Percorsi didattici, interdisciplinari ed innovativi per la Statistica

Maurizio Vichi

Sapienza Università di Roma

Presidente Federazione Europea delle Società Nazionali di Statistica

Scuola Estiva di Matematica per i Docenti delle Scuole Secondarie di 2°Grado
Montegrotto Terme, 22-25 Luglio 2014

Statistica e Informatica

USO DELLE BASI DATI PER UNA INDAGINE STATISTICA

- Il Questionario di una indagine e la matrice dei dati – Maschera di immissione dati e la base dati informatica per raccogliere i dati
- Schematizzazione di un fenomeno statistico e lo schema Entità/Relazione
- Distribuzioni di frequenze viste come una query di un database

PARALLELISMO TRA STRUTTURE STATISTICHE ED INFORMATICHE

- le strutture di teoriche della statistica (unità, variabile, popolazione, campione) e le strutture informatiche dei dati (i record, i file, le tabelle di un database)
- una unità statistica multivariata può essere vista come un «record di dati»
- La matrice di dati (unità statistiche x variabili) può essere vista come un «file di dati»

PARALLELISSMO TRA INDICI STATISTICI E ALGORITMI di CALCOLO

- indicatori di sintesi dei dati: medie, indici di variabilità di correlazione, di regressione viste come algoritmi di calcolo

Primi elementi per la realizzazione di una indagine statistica

USO DELLE BASI DATI PER UNA INDAGINE STATISTICA
- Schematizzazione di un fenomeno statistico e lo schema Entità/Relazione
- Il Questionario di una indagine e la matrice dei dati – Maschera di immissione dati e la base dati informatica per raccogliere i dati
- Distribuzioni di frequenze viste come una query di un database

PARALLELISMO TRA STRUTTURE STATISTICHE ED INFORMATICHE

-le strutture di teoriche della statistica (unità, variabile, popolazione, campione) e le strutture informatiche dei dati (i record, i file, le tabelle di un database)
-una unità statistica multivariata può essere vista come un «record di dati»
-La matrice di dati (unità statistiche x variabili) può essere vista come un «file di dati»

I Fase: Schematizzazione di un fenomeno statistico mediante lo Entità/Relazione

Il **modello di rappresentazione Entità-Relazione** è stato proposto da P. P. Chen nel 1976 in ambito informatico allo scopo di facilitare la progettazione di una base di dati. Come vedremo più avanti, il modello Entità-Relazione può essere utilizzato proficuamente anche per una formale e precisa descrizione qualitativa di un fenomeno collettivo complesso.

Mediante tale modello si può costruire un grafico, ossia uno schema che costituisce una fotografia del fenomeno. Ciascuno schema è il risultato della composizione logica di cinque tipi di **strutture di rappresentazione** che si avvalgono di regole formali e di una simbologia grafica molto semplice ma estremamente chiara. Tali strutture sono: l'entità, la relazione, l'attributo, il sottoinsieme, la gerarchia di generalizzazione.

Per facilitare la comprensione della denominazione informatica data alle strutture di rappresentazione affiancheremo, la terminologia statistica che individua lo stesso tipo di struttura.

Esaminiamo la simbologia del modello Entità-Relazione.
Il **rettangolo** rappresenta una **entità**, ossia un insieme di oggetti di cui sono note alcune caratteristiche. In termini statistici, il rettangolo rappresenta un collettivo di unità statistiche, denominate nell'ambito dell'informatica, **istanze** dell'entità.

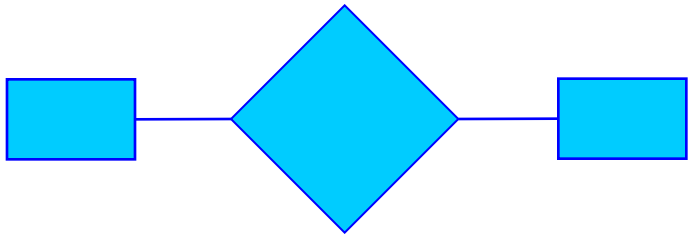
Struttura di rappresentazione	**Terminologia Informatica**	**Terminologia Statistica**
	**Entità**	**Collettivo di unità statistiche** **Popolazione**

Il **piccolo cerchio all'estremità di un segmento** rappresenta un **attributo**, ossia una proprietà elementare dell'entità, normalmente collegato a un'entità (un rettangolo) in quanto rappresenta una sua caratteristica. L'attributo è in realtà una variabile statistica del fenomeno analizzato. L'insieme dei valori che può assumere un attributo è detto **dominio** e rappresenta l'insieme delle possibili modalità che un carattere può presentare, ossia la scala delle modalità del carattere.

Struttura di rappresentazione	Terminologia Informatica	Terminologia Statistica
	**Attributo di una entità**	**Variabile o carattere**

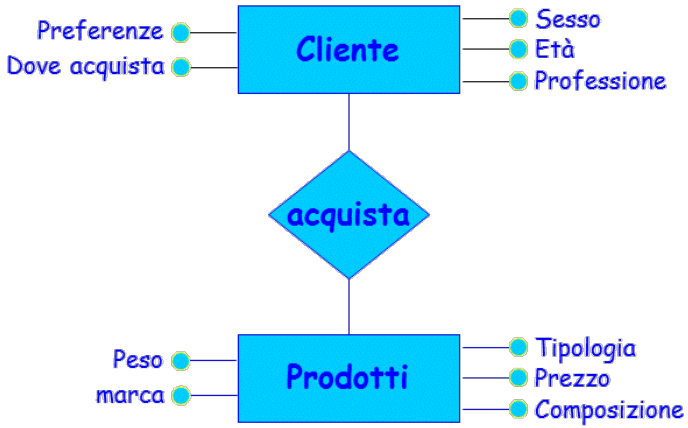
Esempio 1.1. **Schema E-R per collettivo occupati**

Il **rombo** corrisponde ad una **relazione**, indicando un legame logico tra le istanze di due o più entità. Nella rappresentazione grafica, il rombo è collegato a due rettangoli (entità o collettivi).

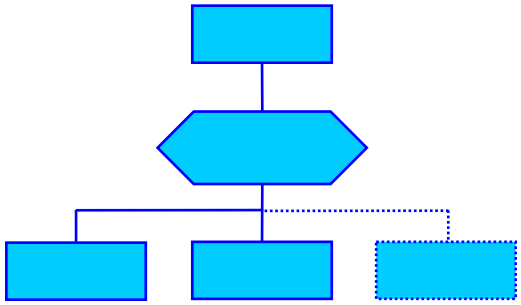
Struttura di rappresentazione	Terminologia Informatica	Terminologia Statistica
	Relazione fra due entità	Relazione fra collettivi

Come si può immediatamente osservare alla struttura "relazione" non corrisponde una struttura statistica, come invece risulta per l'entità e gli attributi, e quindi per essa non si intuisce immediatamente come possa aiutare a descrivere, con successivi mezzi statistici, il fenomeno in esame. La relazione invece ipotizza che tra i due collettivi e i dati ad essi associati esista un collegamento causale (relazione statistica) e che questa debba essere investigata, con le idonee metodologie statistiche.

Esempio 1.2. **Relazione fra prodotti e clienti**



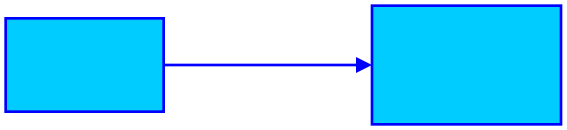
Un **esagono** che collega un rettangolo ad altri due o più rettangoli, rappresenta la **gerarchia di generalizzazione**, ossia una partizione del collettivo di unità statistiche in sottocollettivi omogenei secondo una o più variabili di stratificazione o di classificazione.

Struttura di rappresentazione	Terminologia Informatica	Terminologia Statistica
	Gerarchia di Generalizzazione tra entità	Partizione di un collettivo in sottocollettivi

Nel caso della gerarchia di generalizzazione, le unità statistiche che costituiscono il collettivo posto in alto, vengono ripartite nei collettivi di unità statistiche posti in basso. Pertanto, tutti i collettivi posti in basso nella gerarchia di generalizzazione, denominati **entità figlie** o **collettivi figli**, possiedono tutte le caratteristiche del collettivo posto in alto, detto **entità padre** o **collettivo padre**.

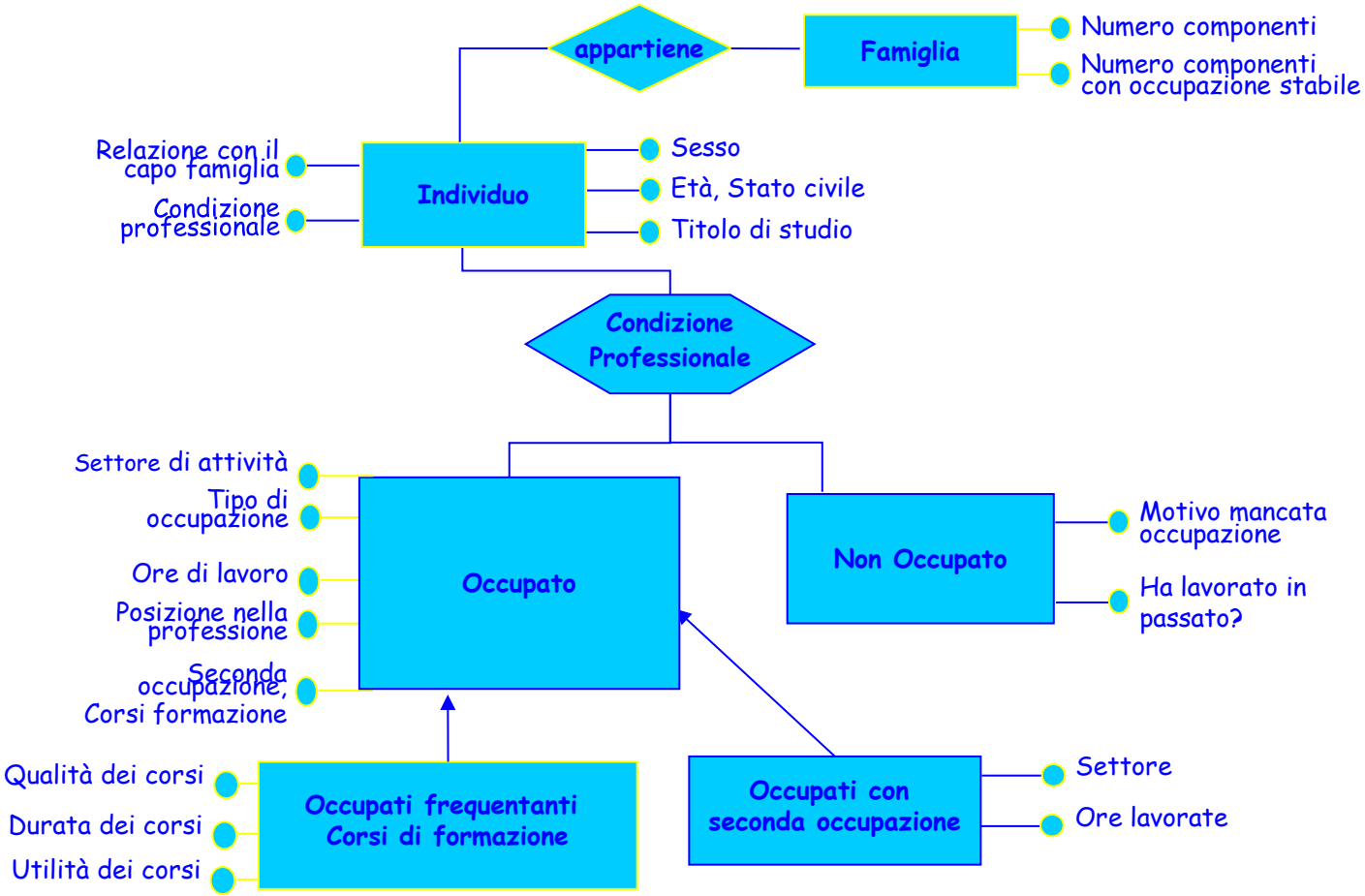
Esempio 1.3. **Gerarchia di generalizzazione**

Si noti che la variabile di stratificazione si suppone sia una delle variabili rilevate nel collettivo padre anche se nell'entità padre spesso non viene riportata esplicitamente, ma solo indicata nell'esagono.
Una **freccia** che collega due rettangoli, rappresenta un legame di sottoinsieme tra due entità.

Struttura di rappresentazione	**Terminologia Informatica**	**Terminologia Statistica**
	**Sottoinsieme di un insieme**	**Sottocollettivo di un collettivo**

Nel caso del sottoinsieme, diversamente dalla gerarchia di generalizzazione, non tutti gli oggetti del collettivo posto in alto fanno parte del collettivo o dei collettivi posti in basso; i collettivi figli, ossia quelli posti in basso, possiedono comunque tutte le caratteristiche del collettivo padre.
Si noti che anche per identificare il sottoinsieme è necessaria una variabile di stratificazione che si suppone sia una delle variabili rilevate nel collettivo padre anche se nell'entità padre spesso non viene riportata esplicitamente.

Mediante il collegamento logico di questi semplici strutture di rappresentazione si perviene ad uno **schema concettuale del fenomeno** in esame. Vediamo lo schema che si riferisce al fenomeno "forze di lavoro"



Fase 2: Progettazione questionario mediante schema E-R

Lo schema concettuale si può utilizzare per la progettazione delle domande del questionario al fine di individuarne le proposizioni e/o i quesiti da inserire.

Lo schema concettuale ci aiuta a individuare l'esistenza di sezioni omogenee all'interno di un questionario.

Si noti che ad uno schema concettuale possono anche corrispondere più questionari le cui informazioni vengono tra loro collegate attraverso le relazioni tra le entità.

Nella prima fase della progettazione del questionario, con l'ausilio dello schema Entità-Relazione, a ciascuna variabile, riportata nello schema si fa corrispondere una o più domande.

A ciascun collettivo, rappresentato nello schema, si fa corrispondere una sezione del questionario omogenea per argomento trattato.

A ciascun collettivo figlio in una gerarchia di generalizzazione o a ciascun collettivo figlio in una relazione di sottoinsieme si fa corrispondere una sottosezione del questionario.

Ad ogni gerarchica di generalizzazione o ad ogni sottoinsieme deve corrispondere una domanda filtro

Abbiamo utilizzato come identificatori delle domande una sequenza di caratteri che indica il tipo di domanda, seguita da un numero, come segue:

Le domande di un questionario e le variabili statistiche

Ad ogni domanda di un questionario è associata una variabile statistica.
Le domande di un questionario sono:
1. Libera a risposta aperta; 2. strutturata a risposta chiusa; 3. a risposta mista.
Nelle **domande a risposta aperta** o semplicemente **domande libere**, l'intervistato ha la facoltà di rispondere esprimendosi nella forma e con i termini che preferisce.

Le **domande a risposta chiusa** o anche dette **strutturate**, prevedono delle alternative fisse di risposta. L'intervistato risponde alle domande scegliendo, tra le diverse alternative indicate, quella che più si avvicina al suo pensiero.

Le domande sono con **una sola risposta** o con più di una risposta (multiresponse). Tra quest'ultime, ci sono quelle con più risposte gerarchizzate
Nelle **domande con una sola risposta** l'intervistato deve indicare, nel caso di domande strutturate, la risposta la unica corrispondente al proprio pensiero

Nelle **domande con più di una risposta** o **domande multiresponse**, l'intervistato indica una o più risposte fra le varie alternative proposte in una domanda strutturata, ovvero formula una o più risposte, nel caso di una domanda libera.

Nel questionario si introducono le **domande filtro** che indirizzano gli intervistati verso le domande a loro pertinenti saltando altre domande che non li riguardano. Le domande di questo tipo identificano sottoinsiemi di intervistati, aventi particolari caratteristiche in comune.
Dopo ciascuna domanda filtro si deve riportare una nota che indichi, secondo la risposta data alla domanda filtro, a quale successivo quesito l'intervistato deve rispondere. Le domande filtro sono con due alternative; individuano due gruppi di rispondenti uno alternativo all'altro, Si possono incontrare domande filtro con più di due alternative.

Esempi di domande

1. Domanda libera
2. Domanda strutturata
3. Domanda strutturata in classi
4. Domanda a risposta mista
5. Domanda con una sola risposta
6. Domanda con più di una risposta
7. Domanda gerarchizzata
8. Domanda filtro
9. Domande codificate

Esempio 1. Domanda libera o a risposta aperta

Quale professione svolge?

.....

.....

Che cosa ha mangiato a pranzo?

.....

.....

Che cosa pensa della pena di morte?

.....

.....

Nella formulazione di una domanda libera il ricercatore deve essere attento a formulare la domanda in modo da non influenzare il rispondente

Fine Esemp

Esempio 2. Domanda strutturata

Nell'indagine sulle forze di lavoro dell'ISTAT si incontra la seguente domanda strutturata

Per quale motivo ha lasciato l'occupazione ?
Licenziamento . . . . . □ 1
Fine di un lavoro a tempo determinato . . . . . □ 2
Dimissioni o cessazione di attività . . . . . □ 3
Pensionamento anticipato per motivi economici . . . . . □ 4
Ritiro dal lavoro per motivi di salute o maternità . . . . . □ 5

Pensionamento per raggiunti limiti di età o per
motivi diversi da quelli economici e di salute. . . . . □ 6
Servizio di leva . . . . . □ 7

La variabile associata alla precedente domanda è naturalmente *il motivo per cui ha lasciato l'occupazione*; le sette risposte corrispondono alla scala delle modalità della variabile. In una domanda a risposta chiusa si deve verificare se la domanda e le risposte sono state formulate in modo da non influenza il rispondente e se le risposte sono tutte quelle possibili.

Esempio 3. Domanda strutturata in classi

Una domanda con risposte strutturate è riportata di seguito.

Ore di lavoro settimanali complessivamente svolte
Meno di 10 ore . . . . . □ 1
Da 10 a 20 ore . . . . . □ 2
Da 21 a 40 ore . . . . . □ 3
Oltre le 40 ore . . . . . □ 4

Nella seguente classe si noti l'evidente perdita di dettaglio nell'informazione dovuta all'eccessiva ampiezza delle classi prescelte.

Quale professione svolge ?

Imprenditore e libero professionista . . . . . □ 1
Lavoratore in proprio . . . . . □ 2
Impiegato e lavoratore dipendente . . . . . □ 3
Casalinga . . . . . □ 4
Studente. . . . . □ 5
In condizione non professionale . . . . . □ 6

Fine Esemp

Esempio 4. Domanda a risposta mista

Per la domanda sugli alimenti consumati nel pasto, introdotta nell'esempio 2.3, possiamo riportare le voci alimentari più frequentemente consumate in Italia ed aggiungere la risposta "altro da specificare".

Che cosa ha mangiato a pranzo?

Pane e pizza . . . . . □ 1
Pasta. . . . . □ 2
Riso. . . . . □ 3
Carne. . . . . □ 4
Salumi . . . . . □ 5
Prodotti ittici freschi e surgelati . . . . . □ 6
Formaggi . . . . . □ 7
Uova . . . . . □ 8
Verdure e ortaggi . . . . . □ 9
Olio di oliva . . . . . □ 10
Vino . . . . . □ 11
Birra . . . . . □ 12
Acqua minerale. . . . . □ 13
Altro, da specificare . . . . . .

Fine Esemp

Esempio 2.5. Domanda con una sola risposta

Qual è la sua età in anni compiuti?

Questa è una domanda a risposta aperta che, al momento della somministrazione del questionario, prevede una sola risposta.

Qual è il suo stato civile?

celibe/nubile . . . . . □ 1
coniugato/a . . . . . □ 2
vedovo/a . . . . . □ 3
separato/a, divorziato/a, già coniugato/a . . . □ 4

Questa è una domanda strutturata che, al momento della somministrazione del questionario, prevede una sola risposta.

Fine Esemp

Esempio 2.6. Domanda con più di una risposta

Quali legge tra i seguenti giornali?
(sono previste più risposte)

La Repubblica . . . . . □ 1
Paese Sera . . . . . □ 2
Il Tempo . . . . . □ 3
Il Corriere della Sera . . . . . □ 4
Il Messaggero . . . . . □ 5
Il Giorno . . . . . □ 6

In questa domanda si prevede che un individuo possa leggere più di un giornale, quindi l'intervistato bifferà tutte le caselle corrispondenti ai giornali che legge.

Quali sono le trasmissioni televisive che segue?
(sono previste più risposte)

telegiornali . . . □ 1
films . . . . . □ 2
commedie . . . . . □ 3
documentari . . . . . □ 4
sport . . . . . □ 5

In tal caso, sapendo che la maggior parte degli individui può essere interessata a più di un programma televisivo, si chiede di esprimere tutte le proprie preferenze.

Le domande multiresponse possono essere formulate in modo da prevedere una sola risposta

Qual è il giornale che legge prevalentemente?

Fine Esemp

Esempio 2.7. Domanda gerarchizzata

La seguente domanda è gerarchizzata.

Indichi, in ordine di preferenza, con i numeri da 1 (preferito) a 4 (meno preferito), il suo gradimento fra i seguenti programmi televisivi.

telegiornali . . . . □
films . . . . . □
commedie . . . . . □
documentari . . . . . □

Fine
Esemp

Esempio 2.8. Domanda filtro

Nell'indagine sulle forze di lavoro dell'ISTAT, viene posta la seguente domanda filtro a tutti gli intervistati con 14 anni e più.

Quale era la sua condizione professionale nella settimana di riferimento?

Ha svolto ore di lavoro retribuito . . . . . □ 1
Ha un'occupazione, ma non ha svolto ore di lavoro . . . . . □ 2
Non ha lavorato perché sospeso dal lavoro (CIG)* . . . . . □ 3
E' militare di leva oppure in servizio civile sostitutivo . . . . . □ 4
Non ha svolto ore di lavoro e non possiede occupazione . . . . . □ 5

* *Cassa Integrazione Guadagni.*

Ogni intervistato che biffa le caselle 1 o 2 è considerato occupato e deve rispondere a tutte le domande della sezione degli occupati. Quelli che barrano una delle caselle dalla 3 alla 5 sono considerati non occupati, devono quindi saltare la sezione degli occupati e rispondere alle domande della sezione non occupati.

Fine Esemp

Esempio 2.9. Domande codificate

Nel questionario ISTAT per il Censimento Generale della Popolazione, vengono richieste, fra le altre, le seguenti notizie:

ACQUA CALDA Indicare se l'abitazione dispone di impianto per la produzione di acqua calda per uso igienico sanitario

si □ 6 no □ 7

In caso di risposta affermativa indicare se l'impianto di produzione è comune con quello del riscaldamento

si □ 8 no □ 9

TELEFONO Indicare se l'abitazione dispone di telefono

si □ 1 no □ 2

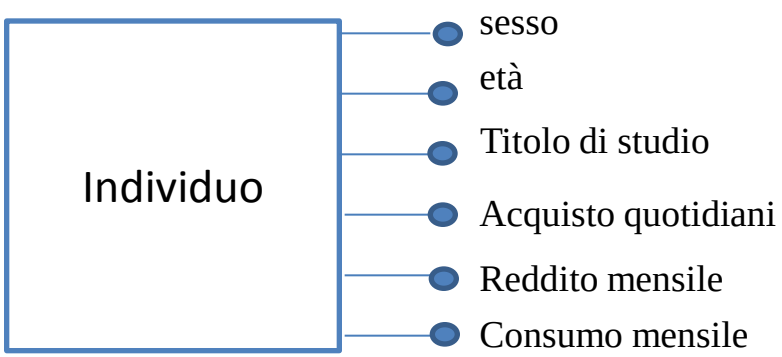
Si noti che le tre domande riportate hanno tutte due risposte precodificate. Per ciascuna sono stati utilizzati codici diversi per ovviare a problemi di disallineamento delle risposte.

Fine Esemp

Indagine statistica: Schema ER → Questionario → Base-dati (file) →Matrice Dati

Uso di ACCESS per la costruzione delle BASE DATI dell'Indagine Statistica

1 Schema ER del fenomeno



2 Questionario individuale

Sesso	**M=0; F=1 Età**
Titolo di studio	**Elem.=1, Media inf.=2, Media sup.=3, Laurea=4, Dottorato=5**
Acquisto quotidiano	**SI=1; NO=0**

Reddito mensile ……EUR Consumo mensile …….. EUR

3 Base dati mediante ACCESS e maschera di input Dati

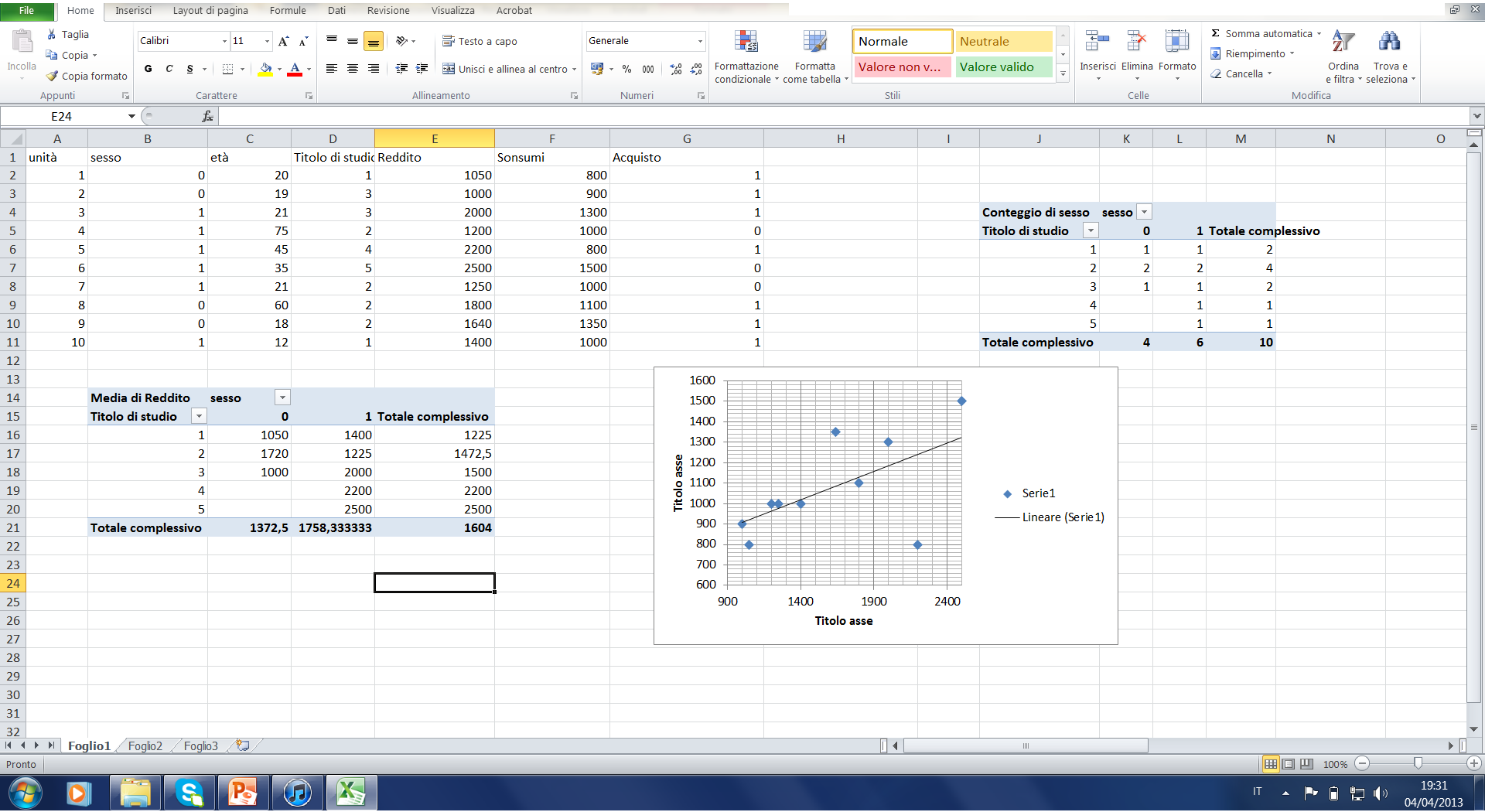
Unità	*sesso*	*Età*	*Titolo di studio*	*Reddito mensile*	*Consumo mensile*	*Acquisto. quotidiano*
1	0	20	1	1050	800	1
2	0	19	3	1000	900	1
3	1	21	3	2000	1300	1
4	1	75	2	1200	1000	0
5	1	45	4	2200	800	1
6	1	35	5	2500	1500	0
7	1	21	2	1250	1000	0
8	0	60	2	1800	1100	1
9	0	18	2	1640	1350	1
10	1	12	1	1400	1000	1

4 Matrice dei dati

$$X=\begin{bmatrix} 0 & 20 & 1 & 1050 & 800 & 1 \\ 0 & 19 & 3 & 1000 & 900 & 1 \\ 1 & 21 & 3 & 2000 & 1300 & 1 \\ 1 & 75 & 2 & 1200 & 1000 & 0 \\ 1 & 45 & 4 & 2200 & 800 & 1 \\ 1 & 35 & 5 & 2500 & 1500 & 0 \\ 1 & 21 & 2 & 1250 & 1000 & 0 \\ 0 & 60 & 2 & 1800 & 1100 & 1 \\ 0 & 18 & 2 & 1640 & 1350 & 1 \\ 1 & 12 & 1 & 1400 & 1000 & 1 \end{bmatrix}$$

Microsoft Excel

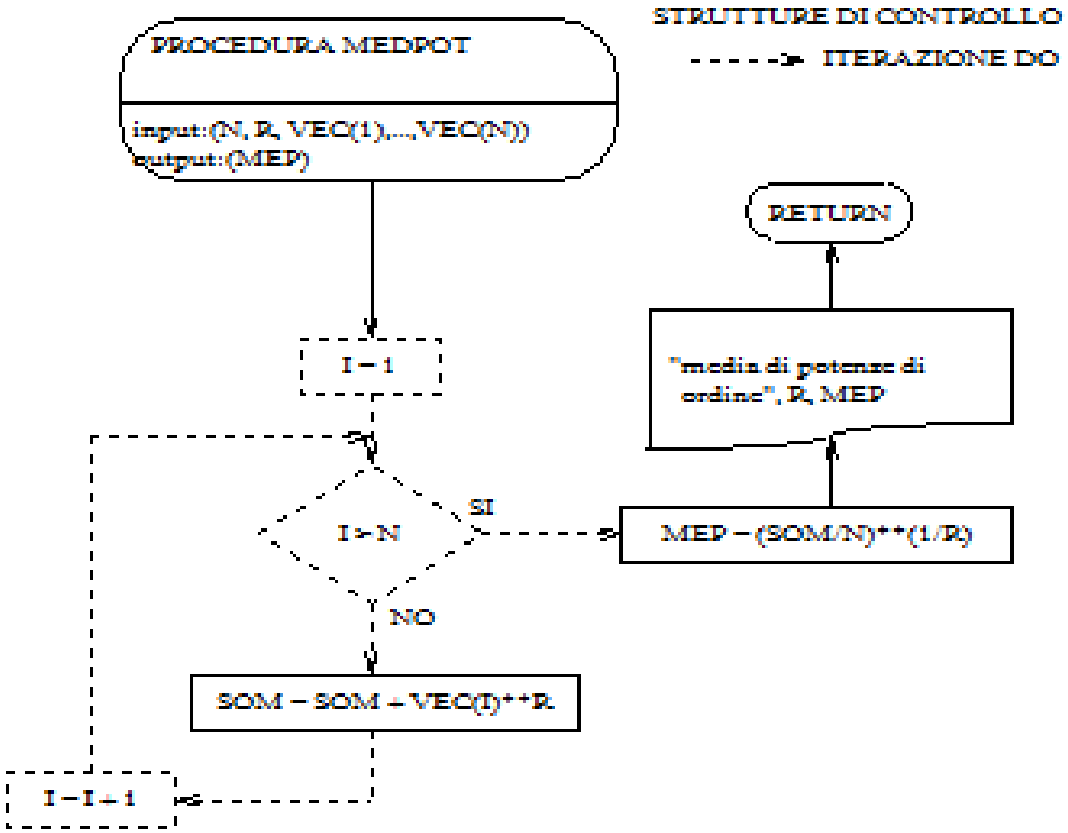
Unità	sesso	Età	Titolo di studio	Reddito mensile	Consumo mensile	Acquisto. quotidiano
1	0	20	1	1050	800	1
2	0	19	3	1000	900	1
3	1	21	3	2000	1300	1
4	1	75	2	1200	1000	0
5	1	45	4	2200	800	1
6	1	35	5	2500	1500	0
7	1	21	2	1250	1000	0
8	0	60	2	1800	1100	1
9	0	18	2	1640	1350	1
10	1	12	1	1400	1000	1



INDICI STATISTICI come ALGORITMI di CALCOLO

- indicatori di sintesi dei dati: medie, indici di variabilità di correlazione, di regressione viste come algoritmi di calcolo

$$Mr = \sqrt[r]{\frac{1}{n}\sum_{i=1}^{n} x_i^r}$$




```
Program MediaVettore;
Uses Crt;
Const Max=30;
Var V:ARRAY [1..Max] OF integer;
    Dim,i,Tot:integer;
    Med:real;

Begin
 Clrscr;
 Writeln ('Programma per calcolare il valore
medio di un vettore numerico');
 REPEAT
  writeln;
  Write('Inserisci la dimensione del vettore: ');
  Readln (N);
  Writeln;
 UNTIL (N>=1) AND (N<=Max);
 FOR i:=1 TO N DO
  Begin
   Write ('Dammi il numero di posizione ',i,': ');
   Readln (VEC[i]);
  End;
 SOM:=VEC[1];
 Med:=0;
 FOR i:=2 TO N DO SOM:=SOM+VEC[i];
 Med:=(SOM/N);
 Writeln ('La media è: ',Med:5:2);
 Readln;
End.
```