Orientamento consapevole
Statistica
a.a. 2021-2022

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
DIPARTIMENTO DI ECONOMIA E FINANZA

Facciamo statistica! Conoscere la realtà che ci circonda attraverso i numeri

Materiale didattico

prof.ssa Nunziata Ribecco *nunziata.ribecco@uniba.it*
prof.ssa Angela Maria D'Uggento *angelamaria.duggento@uniba.it*

Oltre a descrivere i fenomeni, la Statistica è utile anche per **esplorare le relazioni** tra le variabili…

indipendenza, dipendenza e interdipendenza?

Analisi statistica univariata

Considera una sola variabile/mutabile per volta, ad esempio età, voto, numero iscritti, altezza, residenza, titolo studio, genere, ecc.

Gli strumenti

1. Medie e rappresentazioni grafiche
2. Variabilità
3. Numeri indici

Analisi statistica bivariata

Studia due variabili/mutabili contemporaneamente con scopo descrittivo o per individuare le eventuali relazioni tra le variabili quali ad esempio peso/altezza, voto X/voto Y, numero iscritti/sesso, titolo studio/voto laurea, risparmio/consumo, ecc.

Le relazioni possibili sono: Indipendenza; Regressione semplice; Correlazione semplice.

Analisi statistica multivariata

Studia l’influenza contemporanea di tre o più variabili

Tecniche statistiche

1. Analisi fattoriale
2. Cluster analysis
3. Scaling Multidimensional
4. Regressione multipla
5. Correlazione canonica
6. Analisi corrispondenze

Relazioni tra variabili

REGRESSIONE → Analisi della **DIPENDENZA** tra due variabili statistiche X e Y

CORRELAZIONE → Analisi della **INTERDIPENDENZA** tra due variabili statistiche X_1 e X_2

INDIPENDENZA → Il variare di una delle due variabili statistiche X_1 e X_2 non produce alcun effetto sull'altra

L’analisi della dipendenza o regressione

Con la regressione si analizza la dipendenza della variabile Y in funzione della variabile X

X = Variabile INDIPENDENTE o REGRESSORE O ANTECEDENTE

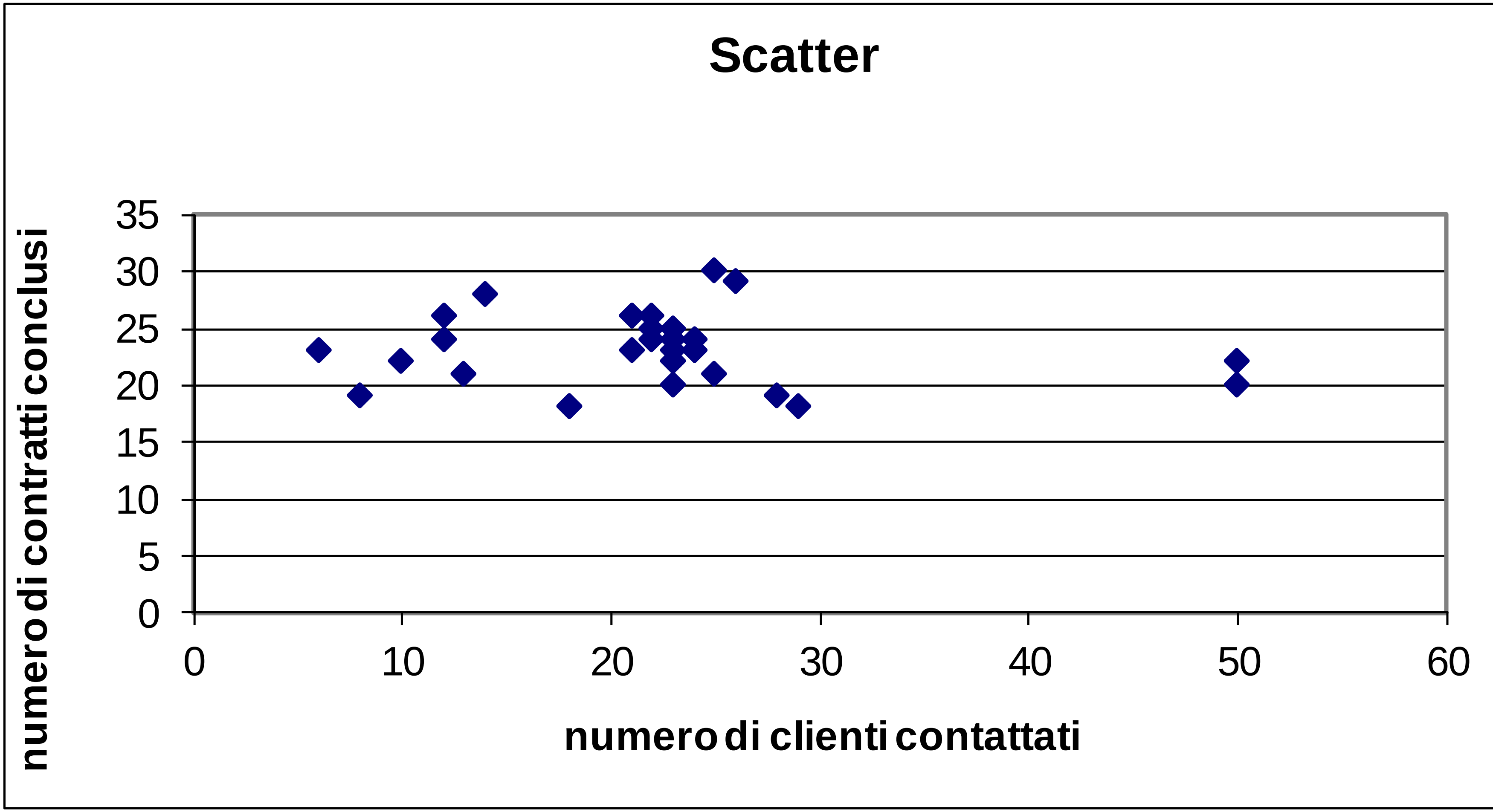
Y = Variabile DIPENDENTE o CONSEGUENTE

Lo scopo della regressione è quello di individuare la eventuale relazione tra le due variabili (rappresentata graficamente dalla nuvola di punti o scatter) esplicitandola attraverso una funzione matematica

Esempi:

X = n.ro clienti contattati; Y = n.ro contratti conclusi

X= reddito; Y=consumo



Scatter
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
numero di contratti conclusi
numero di clienti contattati

Regressione lineare

La relazione tra le due variabili è, per semplicità, supposta lineare, pertanto sarà studiata attraverso il modello della retta di regressione

$$y^* = a + bx$$

La stima dei parametri incogniti a e b della funzione di regressione avviene con il **metodo dei minimi quadrati.** Tale metodo consiste nel rendere minima la differenza al quadrato tra valori teorici (valori da modello) e valori empirici (dati rilevati).

Regressione lineare

Il parametro *a* esprime il valore che assume Y quando X è pari a 0

Il parametro *b* indica come varia **IN MEDIA** il carattere Y al variare di **una unità** del carattere X

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \end{cases}$$

Con Excel: b si calcola con funzione «pendenza», a con funzione «intercetta» . L'equazione si inserisce insieme a R^2 dal grafico a dispersione

Il significato dei parametri della retta di regressione

a è l'intercetta o termine noto ed esprime il valore di Y quando x=0;

b è il coefficiente angolare della retta e quindi

$$-\infty \leq b \leq +\infty$$

E' anche detto coefficiente di regressione ed esprime la variazione media del carattere Y al variare **unitario** del carattere X.

Se b>0 c'è dipendenza diretta tra X e Y, cioè Y aumenta in media all'aumentare di X;

se b<0 c'è dipendenza inversa tra X e Y, cioè Y diminuisce in media all'aumentare di X;

se b=0 vi è indipendenza di Y da X .

L’interdipendenza o Correlazione semplice

La correlazione misura l’interdipendenza tra due caratteri X_1 ed X_2 in termini di concordanza o discordanza.

In tal caso non è possibile distinguere il carattere dipendente da quello indipendente.

Una misura assoluta della concordanza/discordanza è la codevianza.

$$Codev(X,Y) = \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$$

La correlazione si misura con il coefficiente di correlazione r di Bravais Pearson

La Correlazione semplice

Codevianza

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

Devianza di X e Devianza di Y

=+1 max concordanza

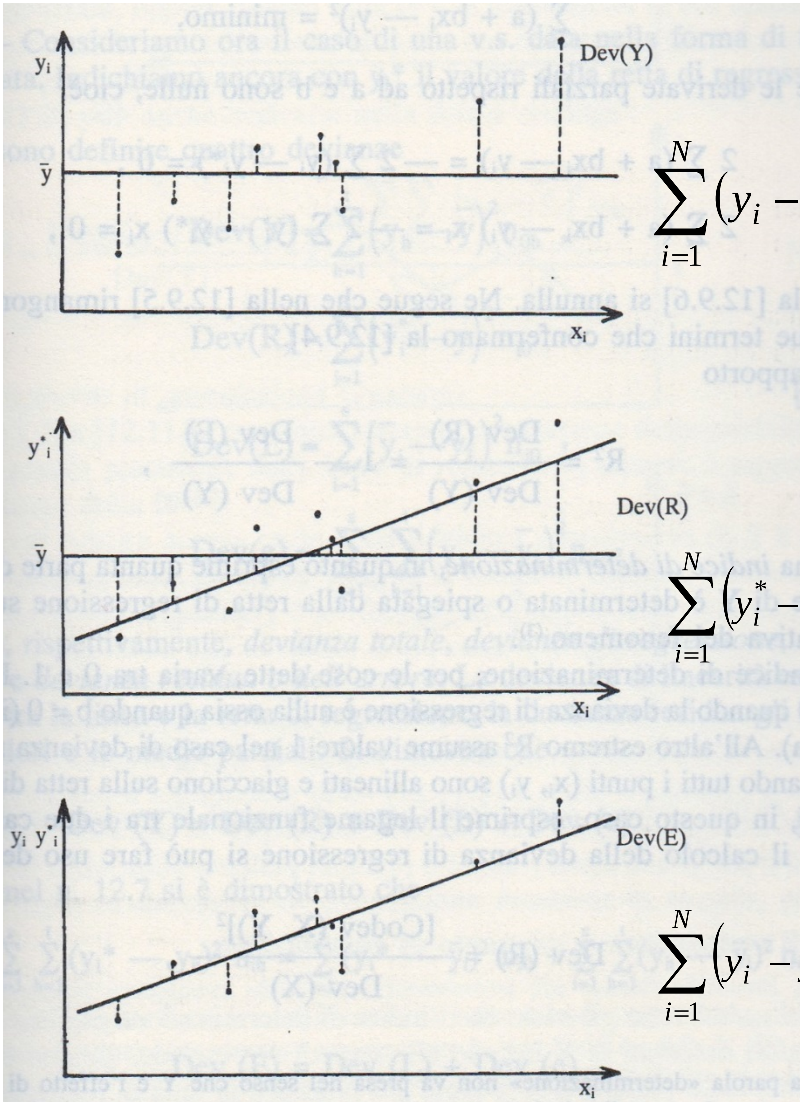
=0 indifferenza

=-1 max discordanza

$$-1 \leq r \leq +1$$

Con Excel: il coefficiente di correlazione si calcola con la funzione «correlazione»

Varianza di regressione



$$\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2$$

$$\sum_{i=1}^{N}\left(y_i^*-\bar{y}\right)^2$$

$$\sum_{i=1}^{N}\left(y_i-y_i^*\right)^2$$

12.9.1 - *Schema grafico delle devianze totale, di regressione e residua.*

Analogamente a quanto accade per la media, è possibile studiare la dispersione dei valori osservati di Y intorno alla retta di regressione. Dispersione elevata significa limitata attendibilità delle previsioni fatte con il modello scelto (retta di regressione); il contrario se la dispersione è bassa.

L'adattamento si misura con R^2.

Le tre devianze sono date rispettivamente dalla somma dei quadrati dei segmenti verticali tratteggiati.

Si dimostra che

Dev(Y) = Dev(R) + Dev(E)

Fitting del modello: l'indice di determinazione R²

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \frac{\sum_{i=1}^{N}\left(y_i^* - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}$$

$$R^2 = \frac{Dev(R)}{Dev(Y)} = 1 - \frac{Dev(E)}{Dev(Y)} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i^*\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}$$

R^2 esprime la bontà di adattamento del modello di regressione cioè quanta parte della devianza totale di Y è spiegata dalla retta scelta

$$0 \leq R^2 \leq 1$$

$R^2 = +1$ quando $Dev(E) = 0$ cioè tutti i punti sono perfettamente allineati su retta regressione

$R^2 = 0$ quando $Dev(R) = 0$ cioè $b = 0$ cioè indipendenza in media