Analisi dei dati

corso sperimentale per Scienze turistiche/Specialistica

(prof. A. TARSITANO)

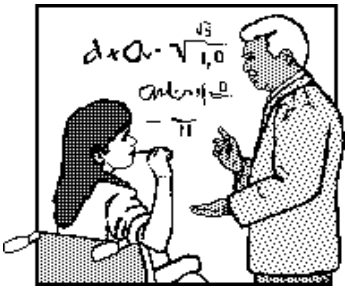
Le informazioni sul corso sono reperibili nel sito

http://www.ecostat.unical.it/Tarsitano/teds.htm

La statistica

La statistica è una scienza che raccoglie tutti i metodi e le tecniche che hanno come obiettivo

- **LA SCOPERTA**
- **LA NEGAZIONE**
- **L'ESTRAZIONE**



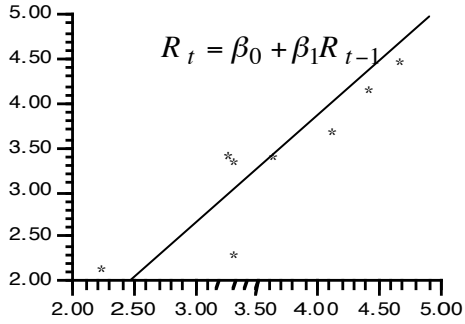
Del contenuto *informativo* di un insieme di dati

Uso della statistica

Le situazioni studiate dalla statistica sono reali ovvero sono connesse a fatti concreti

Il rendimento netto di un fondo azionario al tempo "t" indicato con R_t è legato a quello del tempo "t-1".

t	R_{t-1}	R_t
1990	2.10	2.26
1991	2.26	3.34
1992	3.34	3.29
1993	3.29	3.35
1994	3.35	3.65
1995	3.65	4.13
1996	4.13	4.42
1997	4.42	4.69



I valori mostrano un trend raffigurato da una retta. Stimati i parametri si potrà prevedere quale sarà il rendimento del prossimo anno, noto quello dell'anno attuale. Nel caso in esempio, per il 1998, si passerà da 4.69 a 4.63.

Dove non c'è statistica

Certamente la statistica ha poco a che vedere con gli "statisti" e con la "statica", ma ci sono anche altri casi

CASI LIMITE

Monterone (Co) è il comune più piccolo d'Italia (29 ab.) ed è una curiosità per gli studiosi di statistica.

I casi isolati o i casi singoli non interessano la statistica che infatti si presenta come scienza dei collettivi

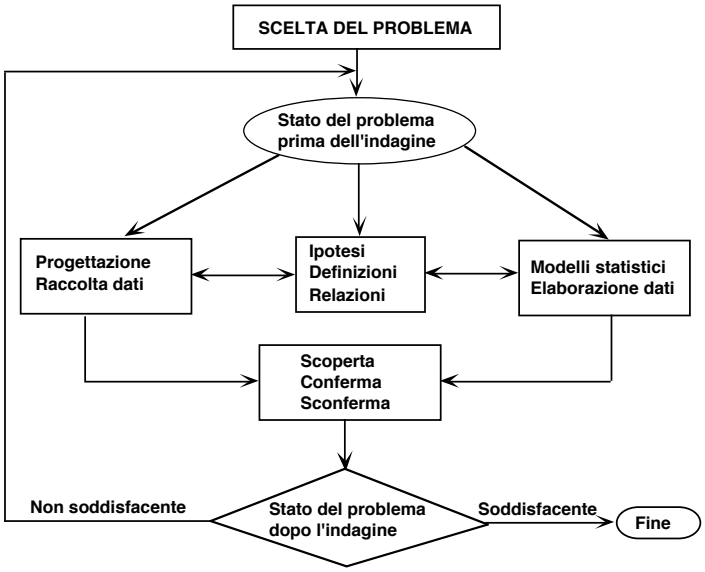
ACCOSTAMENTI FORZATI

La gazzetta dello sport ha un angolo intitolato: *"Per gli amanti della statistica"* dove riporta dati relativi ai precedenti incontri tra due squadre.

Non si capisce bene quale sia il collegamento se non che i due club hanno lo stesso nome

il paradigma di lavoro

Si effettua un'indagine statistica per dare sostegnoa teorie incerte



L'insieme delle conoscenze teoriche ed empiriche ed un SANO scetticismo aiutano a spiegare le variazioni tra due stati: prima e dopo l'indagine

L'indagine statistica

Se la trattazione del problema costringe a cercare nuovi dati, questi debbono essere rilevati con uno schema appropriato.

La rilevazione si articola in una sequenza ordinata di casi o repliche dette OSSERVAZIONI che hanno tanti elementi in comune da essere considerati facenti parte di un unico processo: l'indagine statistica.

L'osservazione è composta da DATI:

"ciascuno degli elementi di fatto (notizia, comunicazione, messaggio, rilevazione strumentale) utilizzabile per la soluzione di un problema"

Ogni indagine ha il suo piano di realizzazione legato alle peculiarità della disciplina in cui il problema è sorto.

Elementi costitutivi del Dato

La statistica è centrata sul dato che studiamo nei suoi elementi costitutivi:

- **L'UNITA' SU CUI E' RILEVATO**
- **LA VARIABILE STUDIATA**
- **LA SCALA DI MISURAZIONE**
- **IL CRITERIO ORGANIZZATIVO**

ESEMPIO

Nell'idea che i disavanzi delle aziende pubbliche si concentrino in particolari regioni a fianco c'è la tabella che li riporta, in milioni, per alcune regioni.

La caratterizzazione dei dati è ora: {Regione, Disavanzo, Milioni di lire, Ordinamento alfabetico};

Regioni	Disavanzo
Abruzzi	110558
Calabria	49991
Campania	2189901
Emilia R.	478704
Lazio	2739464
Liguria	378193
Lombardia	1111113
Marche	83445
Piemonte	342798
Puglia	360113
Toscana	562888
Umbria	143723
Veneto	600062
Totale	**9150955**

AT94

L'unità statistica

L'unità è il soggetto elementare cui l'indagine si rivolge: una persona fisica oggetto, azienda, o un gruppo di entità che, dal punto di vista dell'indagine, formino un tutt'uno.

Le unità devono essere obiettivamente distinguibili e deve pure essere stabilito quali siano quelle che interessa rilevare e quali debbano invece tralasciarsi.

ESEMPI

a) Interessi maturati su di un conto corrente **(Il conto corrente)**

b) Tipo di riscaldamento di un appartamento **(L'appartamento)**

c) Numero di testi consigliati in un corso **(Il corso)**

d) Emissione di gas tossici da un automobile **(L'automobile)**

f) Numero di arresti per agente di polizia **(L'agente)**

Problemi di definizione

INDAGINE SULLE FAMIGLIE

Come si considerano i "single", le coabitazioni, le comunità?

PUBBLICITA' TURISTICA

Non è raro leggere o sentire messaggi promozionali del tipo: 30 giorni di sole nel mese X. Il problema è capire cosa si intende per "giornata di sole": ad esempio nelle ore diurne una sequenza di almeno otto ore di sereno e senza nebbia.

SONDAGGI PRE-ELETTORALI

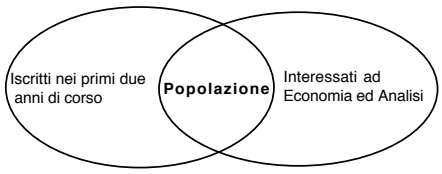
Un intervista telefonica agli abbonati di "La Gazzetta del Sud" può solo indicare come la pensano gli abbonati che hanno risposto alle telefonata.

La popolazione

Dicesi popolazione o UNIVERSO l'insieme di tutte e solo le unità che si è interessati ad osservare in una certa indagine.

ESEMPIO:

Alcuni studenti intendono finanziare le spese di frequenza universitaria avviando un programma di ripetizioni ben fatte ed a basso costo. Quale sarà la popolazione?



E' chiaro che non possono essere tutti gli studenti iscritti. Ci si può limitare agli studenti dei primi due anni.

Occorre poi determinare le materie per cui esistono le competenze: diciamo i corsi fondamentali di economia e matematica.

La delimitazione dell'universo è ora chiara: studenti del biennio che non hanno sostenuto economia e/o analisi.

La popolazione/2

NON è un gruppo di persone che risiedono in una certa zona.

il termine POPOLAZIONE ha una accezione più ampia e più astratta : tutte e solo le unità che hanno in comune una o più proprietà rilevanti per il problema.

La caratteristica unificante deve essere evidente cosicché il riconoscimento avvenga con il minimo di incertezza tenuto conto delle difficoltà create da unità congiunte o sfocate.

Se disponiamo di un elenco del fatturato di 500 imprese edili ciò che studiamo non è la popolazione delle imprese, ma la popolazione dei fatturati.

Microdati e macrodati

L'unità per cui si cercano i dati (unità di rilevazione) non sempre coincide con quella oggetto di studio (unità di indagine)

Esempio:
La rilevazione delle scuole materne può essere effettuata per comuni, ma essere poi elaborata per provincie

I microdati sono i valori riferiti all'unità elementare che non può essere ulteriormente scomposta.

I macrodati sono i valori ottenuti o direttamente o dalla aggregazione di più dati elementari.

I microdati sono un sistema di rilevazione comodo quando non si è sicuri della scala di aggregazione che poi potrà servire

La variabile

E' l'aspetto si intende studiare nel dato.

Può essere una distanza, una numerosità, una forma, un atteggiamento, un grado od anche una composizione di caratteristiche da trattare in modo aggregato.

I simboli più diffusi sono:

X, Y, W, Z

Che sono la codifica della variabile

La codifica è l'espressione abbreviata con cui le informazioni sulle variabili acquisite dalle unità sono trasferite sui supporti di elaborazione o nei ragionamenti astratti

La variabile/2

Perché una generica qualità o quantità sia definita "variabile" occorre...

- **ATTINENZA con la realtà di interesse la cui comprensione aumenta (anche di poco) per la disponibilità di dati sulla variabile.**

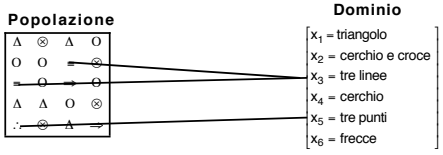
- **ESSERE SOGGETTA A VARIAZIONI: cioè possa presentarsi con almeno due valori o categorie distinte nell'ambito della popolazione.**

- **ESSERE ACCERTABILE e cioè essere rilevabile strumentalmente senza ambiguità**

Si presuppone inoltre che la variabile possa essere osservata/misurata in modo separato da altre variabili che pure incidono sull'unità.

il dominio della variabile

Individuata la variabile occorre definire l'insieme di tutti e solo i valori o modalità della variabile X (il dominio) riscontrabili nella popolazione:



Ad ogni unità della popolazione sarà associata una ed una sola modalità del dominio.

In questo caso, una delle sei diverse forme presenti. Unità diverse possono presentare la stessa modalità

il dominio della variabile è un insieme di "k" elementi con "k" finito od infinito

$$S=\{X_1, X_2, \ldots, X_k\}$$

L'abbinamento unità/modalità si effettua confrontando ciascuna delle unità è con il dominio "S" ed associandola ad una delle Xi in base ad una regola di classificazione o misurazione.

il dominio della variabile/2

Perché non insorgano ambiguità è necessario che le modalità siano

UNIVOCHE: sia possibile osservarne una sola per ogni unità e sia subito chiaro quale

ESAUSTIVE: non sia possibile osservarne di diverse da quelle già in S

RIPRODUCIBILI: la rilevazione dovrà dar luogo sembre allo stesso schema di attribuzione.

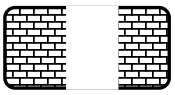
$$a)\ Incompatibilità: \quad X_i \neq X_j \ \ per\ ogni\ i \neq j\ e X_i, X_j \in S$$

$$b)\ Esaustività: \quad per\ ogni\ u \in P\ \ X(u) \in S;$$

$$c)\ Riproducibilità: \quad X = X_i \ \ se\ \ e\ \ solo\ \ se\ \ X(u) = X_i$$

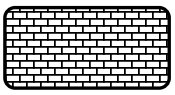
Dominio chiuso o aperto

L'insieme dei valori ammissibili "S" può essere



APERTO **Quando il fenomeno descritto non ha un limite minimo e/o massimo ben definito prima che sia completata la rilevazione**

Esempio: Reddito (che può anche essere negativo)



CHIUSO **Quando le sue modalità sono definite e note in anticipo e non possono cambiare durante la rilevazione**

Esempio: Stato civile

il dominio aperto comporta problemi di elaborazione. Quello chiuso consente dei controlli di validità dei dati

La definizione operativa

E' l'insieme di regole con cui classificare un concetto, determinarne la misura o, in generale, per aggangiarlo alla realtà osservabile.

cioè impone la definizione operativa solo con variabili di cui sia possibile seguire con facilità il meccanismo di conversione di una proprietà delle unità in una categoria o valore del dominio.

ESEMPIO

il concetto di interconnessione tra due centri abitati, diciamo "A" e "B" , è misurato con la semisomma degli automezzi che si sono spostati da "A" a "B" e quelli che da "B" sono andatti ad "A".

Classificazione e misurazione

L'acquisizione dei dati può avvenire classificando in categorie distinte la proprietà di cui l'unità è portatrice oppure misurandola in base ad una determinata unità di misura.

con la CLASSIFICAZIONE si identificano l'unità (e le modalità numeriche del dominio sono equivalenti ad ogni altro insieme di simboli);

con la MISURAZIONE si quantifica una proprietà posseduta ed i numeri sono utilizzati in quanto inseriti in un sistema di numerazione.

UN ATTEGGIAMENTO NON SELETTIVO

Classificazione e misurazione/2

La classificazione e la misurazione possono scaturire da due procedure di assegnazione dei valori: enumerazione delle unità rispetto alla proprietà posseduta

oppure comparazione della proprietà studiata rispetto ad un ventaglio di possibilità che, identico per tutte le unità, non dipende né dal numero.

ASSEGNAZIONE VALORI

RILEVAZIONE DEL DATO	**Enumerazione**	**Comparazione**
Classifizione	Nominazioni	Scala nominale
Misurazione	Graduatorie	Scale ordinali semplici Scale ordinali graduate Scale intervallari Scale proporzionali

Nominazioni e Variabili nominali

Le modalità di queste variabili esprimono categorie, qualità, status: le $\{X_i\}$ in "S" hanno la sola funzione di etichettare le unità per formarne un elenco o per raggrupparle in classi omogenee:.

ESEMPI:

Nominazione: La variabile "Regione" si manifesta con le usuali 20 modalità S={Calabria, Sicilia, ..., Val d'Aosta, Piemonte}.

variabile nominale: Un'impresa può ricadere nel settore {agricoltura, industria, altre attività}.

Le differenze possono essere accertate, ma non ordinate né misurate: si possono scambiar di posto senza che ciò influisca sulla validità della classificazione

Uso dei numeri

la codifica delle modalità porta ad usare dei numeri. Questo però non significa che siano lecite delle operazioni aritmetiche:

i ruoli di una squadra di calcio sono indicati con dei numeri, ma non si può dire che l'ala sinistra ("11") sia maggiore dello stopper ("5") o che l'unità di misura "1" dei calciatori sia il portiere;

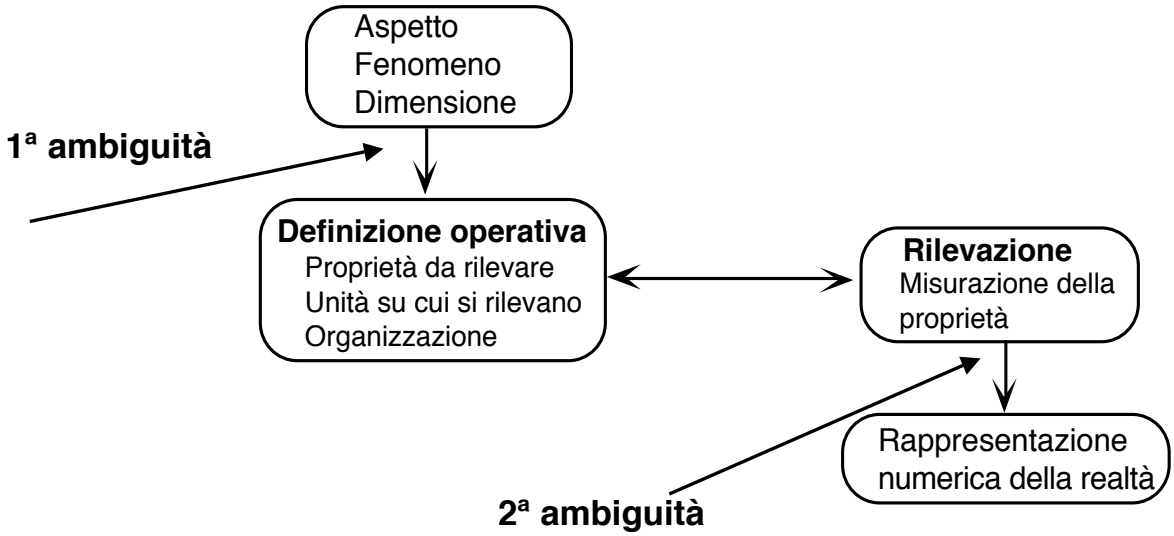
ESEMPI

il numero civico delle abitazioni:

Non ha significato la eventuale progressione delle modalità;

Tecniche di misurazione

il concetto di misurazione è uno dei più controversi tanto che oggi, dopo più di 50 anni, il dibattito è sempre aperto.



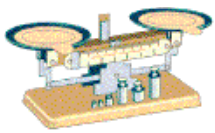
La scala di misurazione

Qui esiste una sovrapponibilità tra una categoria e la successiva che, oltre a contenere quella che la precede, vi aggiunge un quantum di proprietà che la differenzia dalla prima senza cancellarla, anzi inglobandola

ciò che distingue le scale di misurazione è il diverso grado di formalizzazione che si può dare al meccanismo dell'aggiunta del quanto di proprietà.

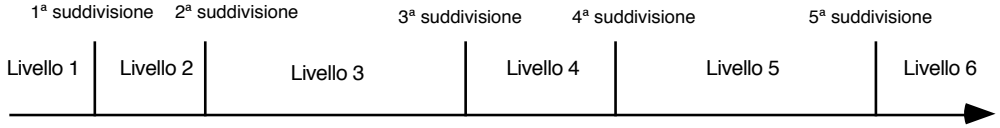
1. **Ordinamento tra valori senza distinguibilità degli scarti;**
2. **Ordinamento tra valori con ordinamento degli scarti;**
3. **Quantificazione dei valori con parità tra scarti: 7-5=3-1;**
4. **Quantificazione dei valori con parità tra rapporti: 8:4=6:3.**

Scala=bilanciamento



Continuo percettivo

L'intensità con cui si avverte una sensazione varia in una successione continua di stati: al tessuto uniforme del concetto si sovrappone una griglia più o meno regolare



una unità che sia X_i in una rilevazione ed X_j in una successiva con $X_i<X_j$ sarà passata per tutti gli stati intermedi tra X_i ed X_j.

Le suddivisioni non sono però oggettive: osservatori diversi scelgono divisioni diverse ovvero lo stesso punto di separazione ha senso diverso.

N.B. Talvota la proprietà studiata ha natura discontinua: si modifica con una scansione non frazionabile per un numero finito di stati che sono i soli a poter essere osservati.

Ordinamenti

Il termine "scala" ha senso se tra le modalità di "S" sono possibili degli ordinamenti.

$$1)\ X_i < X_j \quad oppure \quad X_i > X_j \quad per \quad ogni \quad i \neq j$$

$$2)\ X_i < X_j \Rightarrow X_i \neq X_j$$

$$3)\ X_i < X_j \quad e \quad X_j < X_k \Rightarrow X_i < X_k \quad per \quad ogni \quad i < j < k$$

Maggiore è il contenuto di "fenomeno" maggiore è la modalità che la rappresenta; esiste perciò una disposizione delle modalità che non può essere alterata senza che ne risulti modificata la rilevazione.

Il dominio si esprime con interi consecutivi:

$$S = \{a, a+1,\ a+2,\ \ldots, a+k-1\}$$

Graduatorie

Le modalità "S" sono i ranghi corrispondenti alle posizioni in graduatorie delle unità per i valori possibili sono dati dalla numerosità della rilevazione.

$$S = \{1^{\mathrm{a}},\ 2^{\mathrm{a}},\ \ldots, k^{\mathrm{a}}\}$$

Il processo di misurazione è ad un livello molto superficiale, con possibilità elaborative limitate, essenzialmente basate su confronto e sintesi delle posizioni che le unità occupano rispetto a variabili diverse.

ESEMPIO

Per stare in testa occorrono buone posizioni su entrambe le graduatorie

Stud.	Grad. Scritto	Grad. Orale	Totale
A	3	1	4
B	2	3	5
C	1	7	8
D	7	2	9
E	5	4	9
F	4	6	10
G	6	5	11

Variabili ordinali

i ranghi sono dei voti che esprimono la stima della proprietà posseduta: ogni unità è confrontata con una linea di valutazione che incasella l'unità in una data categoria di valore a prescindere da quello che succede alle altre unità.

Spesso, le modalità di una variabile ordinale esprimono soglie di vicinanza ad un ideale che fungerebbe da "metro" o "campione" di misurazione del concetto.

ESEMPI:

a) Voti di un giudice: S={0, 1, 2, …, 10};

b) Ammontare di punti da ripartire: {0 -100};

c) Voti grafici: {++, +0, 0+, 00, -0, 0-, - -};

d) Quantificatore verbale: { pianura, collina, montagna}

Invarianza rispetto a trasformazioni monotòne

$$f(X_i) < f(X_j) \quad se \quad X_i < X_j$$

Numero di modalità e Posizione

Non esiste un numero ottimale di livelli: k=7±2 o k=6 sono considerati uno standard nelle ricerche di mercato (Kinnear eTaylor, 1979, p. 30, Malhotra 1996, p. 298).

3 o 4 gradini comportano risultati confusi per l'accorpamento di giudizi eterogenei; d'altra più di sei è utile solo per acquisire variazioni di quantità molto piccole di cui non sempre si ha bisogno.

Anche la disposizione deve essere equilibrata:

ESEMPIO:

Quale delle tre seguenti moltiplicazioni

$P1.\ 9*7*8*6*5*4*3$

$P2.\ 3*4*5*6*7*8*9$

$P3.\ 7*3*8*4*6*5*9$

Effetto posizione

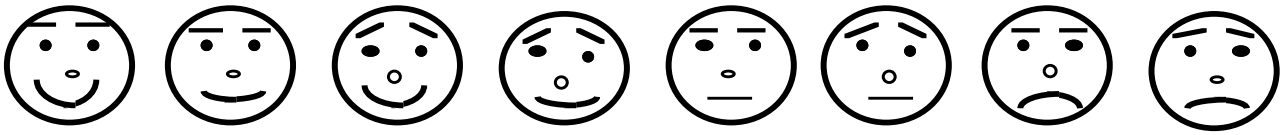
darà il risultato più alto?

Differenziale semantico

Per attenuare le ambiguità derivanti delle scale ordinali si possono usare delle scale bipolari in cui sono inserite solo le valutazioni più opposte dell'aspetto indagato collocando tra di esse, ad opportune interdistanze, una serie di riquadri

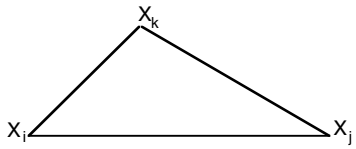
Chi risponde dovrà poi indiviudare il punto più prossimo al suo giudizio ovvero indicare quale descrizione numerica o verbale si adatti al proprio sentire.

ESEMPIO:
"Come giudicate l'operato dei rappresentanti degli studenti nel Senato accademico integrato?"



Scale metriche

Date tre qualsiasi modalit à di "S" allora



$$d(X_i, X_j) = 0 \text{ se e solo se } X_i = X_j; \qquad \textit{Identità}$$

$$d(X_i, X_j) > 0 \text{ se } X_i \neq X_j; \qquad \textit{Positività}$$

$$d(X_i, X_j) = d(X_j, X_i); \qquad \textit{Simmetria}$$

$$d(X_i, X_k) + d(X_k, X_j) \geq d(X_i, X_j); \qquad \textit{Disuguaglianza triangolare}$$

Se il dominio della "X" verifica le quattro condizioni allora su di esso si applicano, sia pure con qualche distinguo, tutte le procedure statistiche.

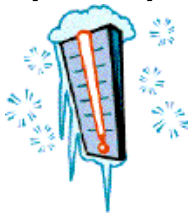
Scale intervallari

Sivaluta ciò che succede al fenomeno ponendolo in relazione con un movimento lungo un'asta graduata.

Le tacche sono regolari e separate -al livello minimo- da una unità convenzionale che può essere variata senza interferire con ciò che si misura.

L'origine o punto-zero della scala ha un ruolo marginale dato che agisce solo come riferimento e può essere sostituita, senza alcuna conseguenza sull'esito della rilevazione.

Un incremento assoluto tra due misurazioni ha lo stesso significato qualunque sia il livello da cui si calcola l'incremento.



Scale intervallari/2

La differenza tra 40C e 30C gradi è la stessa di quella tra 30C e 20C, ma non si può dire che ad 40C faccia due volte più caldo che a 20C. Se le temperature sono convertite in gradi Fahrenheit si avrà:

$$30\,C \rightarrow \frac{9*30}{5}+32 = 86F; \quad 40C \rightarrow \frac{9*40}{5}+32 = 104\,F; \quad 20\,C \rightarrow \frac{9*20}{5}+32 = 68F;$$

La differenza tra due temperature ha lo stesso significato qualunque sia il livello, ma nessuna asserzione può farsi sul loro rapporto dato che C=0 o F=0 non significa "totale assenza di calore".

Se la "X" è misurata su scala intervallare è lecito -se preferibile- usare in sua vece la variabile ottenuta come trasformazione lineare.

$$Y = a + bx; \quad b > 0$$

Scale proporzionali

Ad un incremento relativo nella misura, corrisponde un incremento relativo di eguale entità in ciò che si misura.

ESEMPIO

La misura di due centimetridi un segmento è -senza incertezze- il doppio di uno con lunghezza pari ad un centimetro;

se la misura aumenta del 50% anche il segmento si allunga di una estensione pari alla sua metà.

E' ammessa ogni trasformazione del tipo:

$$Se \;\; X_j \neq 0 \;\; \frac{X_i}{X_j} = c\left[\frac{f(X_i)}{f(X_j)}\right]; \quad per\; f(X_j) \neq 0$$

Variabili discrete

Derivano da un processo di conteggio o di numerazione:

riviste per numero di abbonati
partiti politici per numero di iscritti
catene commerciali per numero di punti vendita affiliati,

Le modalità sono presentate usualmente, ma non sempre, in ordine crescente:

$$X_1 < X_2 < \ldots < X_k$$

Il simbolo "<" ha qui il significato aritmetico di "minore".

La differenza tra due modalità ha significato costante, ma nulla si può dire sul rapporto tra di esse.

In modo alternativo si può dire che le modalità della variabile discreta possono essere contate ovvero poste in corrispondenza biunivoca con l'insieme dei numeri naturali.

Discrete frazionarie e dense

Una variabile può essere discreta, ma espressa con dei numeri decimali

ESEMPIO: lancio di due dadi. Modalità= semisomma dei punti sulle facce superiori. L'uscita di un "6" e di un "5" o di un doppio "6" danno luogo alle modalità

$$\frac{6+5}{2} = 5.5; \quad \frac{6+6}{2} = 6$$

La variabile è discreta perchè tra "5.5" e "6" la variabile non può assumere alcun valore. Le sue modalità sono tutte ISOLATE: E' sempre possibile trovare un intervallo, per quanto piccolo, che contiene una sola modalità.

La variabile DENSA è discrete per natura, ma ha una unità di misura è molto piccola rispetto all'ordine di grandezza con cui si manifesta

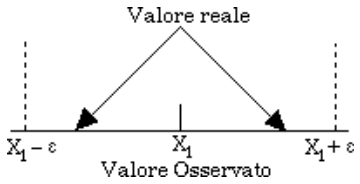
reddito in lire;
circolazione di vetture per numero di auto;
nazioni per numero di abitanti;

La trattazione dei caratteri densi è simile a quella dei caratteri continui

Variabili continue

Non possono essere rilevate puntualmente; il valore assunto è il centro dell'intervallo

$$\left[X_1 - \varepsilon, X_1 + \varepsilon\right]$$



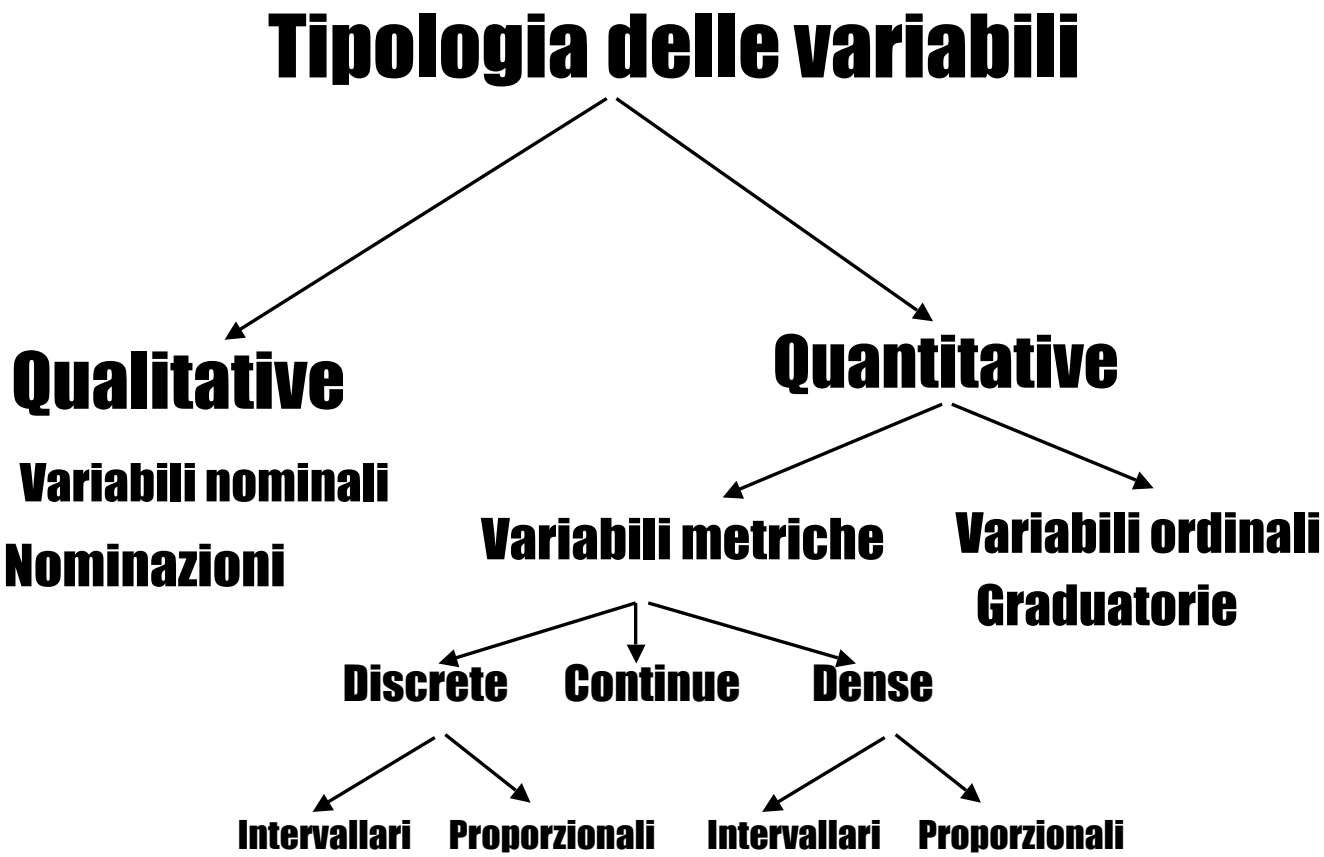
Dire che $X = X_1$ significa dire che $|X - X_1| < \varepsilon$ cioé che si è osservato uno qualsiasi degli infiniti valori compresi in

$$\left[X_1 - \varepsilon, X_1 + \varepsilon\right]$$

L'ampiezza del sottointervallo dipende dalla precisione degli strumenti di rilevazione. (Questo è un limite degli strumenti di misurazione non della variabile misurata)

Talvolta le modalità sono presentate come interi. Per distinguerle da quelle di una variabile discreta basta ricordare che:

Tra due modalità discrete non ve ne può essere un'altra, tra due modalità continue se ne trovano infinite



Criterio organizzativo

Ogni unità si inserisce in un contesto in cui si distingue e che consente di attribuirle la modalità corretta.

Le tecniche statistiche sono anonime e trascurano la localizzazione dell'unità rispetto alle altre.

In alcune analisi è necessario che l'unità sia ben collocata -nel tempo o nello spazio- ed il suo esame prima o dopo di un altra è rilevante.

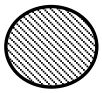
Gli ordinamenti possibili sono diversi, ma noi consideriamo solo

SERIE SPAZIALI (ordinamento geografico)

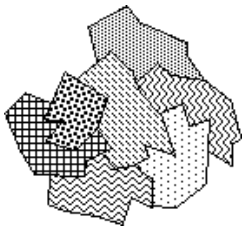
SERIE STORICHE (ordinamento temporale)

Le serie territoriali

Si ritiene che la modalità o intensità raggiunta dipenda dalla sua posizione topografica. Qui conta il tipo di unità considerata



UNITA' AREALI: rappresentata da una poligolane chiusa
entità fisiche: isola, lago, continente, etc.
entità amministrative: comuni, regioni, nazioni
entità funzionali: distretti sanitari, telefonici, scolastici



Le unità si considerano omogenee al loro interno anche se la rilevazione del carattere si effettua in più punti

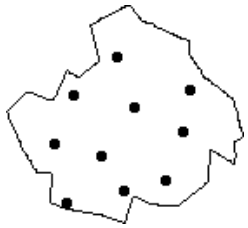
Esempio:
Percentuale di dipendenti pubblici sul totale occupati

Paese	%
Belgio	20.2
Danimarca	29.8
Germania	15.4
Grecia	10.4
Francia	22.8
Spagna	14.3
Regno Unito	19.5
Irlanda	17.9
Italia	17.4
Lussemburgo	11.6
Paesi bassi	15.1
Portogallo	14.1

Altri tipi di unità (territoriali)

UNITA' PUNTUALI: costituiscono i nodi di una maglia più o meno fitta di punti che coprono un dato territorio

misurazioni atmosferiche e idrogeologiche,
censimenti della popolazione
rilevazione della forza lavoro



Le unità puntuali hanno il grande pregio di visualizzare l'ubicazione delle modalità o intensità rivelandone la disseminazione o la concentrazione nel territorio

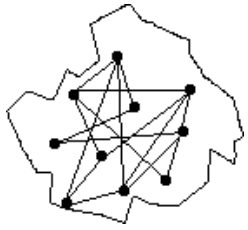
Esempio: Consumi di acqua per uso domestico

Città	Consumo
Bruxelles	108
Amburgo	146
Copenaghen	194
Londra	132
Parigi	147
Roma	220
Lussemburgo	171
Amsterdam	159
Madrid	158

Altri tipi di unità/2

UNITA' RETICOLARI (NETWORK): sono unità che si diramano nel territorio

fiumi, strade, gallerie
direttrici di sviluppo
rotte di navigazione
reti di distribuzione



La rilevazione dei caratteri sui network avviene per punti, in analogia alla osservazione di un flusso che percorre il reticolo (il flusso è spesso la variabile che si intende rilevare)

Esempio:
itinerari turistici calabresi per numero di pro-loco coinvolte

Itinerari	Pro-loco
Catanzaro-S.S.Bruno	5
Catanzaro-Capo Vaticano	7
Catanzaro-Laghi	3
Costa tirrenica	14
Sila e laghi	5
Costa Jonica	14
Aspromonte	5
Costa Viola	13
Magna Grecia	17

Le serie storiche

Spesso si studiano variabili nel tempo e la progressione cronologica dei valori è essenziale per comprendere il comportamento della variabile

VARIABILI DI FLUSSO: procedono in modo continuo. Se il flusso è regolare non importa l'unità di tempo. Se il flusso è erratico mette conto sapere che si rileva per settimane, mesi , trimestri, anni, etc.

Esempio:
Spesa dell'ammnistrazione statale per la cultura
te uto conto che ogni anno non arretra rispetto al periodo precedente

Anno	Spesa
1984	1 055
1985	1 206
1986	1 322
1987	1 954
1988	1 393
1989	1 064
1990	988
1991	947

VARIABILI DI STOCK: si manifestano in un dato istante per poi ripetersi più o meno regolarmente. Ricominciano da zero

Esempio:
Voti validi nelle consultazioni politiche

Anni	Totale
1948	26268912
1953	27092743
1958	29563633
1963	30758031
1968	31803253
1972	33414779
1976	36727273
1979	36671308
1983	36906005
1987	38592383

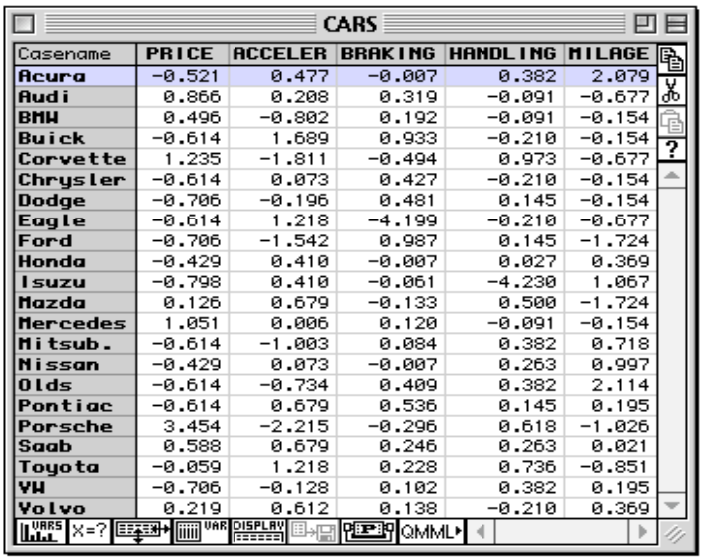
Esempio di data set su foglio elettronico.

Variabili e dati sul Piano integrato Territoriale (PIT) “Serre vibonesi”
N=24, m=13

Codice	NOME	SUP	POPRES	DENS99	VEC98	DIP98	LUADIP	TANALF	VPR9981	TIM	TIN	IMPRLA	TIMPR	DENSOC
101	Acquaro	2532	3018	119.2	104.1	63.7	13.4	14.6	-8.4	-14.4	2.2	47.2	29.4	38.2
106	Arena	3235	1983	61.3	102.1	62.5	15.0	14.3	-15.2	-4.6	0.2	47.9	24.8	30.6
114	Brognaturo	2450	801	32.7	82.5	66.2	16.5	4.9	-0.2	-10.2	3.8	42.0	25.2	57.9
116	Capistrano	2094	1244	59.4	118.9	61.8	8.0	15.7	-4.2	-6.4	0.6	42.1	22.0	26.6
141	Dasa'	619	1378	222.6	164.8	61.1	5.5	11.4	-14.0	-5.7	-2.9	45.4	50.2	71.2
144	Dinami	4406	3222	73.1	68.7	60.0	7.6	12.3	-0.9	-8.3	6.5	59.9	33.5	47.0
146	Fabrizia	3878	2776	71.6	95.8	63.7	6.0	15.6	-17.0	-15.0	2.4	55.5	39.5	58.4
149	Filadelfia	3048	6742	221.2	109.2	57.3	11.1	12.2	-20.6	-24.0	3.5	47.8	35.5	57.9
151	Filogaso	2369	1390	58.7	58.2	53.3	9.5	10.0	18.2	-4.7	6.0	72.1	39.9	97.2
153	Francavill	2825	2670	94.5	95.0	56.4	7.8	9.1	-12.4	-17.3	2.6	33.4	16.3	26.6
157	Gerocarne	4493	2633	58.6	78.6	58.8	7.9	14.1	-12.9	-23.5	5.5	44.8	29.9	38.2
179	Mongiana	2070	848	41.0	86.8	63.8	10.4	10.8	-14.2	-19.1	2.8	44.8	26.8	51.4
182	Monterosso	1816	2063	113.6	147.3	58.5	18.7	9.5	-11.2	-6.9	-2.7	53.2	47.4	64.3
184	Nardodipac	3278	1532	46.7	97.0	63.7	4.7	14.8	-25.8	-17.2	3.2	26.7	15.0	23.4
198	Pizzoni	2323	1440	62.0	128.8	63.3	10.5	16.8	-19.8	-14.9	-2.5	51.8	25.2	36.4
200	Polia	3178	1290	40.6	153.0	78.5	14.6	15.7	-16.9	-16.9	-2.4	48.8	28.2	53.6
212	San Nicola	1932	1727	89.4	164.7	76.0	18.0	16.8	-11.0	-6.4	-3.8	46.1	27.1	35.4
228	Serra San	3958	6894	174.2	106.4	52.8	17.8	9.3	8.2	-2.1	5.3	54.0	44.5	72.6
232	Simbario	1925	1139	59.2	130.4	72.3	20.1	9.6	-20.5	-8.1	-0.6	57.6	28.8	36.2
235	Sorianello	972	1682	173.0	62.0	55.9	10.2	14.7	-0.6	-8.7	8.5	71.3	31.3	57.2
236	Soriano Ca	1517	3154	207.9	77.7	52.7	15.9	8.7	1.6	-9.2	5.9	103.3	65.8	110.9
240	Spadola	958	818	85.4	116.8	54.8	20.3	7.6	6.1	-0.5	-0.7	45.0	69.7	99.3
252	Vallelonga	1753	852	48.6	138.5	66.9	15.0	15.2	1.5	-1.2	-2.8	40.6	30.5	36.3
253	Vazzano	1985	1283	64.6	131.3	56.7	14.1	9.6	4.4	-0.8	-2.8	56.4	31.8	44.2

Esempio di data set su pacchetto applicativo. STATISTICA

Caratteristiche di alcune automobili: m=5 variabili per n=22 unità.



CARS

Casename	PRICE	ACCELER	BRAKING	HANDLING	MILAGE
Acura	-0.521	0.477	-0.007	0.382	2.079
Audi	0.866	0.208	0.319	-0.091	-0.677
BMW	0.496	-0.802	0.192	-0.091	-0.154
Buick	-0.614	1.689	0.933	-0.210	-0.154
Corvette	1.235	-1.811	-0.494	0.973	-0.677
Chrysler	-0.614	0.073	0.427	-0.210	-0.154
Dodge	-0.706	-0.196	0.481	0.145	-0.154
Eagle	-0.614	1.218	-4.199	-0.210	-0.677
Ford	-0.706	-1.542	0.987	0.145	-1.724
Honda	-0.429	0.410	-0.007	0.027	0.369
Isuzu	-0.798	0.410	-0.061	-4.230	1.067
Mazda	0.126	0.679	-0.133	0.500	-1.724
Mercedes	1.051	0.006	0.120	-0.091	-0.154
Mitsub.	-0.614	-1.003	0.084	0.382	0.718
Nissan	-0.429	0.073	-0.007	0.263	0.997
Olds	-0.614	-0.734	0.409	0.382	2.114
Pontiac	-0.614	0.679	0.536	0.145	0.195
Porsche	3.454	-2.215	-0.296	0.618	-1.026
Saab	0.588	0.679	0.246	0.263	0.021
Toyota	-0.059	1.218	0.228	0.736	-0.851
VW	-0.706	-0.128	0.102	0.382	0.195
Volvo	0.219	0.612	0.138	-0.210	0.369

VARS | X=? | VAR | DISPLAY | QMML

Modello relazionale dei dati

Deriva dal concetto matematico di RELAZIONE

Noti gli insiemi $S_1,\ S_2,\ldots,S_m$ **coincidenti ognuno con un dominio**

"d" è una RELAZIONE se si configura come una "m-tupla" ordinata di valori

$$d = (d_1, d_2, \ldots, d_m)$$

tali che $d_1 \in S_1,\ d_2 \in S_2,\ \ldots,\ d_m \in S_m$

E' evidente che "d" coincide con una osservazione

"d" è un elemento del prodotto cartesiano di insiemi

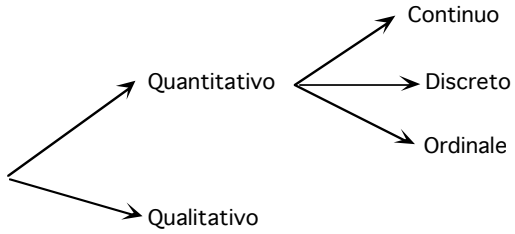
$$D = S_1 \otimes S_2 \otimes \ldots \otimes S_m$$

Che costituisce lo SPAZIO DEI DATI

Lo spazio dei dati

Su ogni unità si rilevano "m" variabili $X_1,\ X_2,\ \ldots\ ,\ X_m$

Ogni variabile ha un suo dominio $S_1,\ S_2,\ldots,S_m$



Si possono analizzare in tutto "N" unità (ma N può essere infinito)

$$P = \{U_1, U_2, \ldots, U_N\}$$

P è la popolazione (o universo) formata da tutte e solo le unità di interesse di Una ricerca

Su ogni unità è possibile rilevare un insieme di "m" informazioni detto vettore della osservazione

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{im}),\ i = 1, 2, \ldots, N$$

La matrice dei dati

Una rilevazione consiste nella osservazione delle variabili sulle unità

Le osservazioni sono i vettori $X_i,\quad i = 1, 2, \ldots, n$

I cui valori formano la MATRICE DEI DATI

ESEMPIO

Lo staff tecnico di una organizzazione è composto da 6 persone: Donne o uomini, laureate o no, residenti, vicini, fuori sede.

SPAZIO DEI DATI

(D,L,R,u1)	(D,L,V,u1)	(D,L,F,u1)	(D,L,R,u2)	(D,L,V,u2)	(D,L,F,u2)	(D,L,R,u3)	(D,L,V,u3)	(D,L,F,u3)
(D,N,R,u1)	(D,N,V,u1)	(D,N,F,u1)	(D,N,R,u2)	(D,N,V,u2)	(D,N,F,u2)	(D,N,R,u3)	(D,N,V,u3)	(D,N,F,u3)
(U,L,R,u1)	(U,L,V,u1)	(U,L,F,u1)	(U,L,R,u2)	(U,L,V,u2)	(U,L,F,u2)	(U,L,R,u3)	(U,L,V,u3)	(U,L,F,u3)
(U,N,R,u1)	(U,N,V,u1)	(U,N,F,u1)	(U,N,R,u2)	(U,N,V,u2)	(U,N,F,u2)	(U,N,R,u3)	(U,N,V,u3)	(U,N,F,u3)
(D,L,R,u4)	(D,L,V,u4)	(D,L,F,u4)	(D,L,R,u5)	(D,L,V,u5)	(D,L,F,u5)	(D,L,R,u6)	(D,L,V,u6)	(D,L,F,u6)
(D,N,R,u4)	(D,N,V,u4)	(D,N,F,u4)	(D,N,R,u5)	(D,N,V,u5)	(D,N,F,u5)	(D,N,R,u6)	(D,N,V,u6)	(D,N,F,u6)
(U,L,R,u4)	(U,L,V,u4)	(U,L,F,u4)	(U,L,R,u5)	(U,L,V,u5)	(U,L,F,u5)	(U,L,R,u6)	(U,L,V,u6)	(U,L,F,u6)
(U,N,R,u4)	(U,N,V,u4)	(U,N,F,u4)	(U,N,R,u5)	(U,N,V,u5)	(U,N,F,u5)	(U,N,R,u6)	(U,N,V,u6)	(U,N,F,u6)

Ciò che era possibile osservare

MATRICE DEI DATI

Persona	Sesso	Titolo	Residenza
u_1	D	L	F
u_2	D	L	V
u_3	D	L	V
u_4	D	L	R
u_5	D	N	R
u_6	U	N	V

Ciò che si è effettivamente osservato

Le dimensioni della matrice dei dati

La matrice dei dati ha dimensioni $(n \;\; x \;\; m)$

n è il numero di righe dove ogni riga (*record*) corrisponde ad una unità

m è il numero di colonne dove ognuna corrispondente ad una variabile

indagine sul *self-service* di una biblioteca

meta-dato

Matrice dei dati = *data set*

Insieme strutturato di informazioni

n=20
m=5

Nome	N.libri	Tempo	Posiz.	Corso	Giudizio
A.C.	6	6	Coll.		Medio
A.R.	10	6	4	DES	Medio
A.G.	6	11	Doc		Pessimo
A.T	5	1	FC	EA	Medio
D.I.	6	5	Dlp		Pessimo
D.S.	7	8	FC	SSA	Medio
F.D.	11	5	Doc		Ottimo
G.A.	1	4	2	DUS	Ottimo
G.G.	10	1	3	DES	Buono
G.L.	2	1	Est.		Medio
G.P.	8	6	4	SSA	Pessimo
G.S.	4	12	Imp		Cattivo
L.F.	2	7	1	EA	Cattivo
M.B.	8	8	Doc		Pessimo
M.P.	8	3	3	DEAI	Ottimo
P.A.	5	5	4	SSA	Medio
P.C.	8	2	FC		Medio
R.B.	6	4	2	DES	Cattivo
R.T.	1	4	2	EA	Buono
S.B.	5	2	Doc		Ottimo

I dati mancanti

I cosiddetti *missing values* sono quelli dovuti a non risposte insanabili.

Derivano anche da mancata rilevazione o rilevazione manife- stamente sbagliata.

L'elaborazione dei dati non consente vuoti nelle celle. Se mancano i dati si adotta un codice convenzionale

ESEMPIO

Numero di permessi sindacali concessi da ammininistrazioni pubbliche.

Le sedi che non hanno risposto sono indicate con "-99"

E' anche interessante capire il perché dei "missing values"

Rilevazione dei dati

133	197	165	214	188	237	188	115	128	213	120
204	**-99**	232	230	236	149	153	112	68	117	153
94	72	222	220	139	219	144	137	98	80	**-99**
209	93	181	249	200	128	82	**-99**	103	182	156
71	182	199	126	127	187	185	87	177	94	92
145	115	**-99**	203	233	64	227	88	67	243	240
204	156	118	**-99**	91	115	243	74	192	74	**-99**
197	245	235	88	141	116	168	204	62	**-99**	128
242	67	130	158	184	114	232	122	70	122	72

Analisi univariata e multivariata

Ogni problema è una ragnatela: se si tocca un filo tutti gli altri vibrano. Lo stesso sucede per le variabili.

Lo studio univariato ha solo scopo didattico. Nella pratica i dati sono sempre multivariati

ESEMPIO: dove vanno gli studenti

		Stessa regione							
		Nord		Centro		Sud		Totale	
		numero	%	numero	%	numero	%	numero	%
		286555	83.6	178692	90.7	253887	74.7	719.124	81.8
	Nord Ovest	18783	5.5	1526	0.8	8378	2.5	28687	3.3
	Nord Est	27308	8.0	4749	2.4	11312	3.3	43369	4.9
Altre	Centro	9149	2.7	9396	4.8	38800	11.4	57345	6.5
regioni	Sud	929	0.3	2756	1.4	27296	8.0	30981	3.5
	Totale	56169	16.4	18427	9.3	85786	25.3	160382	18.2
	Italia	342724	100.0	197119	100.0	339663	100.0	879506	100.0

La lettura di una tabella a più variabili non è difficile.
Lo è la generalizzazione dei risultati

Gli studi multidimensionali sono al momento rinviati. Faremo solo studi univariati.

Col presupposto che si possa avere l'idea di un concetto multilaterale studiando separatamente le sue componenti

I metadati

Sono codici che identificano in modo sintetico e senza ambiguità le unità

Esempi:

Se si tratta di persone il record include nome e cognome e altre informazioni età, sesso, professione

Nel caso di imprese: settore produttivo, forma societaria, dipendenti, sede degli stabilimenti.

Per daati territoriali è inserito il riferimento geografico delle unità.

I metadati sono dei dati per accedere ad altri dati. Sono il mezzo di contatto tra riclevazioni diverse sulle stesse unità

La codifica

Le denominazioni delle modalità sono talvolta lunghe o espresse con termini scomodi che complicano il ragionamento.

Si stabiliscono abbreviazioni (codifica) per facilitarne la trattazione informatica e saranno poi queste a comparire nella matrice dei dati.

ESEMPIO:

In una indagine internazionale sulla distribuzione dei redditi, il grado di copertura della popolazione di cui si sono considerate le entrata venne rilevata con il dominio S={NL, URB, NAG, RRL, AG} che sono abbreviazioni di *{national, urban, nonagricultural, rural, agricultural}*

La codifica è utile per sveltire le operazioni di trasferimento dei dati dai moduli con cui sono acquisite (questionari, schede di richiesta, fogli di controllo, etc.) e per limitare le sviste nella trascrizione.