STATISTICA
FORMULARIO
DESCRITTIVA

BY ANDREA IL MATEMATICO

INDICE

CONCETTI PRELIMINARI

UNITÀ STATISTICA
È l'elemento fondamentale dell'indagine statistica rispetto al quale si studiano i suoi caratteri. Può essere una persona, un'animale oppure un oggetto

POPOLAZIONE
È l'insieme di tutte le unità statistiche.

CARATTERE
È una caratteristica che viene studiata di una unità statistica

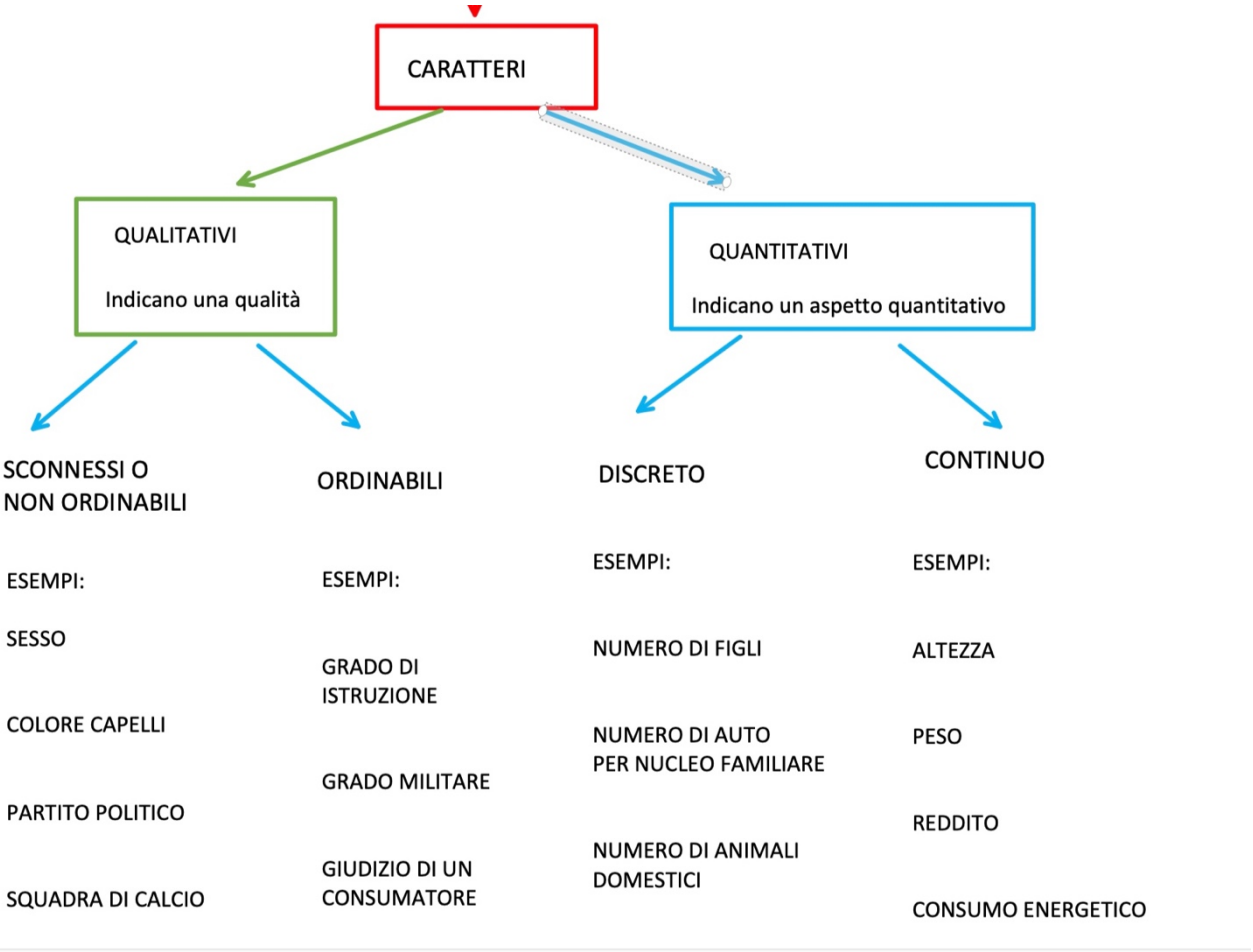
MODALITA'
Rappresentano i diversi modi in cui un carattere può manifestarsi

CLASSIFICAZIONE DEI CARATTERI

I caratteri possono essere QUALITATIVI o QUANTITATIVI.

QUALITATIVI: si distinguono in SCONNESSI e ORDINABILI

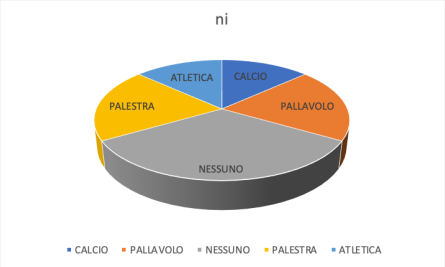
QUANTITATIVI: si distinguono in DISCRETI e CONTINUI



GRAFICI PER RAPPRESENTARE I CARATTERI

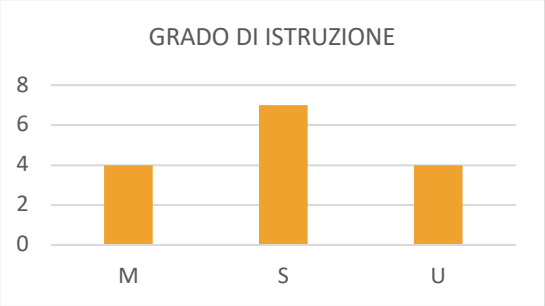
QUALITATIVO SCONNESSO:

grafico a torta



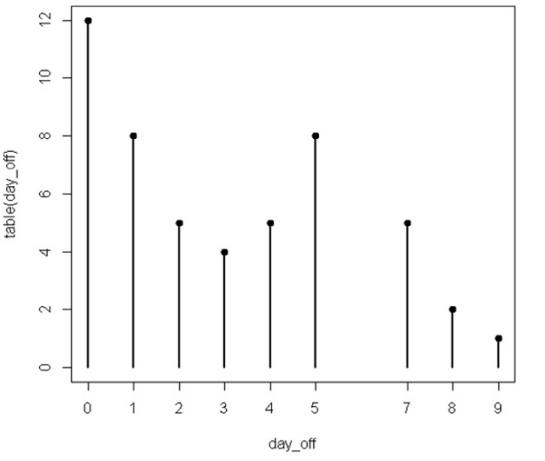
QUALITATIVO ORDINABILE

Grafico a barre o a nastri



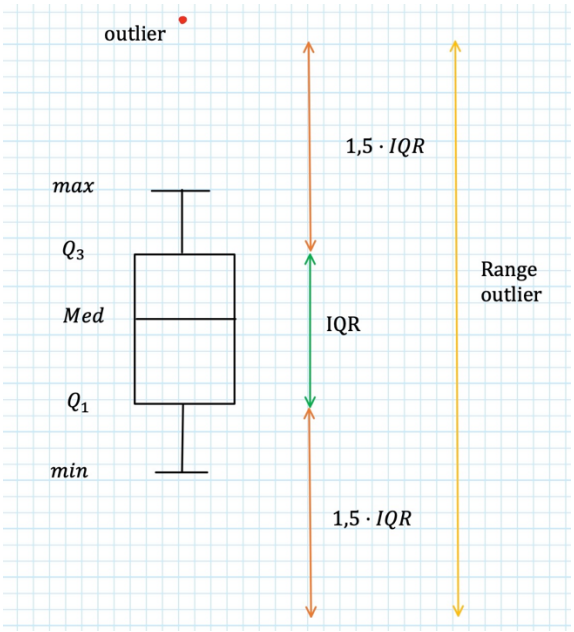
QUANTITATIVO DISCRETO

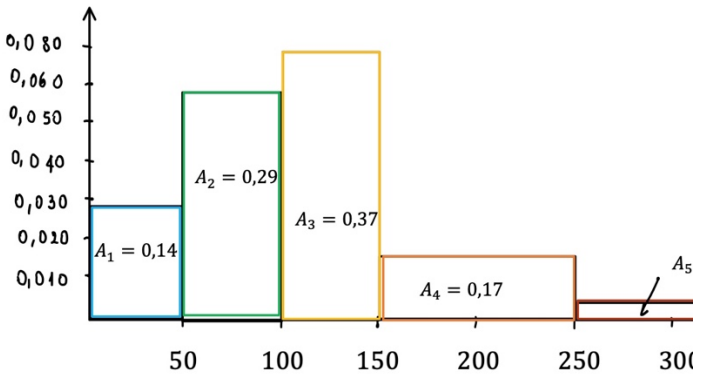
Grafico a bastoncini



QUANTITATIVO CONTINUO

Box-plot e Istogramma





TABELLE

CARATTERI QUALITATIVI SCONNESSI

CARATTERE (X)	ni	fi
X1		
X2		
…		

n_i = frequenza assoluta
f_i = frequenza relativa

CARATTERI QUALITATIVI ORDINABILI E QUANTITATIVI DISCRETI

CARATTERE (X)	ni	fi	Ni	Fi
X1				
X2				
…				

n_i = frequenza assoluta
f_i = frequenza relativa
N_i = frequenza assoluta cumulata
f_i = frequenza relativa cumulata

CARATTERI QUANTITATIVI CONTINUI

CARATTERE (X)	ci	ai	ni	fi	Ni	Fi	di
CLASSE 1 : $[\, x_{\min 1} - x_{\max 1})$							
CLASSE 2 : $[\, x_{\min 2} - x_{\max 2})$							
…							

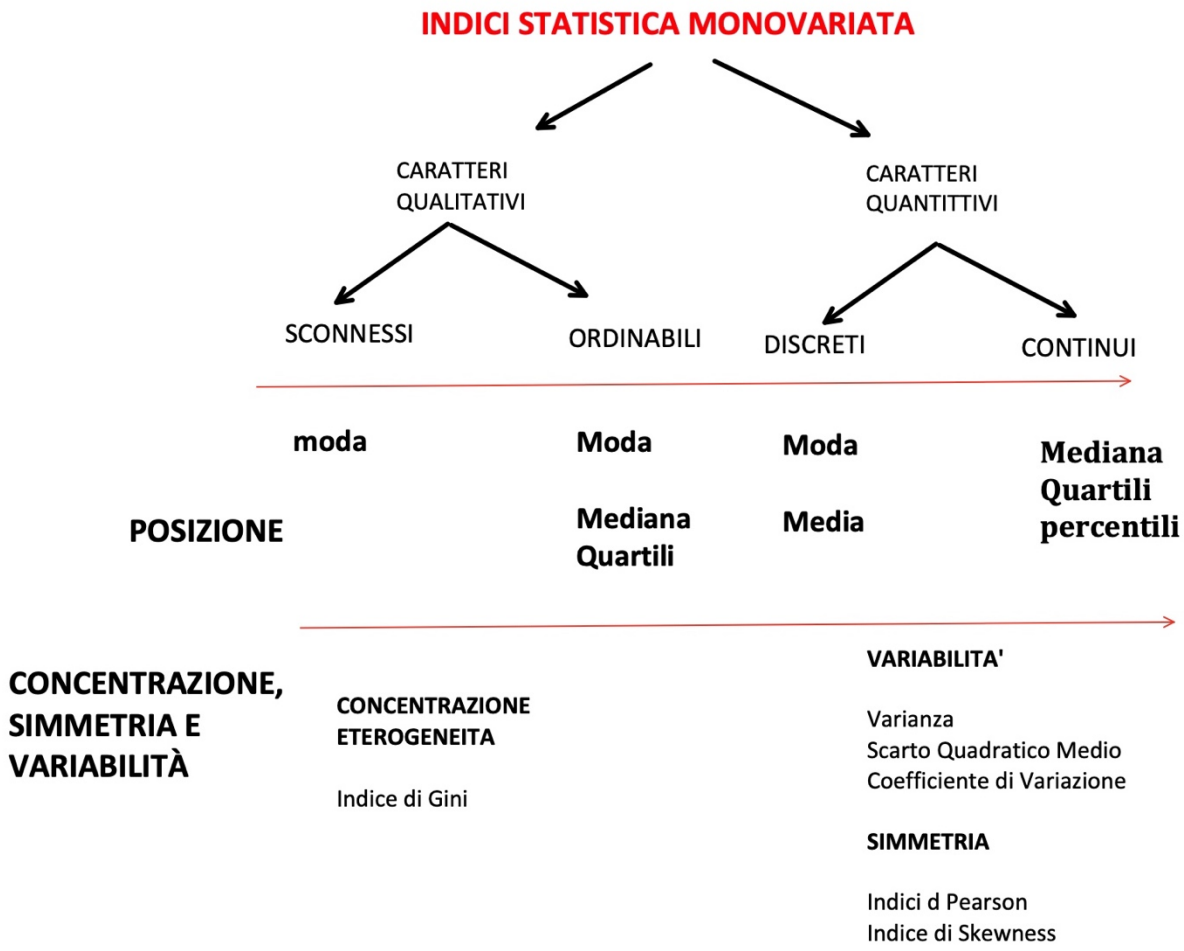
n_i = frequenza assoluta
f_i = frequenza relativa
N_i = frequenza assoluta cumulata
f_i = frequenza relativa cumulata

$c_i = \dfrac{x_{\max i} + x_{\min i}}{2}$ = valore centrale di classe
$a_i = x_{\max i} - x_{\min i}$ = ampiezza di classe

$d_i(\text{ass}) = \dfrac{n_i}{a_i}$ = densità assoluta di classe
$d_i(\text{rel}) = \dfrac{f_i}{a_i}$ = densità relativa di classe

CLASSIFICAZIONE DEGLI INDICI

Gli indici vengono classificati a seconda del tipo di carattere.
A mano a mano che ci spostiamo verso i caratteri quantitativi il numero di indici cresce



MODA

È La modalità che ha frequenza maggiore
Per i caratteri quantitativi continui parliamo di classe modale

MEDIANA

È la modalità che occupa la posizione centrale dei dati ordinati
In generale è calcolata per:

- Caratteri qualitativi ordinabili
- Caratteri quantitativi

Nella teoria di base quando abbiamo i dati in maniera ordinata
N (numero totale di unità statistiche) :
se N è dispari: posizione: $\frac{N+1}{2}$
se N è pari: posizioni: $\frac{N}{2}$ e $\frac{N}{2}+1$

Nel caso di caratteri qualitativi possiamo dunque avere due mode
Nel caso di caratteri quantitativi facciamo la media

Nel caso di caratteri quantitativi continui distribuiti in classi abbiamo la formula:

$$MED = x_{\min(MED)} + \frac{0{,}50 - F_{MED-1}}{F_{MED} - F_{MED-1}} \cdot a_{MED}$$

$x_{\min(MED)}$ = valore minimo della classe mediana
F_{MED} = frequenza cumulata della classe mediana
F_{MED-1} = frequenza cumulata della classe inferiore alla classe mediana
a_{MED} = ampiezza della classe mediana

QUARTILI

Im modo analogo possiamo indagare sui quartili
Q1 è il primo quartile: la modalità che segue il primo 25% dei dati ordinati
Q2 è il secondo quartile (mediana): la modalità che segue il primo 50% dei dati ordinati
Q3 è il terzo quartile: la modalità che segue il primo 75% dei dati ordinati

Nella teoria generale possiamo ricercare la posizione (le posizioni) con le seguenti formule

$$\text{pos}(Q_1) = \frac{1}{4} \cdot (n+1) \quad \text{pos}(Q_2) = \frac{2}{4} \cdot (n+1) \quad \text{pos}(Q_3) = \frac{3}{4} \cdot (n+1)$$

Se la posizione è perfetta allora abbiamo trovato il quartile interessato
Ad esempio se cerchiamo il primo quartile con n=19 abbiamo che la posizione cercata è la quinta.

$$\text{pos}(Q_1) = \frac{1}{4} \cdot (19+1) = 5$$

Se diversamente otteniamo una posizione intermedia avremo:

- 2 quartili per i qualitativi
- Procediamo con interpolazione per i quantitativi

Ad esempio se cerchiamo Q1 con n=20

$$\text{pos}(Q_1) = \frac{1}{4} \cdot (20+1) = 5{,}25$$

$$Q_1 = 0{,}75 \cdot x_{pos(5)} + 0{,}25\, x_{pos(6)}$$

Per i caratteri quantitativi distribuiti in classi abbiamo le seguenti formule:

$$Q_1 = x_{\min(Q_1)} + \frac{0{,}50 - F_{Q_1-1}}{F_{Q_1} - F_{Q_1-1}} \cdot a_{Q_1}$$

$$Q_2 = MED = x_{\min(MED)} + \frac{0{,}50 - F_{MED-1}}{F_{MED} - F_{MED-1}} \cdot a_{MED}$$

$$Q_3 = x_{\min(Q_3)} + \frac{0{,}50 - F_{Q_3-1}}{F_{Q_1} - F_{Q_3-1}} \cdot a_{Q_3}$$

PERCENTILI

I percentili sono una generalizzazione dei quartili

Per la posizione possiamo utilizzare la seguente formula:
$\text{pos}(p) = p \cdot (n+1)$
dove p è un valore compreso tra 0 e 1.
Ad esempio se p=0,9 significa che stiamo cercando il 90-esimo percentile.
Vale sempre anche la regola dell'interpolazione

$$x_p = (1-p)\, x_{\text{int}(pos(p))} + p\, x_{\text{int}(pos(p)+1)}$$

$x_{\text{int}(pos(p))}$ è la modalità che si trova nella posizione: parte intera della posizione di p

Per i caratteri continui raggruppati in classi vale sempre la formula generale:

$$x_p = x_{\min(p)} + \frac{0{,}50 - F_{p-1}}{F_p - F_{p-1}} \cdot a_p$$

LE MEDIE

Le medie si calcolano solamente per i caratteri quantitativi

MEDIA ARITMETICA

La media più conosciuta e più facile da calcolare è la media aritmetica:

Per i dati semplici facciamo la sommatoria delle modalità (dati) e le dividiamo per la numerosità n

$$\mu = \bar{x} = \frac{\sum x_i}{n}$$

Se i dati sono in tabelle di frequenza assoluta (ni) o relativa (fi) usiamo

$$\mu = \bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} x_i \cdot f_i$$

MEDIA GEOMETRICA

Con i dati semplici usiamo la formulazione

$$\mu_g = \overline{x_G} = \sqrt[n]{\prod x_i}$$

Con i dati di frequenza

$$\mu_g = \overline{x_G} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} = \left(\prod_{i=1}^{k} x_i^{n_i}\right)^{\frac{1}{n}} = \prod_{i=1}^{k} x_i^{\frac{n_i}{n}} = \prod_{i=1}^{k} x_i^{f_i} \qquad \text{con } n = \sum_{i=1}^{k} n_i$$

MEDIA ARMONICA

Con i dati semplici

$$\mu_A = \overline{x_A} = \frac{n}{\sum_{i=1}^{k} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{k} (x_i)^{-1}}{n}\right)^{-1}$$

Con i dati di frequenza

$$\mu = \overline{x_A} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{k} n_i (x_i)^{-1}}{\sum_{i=1}^{k} n_i}\right)^{-1}$$

MEDIA POTENZIATA

Con i dati semplici

Con i dati di frequenza

$$\mu_k = \overline{x_k} = \left(\frac{\sum_{i=1}^{k} (x_i)^k}{n}\right)^{\frac{1}{k}}$$

$$\mu_k = \overline{x_k} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{k} n_i (x_i)^{-1}}{\sum_{i=1}^{k} n_i} \right)^{-1}$$

VARIANZA

Con i dati semplici

$$\sigma^2 = \frac{\sum (x - x_i)^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2$$

Con i dati di frequenza

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{\sum_{i=1}^{k} n_i} - \bar{x}^2$$

SCARTO QUADRATICO MEDIO

È la radice quadrata della varianza

$$\sigma = \sqrt{\sigma^2}$$

COEFFICIENTE DI VARIAZIONE

$$CV = \frac{\sigma}{\mu}$$

SCOSTAMENTO SEMPLICE DALLA MEDIA

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

SCOSTAMENTO SEMPLICE DALLA MEDIANA

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

INDICI DI ASIMMETRIA

MEDIA RISPETTO A MEDIANA

$$A_2 = \frac{\mu - Me}{\sigma}$$

MEDIA RISPETTO ALLA MODA

$$A_2 = \frac{\mu - Mo}{\sigma}$$

INDICE DI ASIMMETRIA DI FISCHER

$\gamma = \frac{\bar{\mu}^3}{\sigma^3}$ dove $\bar{\mu}^3 = \frac{\sum(x_i - \mu)^3}{n}$ (media dei cubi degli scarti)

INDICE DI ETROGENEITÀ DI GINI

$$G = 1 - \sum_{i=1}^{k} f_i^2 \quad \rightarrow \quad G_N = G \cdot \frac{k}{k-1}$$

dove è il numero delle modalità del carattere

COEFFICIENTE DI GINI E CURVA DI LORENZ

TABELLA

CLASSE	ni	Ci (mi)	fi	Fi	ti= ci*ni	qi= ti/T	Qi

c_i = valore centrale di classe
m_i = media di classe (più preciso)
$f_i = \frac{n_i}{n}$ = frequenza relativa
$F_i = \sum_{j=1}^{i} f_j$ = frequenza relativa cumulata
$t_i = c_i \cdot n_i$ = valore centrale di classe
$T_i = \sum t_i$ = somma dei totali
$q_i = \frac{t_i}{T}$ = quota sul totale
$Q_i = \sum_{j=1}^{i} q_j$ = quote cumulate

$$G = 1 - \left(\sum (Q_i + Q_{i-1}) \cdot f_i \right) \cdot \frac{n}{n-1}$$

$$G = \frac{area\ concentrazione}{area\ max} = \frac{0{,}5 - 0{,}5 \cdot (\sum (Q_i + Q_{i-1}) \cdot f_i)}{0{,}5 \cdot \frac{n-1}{n}}$$

Le due formulazioni sono identiche

Il numeratore della seconda frazione è l'area di concentrazione data da 0,5 meno la somma delle aree di tutti i trapezi rettangoli compresi tra la curva e l'asse delle frequenze cumulate F
Il denominatore è l'area massima data da 0,5 moltiplicata per (n-1) fratto n

Semplificando la seconda espressione si giunge alla prima formulazione

DISTRIBUZIONE DOPPIA – STATITISTICA BIVARIATA

INDICE CHI QUADRATO

Maggiormente utilizzato per i caratteri qualitativi
Indica la presenta di connessione o dipendenza tra due variabili

TABELLA

	Y1	…	Yj	…	Yk
X1					
…					
Xi			n_{ij}		$n_{i.}$
…					
Xh			$n_{.j}$		n

n_{ij} = frequenza osservata
$n_{i.}$ = totale riga i
$n_{.j}$ = totale colonna j
$n = \sum_{i=1}^{h} n_{i.} = \sum_{j=1}^{k} n_{.j}$ = totale elementi
$t_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$ = frequenza teorica di indipendenza
$c_{ij} = n_{ij} - t_{ij}$ = contingenza

CALCOLO CHI-QUADRATO

$$\chi^2 = \sum_i \sum_j \frac{c_{ij}^2}{t_{ij}} = \sum_i \sum_j \frac{n_{ij}^2}{n_{i.}} - 1$$

CHI-QUADRATO MASSIMO

$$\chi^2_{MAX} = n \cdot \min(h-1, k-1)$$

CHI-QUADRATO NORMALIZZATO

$$\chi^2_N = \frac{\chi^2}{\chi^2_{MAX}} = \begin{cases} 0 & \text{perfetta indipendenza} \\ 1 & \text{perfetta dipendenza} \end{cases}$$

REGRESSIONE

COVARIANZA

$$\text{cov}(x,y) = \sigma_{xy} = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x} \cdot \bar{y}$$

CORRELAZIONE
La correlazione è un indice compreso tra -1 e +1 e indica il grado di correlazione lineare tra due variabili

$$R = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

INDICE DI DETERMINAZIONE
È il quadrato della correlazione
Indica quanta parte della variabilità di y può essere spiegata dalla x

$$R = (\rho_{xy})^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2}$$

BETA 1 E BETA 0
Sono rispettivamente il coefficiente angolare e l'intercetta all'origine della retta di regressione lineare
La x è intesa come variabile indipendente, mentre la y è variabile dipendente

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \rho_{xy} \cdot \frac{\sigma_y}{\sigma_x} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

EQUAZIONE DELLA RETTA DI REGRESSIONE

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

ANALISI ERRORI SULLA SINGOLA UNITA'

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

errore tot. = errore spiegato + errore non spiegato

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

IN TERMINI DI DEVIANZA

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \rightarrow \quad SQT = SQR + SQE$$

INDICE DI DETERMINAZIONE

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$