forall \alpha,\beta$

Vale inoltre:

② $-1 \leq \text{Corr} \leq 1$

DIMOSTRAZIONE ①

$$\text{Cov}(\alpha X, \beta Y) = E\left[(\alpha X - \mu_{\alpha X})(\beta Y - \mu_{\beta Y})\right] =$$

$$= E\left[(\alpha X - \alpha\mu_X)(\beta Y - \beta\mu_Y)\right] =$$

$$= \alpha\beta\, E\left[(X-\mu_X)(Y-\mu_Y)\right] =$$

$$= \alpha\beta\, \text{Cov}(X,Y)$$

Vale poi:

$$\text{Corr}(\alpha X, \beta Y) = \frac{\text{Cov}(\alpha X, \beta Y)}{\sqrt{\text{Var}[\alpha X]\,\text{Var}[\beta Y]}} =$$

$$= \frac{\alpha\beta\,\text{Cov}(X,Y)}{\sqrt{\alpha^2\beta^2}\,\sqrt{\text{Var}[X]\,\text{Var}[Y]}} = \frac{\alpha\beta}{|\alpha\beta|}\,\text{corr}(X,Y)$$

Se $\alpha, \beta > 0$ si ottiene la formula desiderata $\left\{\frac{\alpha\beta}{\alpha\beta}\,\text{corr}(X,Y)\right\}$ ■

DIMOSTRAZIONE (2)

DISUGUAGLIANZA DI HOLDER

$$\left|E[(X-\mu_X)(Y-\mu_Y)]\right| \le \underbrace{\sqrt{E[(X-\mu_X)^2]}}_{\sigma_X}\ \underbrace{\sqrt{E[(Y-\mu_Y)^2]}}_{\sigma_Y}$$

Se si divide tutto per σ_X e σ_Y si ottiene:

$$|\text{corr}(X,Y)| \le 1$$

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

■

TEOREMA Se X e Y sono variabili aleatorie indipendenti allora:

$\text{Cov}(X,Y) = 0$

$\text{Corr}(X,Y) = 0$

Viceversa se $\text{Cov}(X,Y) = 0$ non è detto che X e Y siano indipendenti

TEOREMA Se X e Y sono v.a. indipendenti allora:

$$E[XY] = \mu_X \mu_Y$$

CORRELAZIONE E REGRESSIONE

1° argomento

A partire da dei dati sperimentali $x_1, \dots x_n$ e $y_1, \dots y_n$ organizzati in questo modo:

	X	Y
1	x_1	y_1
⋮	⋮	⋮
n	x_n	y_n

si ha: $\bar{x} = \frac{x_1 + \dots + x_n}{n} \approx E[X]$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \approx \sigma_x^2$$

Considerando n coppie sperimentali $(x_1, y_1) \dots (x_n, y_n)$ si può calcolare la <u>VARIANZA EMPIRICA</u> e il <u>COEFFICIENTE DI CORRELAZIONE EMPIRICO</u>:

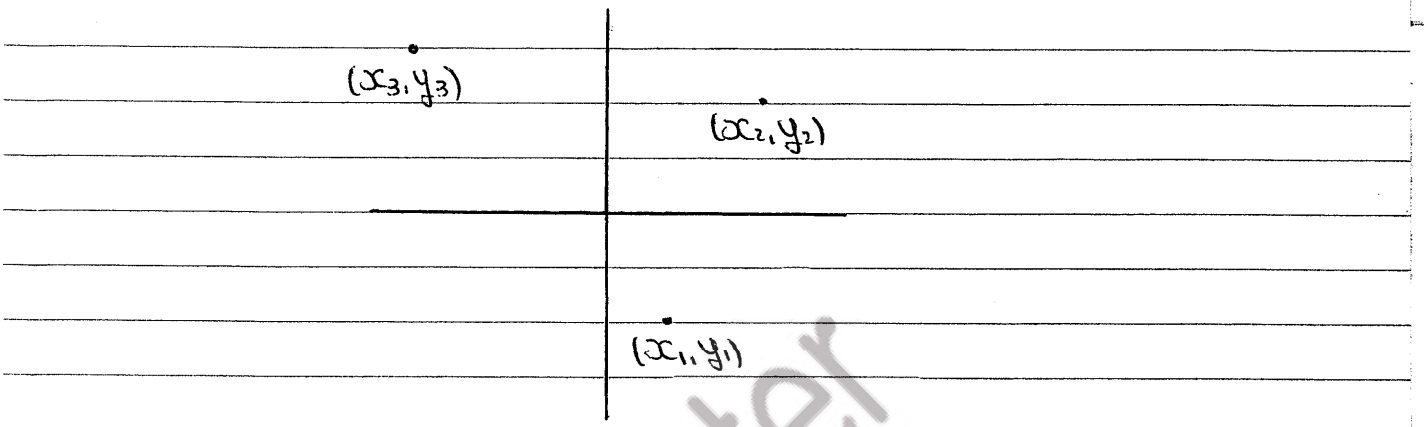
$$\widehat{Cov} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\widehat{Corr} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

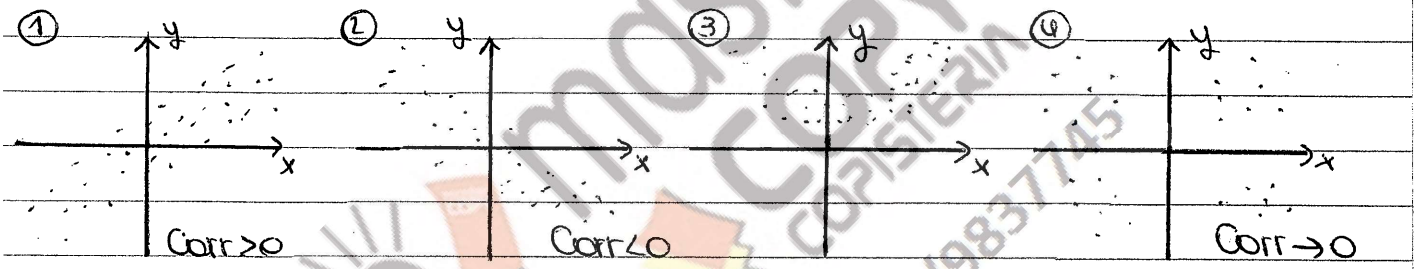
TEOREMA Se $Cov \neq 0$ e $Corr \neq 0$ allora X e Y non sono indipendenti

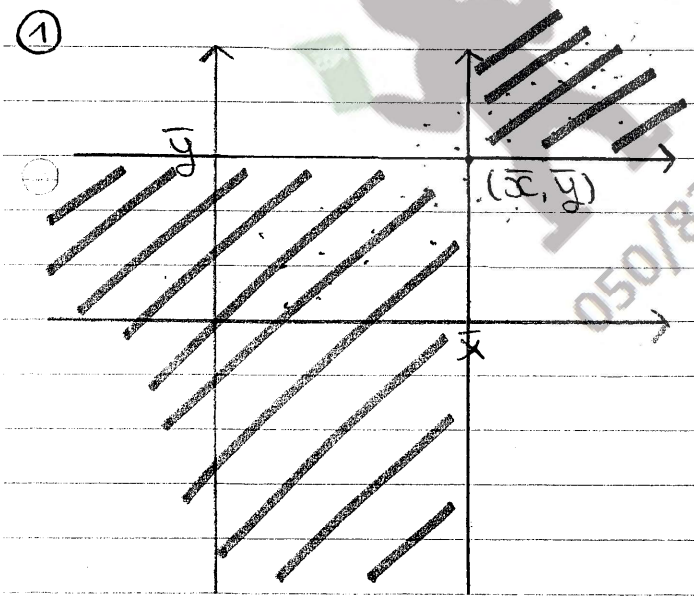
Master COPY COPISTERIA 050/531176 388/9837745

Si riportano queste coppie (x_i, y_i) in un diagramma cartesiano:



Possiamo avere 4 differenti distribuzioni:





Nel quadrante ≡:
$(x_i - \bar{x}) > 0$
$(y_i - \bar{y}) > 0$
Quindi il prodotto $(x_i - \bar{x})(y_i - \bar{y})$ è positivo mentre nella zona complementare è negativo.

Il punto $(\bar{x}, \bar{y})$ sta al centro della nuvola comunque essa sia presa.

master copy COPISTERIA 050/8311226 388/9837745