Convento dell'Osservanza
Radicondoli

ETHOIKOS

Corsi Ethoikos primavera 2010

Quattro lezioni di statistica:

come diavolo si analizzano 'sti dati?

24 - 25 aprile

Docente: Alessandro Giuliani

Corso di approfondimento sulle tecniche di elaborazione statistica e valutazione dei dati raccolti nel corso di indagini scientifiche nel campo dell'etologia, ecologia, psicobiologia, antropologia, biomedicina.

Programma	
Lezione 1: Teoria della misura • Concetto di scala di misura (nominale, ordinale, intervallare) • Serie temporali e spaziali • Limiti di validità delle misure • Quantità di informazione • Scale di grandezza e normalizzazione • Matrici di dati • Principali indici statistici	Lezione 2: Correlazione • Correlazione bivariata • Misure di correlazione • Correlazione nello spazio e nel tempo • Autocorrelazione e cross-correlazioni • Ritmi e stagionalità • Spazi multidimensionali
Lezione 3: Inferenza statistica • Distribuzioni di probabilità • Campionamento • Test parametrici e non parametrici • Ipotesi semplici / Ipotesi complesse • Inferenza e correlazione • Regressione • Modelli matematici	Lezione 4: Analisi multivariata • Componenti principali • Spazi di distanze • Analisi dei cluster • Analisi multivariata di serie temporali • Analisi multivariata di dati spaziali

Appunti di statistica

Alessandro Giuliani

Istituto Superiore di Sanita'

Cap. 1: *Misure*

Se ci chiedessero cosa distingue l'attivita scientifica da tutte le altre attivita', dandole in qualche modo un "marchio di fabbrica", dovremmo certamente rispondere: "l'uso di quantita'misurabili".

Un po' come i critici cinematografici che considerano il cosiddetto "specifico filmico" essere l'attivita' di montaggio e concludono (a ragione) che il codice cruciale del linguaggio cinematografico è la giustapposizione delle immagini nel tempo, allo stesso modo l' attivita' del misurare e' l'aspetto cruciale ed il codice primario di ogni produzione scientifica.

E' quindi quasi una scelta obbligata iniziare il nostro giro esplorativo della metodologia statistica esaminando quale sia il significato dell' attivita' di misura e come l'atto della misura costituisca un filtro necessario e fortemente indirizzante il carattere della conoscenza scientifica.

L' atto del misurare potrebbe schematizzarsi come segue:

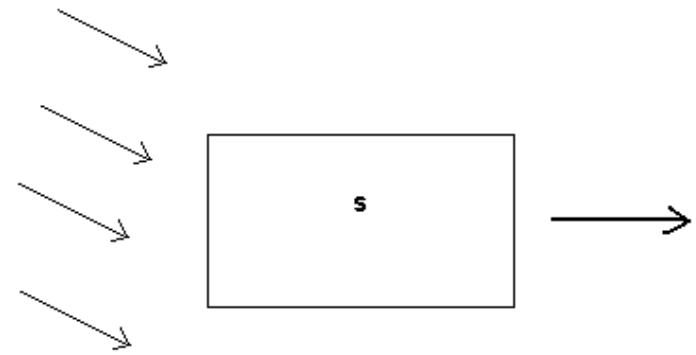


Figura 1.1
Le frecce incidenti indicano un insieme eterogeneo di ingressi esterni costituito da diverse caratteristiche e condizioni (di per se' inconoscibili) del sistema sottoposto a misura che, sotto l'azione del dispositivo S (rappresentato da una scatola rettangolare) concorrono a generare un' unica uscita (la freccia spessa) costituente la misura. Questa risulta quindi l'esito finale dell' interazione tra mondo e dispositivo di misura e fa "collassare" l'iniziale molteplicita' del rappresentato in un' unica rappresentazione sintetica.

La riduzione di complessita' insita in questa mappatura "da tanti ad uno solo", la quale fa si' che stati inizialmente differenziati del mondo risultino, dopo l'operazione di misura, assolutamente indistinguibili, lungi dall'essere una limitazione del discorso scientifico, ne costituisce al contrario la principale forza. Solo in questo modo infatti noi possiamo rinchiudere l'universo del discorso entro limiti finiti e quindi stabilire univocamente se un oggetto A e' piu' simile ad un altro oggetto B piuttosto che ad un oggetto C, D, E. La compressione del molteplice su una ben definita direzione consente invece di impostare correttamente il problema della relativa similitudine degli oggetti rappresentati.

Quale e' allora la natura della compressione operata sulla realta' dal dispositivo di misura che ci permette di esprimerci in maniera comparativa ? La scatola S nello schema rappresenta un insieme di regole, una "ricetta" piu' o meno

complicata per operare delle trasformazioni sui dati in ingresso e trasformarli in grandezze quantitative. Nei casi piu' semplici questo insieme di regole si riduce ad una legge fisica.

Consideriamo la misura della temperatura corporea attraverso i termometri a mercurio che si comprano in farmacia. In questo caso, S non e' altro che la relazione empirica esistente tra la dilatazione lineare dei materiali e la temperatura, espressa da una legge del tipo:

$$Y(t) = Y(t0) + a*t \quad [1.1]$$

Dove Y(t) e' la lunghezza della barretta di mercurio corrispondente alla temperatura t, Y(t0) la lunghezza della barretta a riposo (prima di mettere il termometro sotto l'ascella), t la temperatura effettiva all'interno del corpo e *a* una costante di proporzionalita' specifica per ogni materiale (alcool, mercurio ecc.), che mette in relazione l'allungamento (espresso in millimetri attraverso piccole tacche disegnate sul vetro) e la temperatura (espressa in gradi centigradi). Ovviamente, ogni tacca e' rappresentata in gradi per consentire una semplice lettura dello strumento: in un certo senso la legge che lega temperatura e dilatazione e' materializzata nella scala graduata del termometro e assunta per vera, cosi' che noi non ci dobbiamo preoccupare di tutte queste sottigliezze nel normale uso del termometro.

A noi pero' ora non interessa vedere se abbiamo la febbre, bensì usare il termometro come iniziale guida nel concetto di misura, per cui analizzeremo le considerazioni implicite nella formula [1.1]. Per cominciare, passiamo da una rappresentazione algebrica ad una grafica della [1.1], descritta nella figura 1.2.

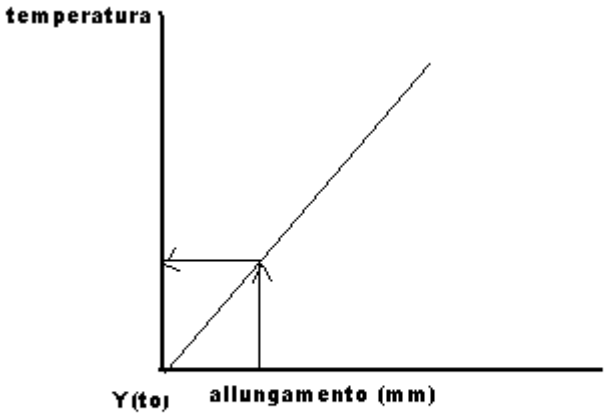


Figura 1.2
Conoscere la relazione [1.1] consente di risalire graficamente , alla temperatura corrispondente ad un certo allungamento della barretta di mercurio come descritto nella figura dalle frecce. Nel caso dei termometri questa relazione si da' una volta per tutte e corrisponde alla scrittura della scala graduata sul vetro.

Tutto molto semplice, tanto piu' che il carattere lineare della relazione implica che, a meno di un valore iniziale Y(t0) il rapporto tra allungamento e temperatura e' fisso (corrisponde al valore *a* della equazione [1.1]), e quindi possiamo pensare di variare come desideriamo la precisione della nostra misura (tralasciando per un momento il fatto che tacche troppo ravvicinate sarebbero di difficile lettura). La legge descritta nel grafico e' detta "legge di regressione lineare" ed avremo modo di incontrarla spesso nel seguito.

Una legge di questo tipo puo' essere "materializzata" e quindi resa operativa sostituendo alle lettere dei numeri:

$Y(t) = 10 + 3*t$ [1.1'].

Dove 10 e' la lunghezza in mm della barra a riposo e 3 la costante di dilatazione lineare, per cui, attraverso la semplice applicazione della formula [1.1'], e' immediato costruire la tabella di conversione seguente:

gradi	**mm**
35	115
36	118
37	121
38	124
39	127
40	130
41	133
42	134

Risulta allora abbastanza ovvio che l'applicazione continuata della [1.1'] consente di risalire esattamente alla temperatura data una qualunque misurazione in millimetri.

Dove sono i limiti della faccenda ? Insomma, fino a dove mi posso spingere ? Questo problema e' espresso dagli scienziati come la ricerca delle "condizioni al contorno" che garantiscono l'ambito di validita' della legge alla base della misura, l'ambito cioe' delle prestazioni del dispositivo S. Un primo estremo delle condizioni al contorno e' gia' presente esplicitamente nella formula [1.1] e corrisponde al valore Y(t0) del termometro a riposo (dopo averlo coscienziosamente sbatacchiato, naturalmente). Questo valore segnera' un limite inferiore di temperatura sotto il quale il mio dispositivo non misura piu'. In altri termini possiamo dire che sotto Y(t0), per limiti fisici legati alla particolare struttura del termometro, la legge empirica lineare di dilatazione dei materiali non e' piu' apprezzabile e quindi la nostra misura impossibile. L' altro estremo dell'intervallo di funzionamento corretto del termometro e' naturalmente legato alle dimensioni finite del termometro stesso, per cui, ad esempio, un allungamento di piu' di 20 cm risulta impossibile. La finitezza del dispositivo definisce quindi univocamente le condizioni al contorno della particolare misura e l'universo del misurabile. All'interno di questo universo noi possiamo dedicarci alla misura della temperatura di liquidi dentro bacinelle, della temperatura corporea dei nostri cari, dell'acqua della vasca e cosi' via riuscendo ad inserire in uno stesso quadro di riferimento oggetti assolutamente disparati come un succo di frutta, gli esseri umani e l'acqua del bagno, e potendo cosi' tranquillamente dire a nostra figlia di tuffarsi nella vasca in quanto la sua temperatura corporea e' leggermente piu' bassa della temperatura dell'acqua o di aspettare a bere il succo di frutta perche' e' ancora troppo freddo essendo appena uscito dal frigo. E' chiaro allora che mentre non ha alcun senso comparare figlie, succhi di frutta e vasche piene d'acqua senza altra specificazione, il confronto di questi enti *sub specie* della loro temperatura acquisisce un' immediata valenza operativa e questa e' proprio l' utilita' della riduzione di complessita' operata dall'operazione di misura.

Immaginiamo adesso di poter costruire, grazie ad espertissimi maestri vetrai di Murano, un termometro lungo alcuni chilometri che, con estrema cautela, verra' sistemato in una vasta e disabitata pianura. Poniamo la punta di questo gigantesco termometro a contatto termico con una sorgente di calore da noi regolabile a piacere (un'immensa caldaia, una centrale termonucleare..) e vediamo, avendo estremamente dilatato i limiti fisici imposti dalla lunghezza dello strumento, fino a che temperatura riusciamo a misurare. Se veramente mettessimo in pratica questo costoso (ed oltremodo pericoloso) esperimento, noteremmo che, molto prima di arrivare ai limiti di lunghezza del termometro gigante, un'altra, ben piu' drastica, condizione al contorno ci fermerebbe: oltre una certa temperatura il mercurio comincerebbe a dilatarsi con una legge diversa e l'equazione [1.1] non sarebbe piu' valida: la sostanza, infatti, cambierebbe lo stato fisico rendendo completamente priva di senso l'idea di misurarne la dilatazione lineare. Insomma, ogni tipo di misurazione e' applicabile entro i limiti stabiliti dal particolare dispositivo di misura adottato, il che pone alla nostra attivita' di misura limiti sia estrinseci che intrinseci: al di fuori di questi limiti la misura semplicemente cessa di essere valida. La finitezza dei limiti di ogni misurazione fissa un orizzonte del discorso scientifico (di quel particolare discorso scientifico) che da' alla scienza il suo sapore particolare, il suo carattere metrico, la possibilita' di stabilire confronti.

Il discorso del termometro e' estensibile a strumenti di misura molto diversi, in cui il dispositivo non e' piu' guidato da semplici leggi fisiche ma da complesse interazioni descrivibili solamente attraverso regolarita' statistiche. A questo riguardo ci riferiremo ad un nostro lavoro scientifico il cui scopo era quello di investigare le possibili cause della marcata variabilita' spaziale sul territorio europeo delle differenze tra i sessi nell'incidenza dei tumori. Il primo problema era quello di costruire un indice di misura sensibile del fenomeno che si desiderava studiare, dove per sensibilita' si intende l'efficienza del dispositivo nel registrare variazioni anche piccole della grandezza d'interesse. Nel caso del termometro, questo significa scegliere un materiale con un elevato valore della costante di dilatazione 'a'; nel caso della nostra indagine, ciò corrisponde ad un indice statistico che tenga conto delle variazioni tra i sessi dell'incidenza di tutte le patologie tumorali, senza essere "distratto" da eventi accessori che ne avrebbero diminuito la potenza.
Questo indice, da noi denominato ΔN e calcolato per ogni regione europea di cui erano a disposizione dati affidabili di incidenza di tumori, e' stato definito come:

$$\Delta N = (Pm - Pf) / Pm \qquad [1.2]$$

Nella [1.2] Pm = Incidenza di tumori nella popolazione maschile per 10000 abitanti, Pf = Incidenza di tumori nella popolazione femminile per 10000 abitanti

Osservando la formula (1.2) possiamo immediatamente notare alcune proprieta' dell' indice ΔN comuni a tutti gli strumenti di misura. La prima proprieta' e la riduzione di complessita' apportata dall'indice sulla realta' in esame: nei registri tumori sono raccolti i dati di incidenza relativi a circa 50 tipi di diverse patologie tumorali per i due sessi: i valori di Pm e Pf , quindi, costituiscono la risultante di circa 100 diverse sorgenti di informazione collassate in due soli indici. Questa drastica riduzione di informazione "alla sorgente", e' giustificata dal fatto che eravamo interessati alla valutazione di un indice sintetico delle differenze tra i

sessi nelle patologie tumorali e non ad una disamina accurata di ogni singolo tipo di tumore.

Esaminando la formula [1.2] notiamo poi che al numeratore è riportata la differenza fra le incidenze dei tumori nei due sessi (Pm – Pf) ed al denominatore l'incidenza nei maschi (Pm). Perche' non basarsi semplicemente sulla differenza (Pm – Pf) piuttosto che inserire l'ulteriore complicazione di dividere il risultato per Pm ? Il motivo e' che ci si voleva concentrare sulle differenze fra i sessi nelle diverse aree senza essere distratti da altri effetti che avrebbero potuto mascherare la significatività del risultato. Il piu' importante di questi possibili effetti di mascheramento ("confondenti", nella letteratura statistica) e' la diversa incidenza assoluta di tumori nelle differenti aree esaminate. Il denominatore dell' espressione [1.2] funge insomma da unita' di misura che vincola l'indice a variare in uno spazio finito in cui la presenza di limiti automaticamente indotti dalla stessa struttura della formula (e quindi dalla logica di funzionamento dello strumento di misura) ci consentono di apprezzare a prima vista l'entita' del fenomeno studiato. Questa operazione e' detta "normalizzazione" e consiste nell'imporre una scala predefinita alla nostra misura che ne regoli i limiti di variazione entro un regime ottimale per gli scopi conoscitivi prefissati. Nel nostro caso, la completa indifferenza tra le incidenze nei due sessi porta ad un valore pari a zero dell'indice ΔN; la deviazione verso valori via via piu' elevati indica uno squilibrio verso una maggiore incidenza nella popolazione maschile, laddove valori di ΔN negativi indicherebbero una maggiore incidenza nella popolazione femminile. L'applicazione dell' indice sulle 48 aree selezionate non ha mai evidenziato una maggiore incidenza nelle femmine, ha invece mostrato delle grandi disparita' tra aree, andando da una sostanziale parita' fra sessi nei paesi Scandinavi (Danimarca, $\Delta N = 0.06$, Svezia, $\Delta N = 0.07$) ad un marcato squilibrio verso una incidenza maschile superiore di quasi il 50% rispetto a quella femminile in alcune aree francesi (Calais, $\Delta N = 0.47$, Somme, $\Delta N = 0.46$).

Quello che qui conta sottolineare, comunque, e':

- il procedimento di costruzione di uno strumento di misura, che deve rispondere ai requisiti imposti dal carattere del problema da risolvere,
- come il suddetto procedimento sia parte integrante del lavoro scientifico e definisca il punto di vista con il quale lo scienziato guarda al reale.

In altre parole, una volta scelto un dispositivo di misura, il nostro sguardo restringe la sua visuale ma diventa piu' acuto. La misura organizza il mondo in una direzione, spezzando la simmetria dei vari punti di vista. Torneremo piu' approfonditamente su questo punto nei successivi capitoli; qui basti sottolineare lo stretto legame tra misura e ricerca di asimmetrie, cioe' di ordinamenti di oggetti (in questo caso aree geografiche) lungo un cline che li differenzi secondo una caratteristica specifica, allontanandoli dall' equivalenza.

Cap. 2: *Forme e Distanze*

Nel capitolo precedente avevamo introdotto il concetto di spazio metrico come generato da una definizione operativa di misura. La costruzione di uno spazio metrico dava vita ad un sistema di riferimento in cui era possibile farconfronti tra differenti oggetti e scoprire in maniera univoca chi fosse piu' vicino a chi. Ora si tratta di precisare ulteriormente la definizione di

metrica, mostrando come sia strettamente legata all' idea intuitiva di forma. Per procedere ben saldi su un terreno a tratti scivoloso, e' importante definire univocamente gli oggetti di cui si parla, cosi' che quando parleremo di "distanza tra oggetti" o "forma di un oggetto complesso" possiamo avere ben chiaro a cosa ci stiamo riferendo.

Qualsiasi argomentazione scientifice si riferisce allo stesso materiale di partenza: una matrice di dati. La matrice dei dati, nella sua forma piu' generale e' una tabella rettangolare che presenta nelle righe le unita' statistiche (gli oggetti "atomici", non ulteriormente divisibili, della particolare opera scientifica) e, nelle colonne, le variabili, cioe' le misure eseguite sulle unita' statistiche.
Di seguito e' riportato un esempio di matrice di dati.

Tabella 2.1: Indicatori socio-economici di alcuni paesi balcanici e mediterranei

Nazione	popolazione	Speranza di vita	Tasso alfabetizzazione	PIL/abitante
Israele	6	77.5	95	16.7
Cipro	0.7	77.2	94	13.4
Malta	0.4	76.5	91	13.3
Grecia	10	77.9	97	11.6
Slovenia	2	73.2	96	10.6
Libia	5.4	64.3	76	6.4
Algeria	28	68.1	62	5.6
Turchia	61	68.5	82	5.5
Siria	14	68.1	71	5.4
Tunisia	9	68.7	67	5.3
Libano	4	69.3	92	5
Bulgaria	8	71.2	98	4.6
Romania	23	69.6	98	4.4
Giordania	4	68.9	87	4.2

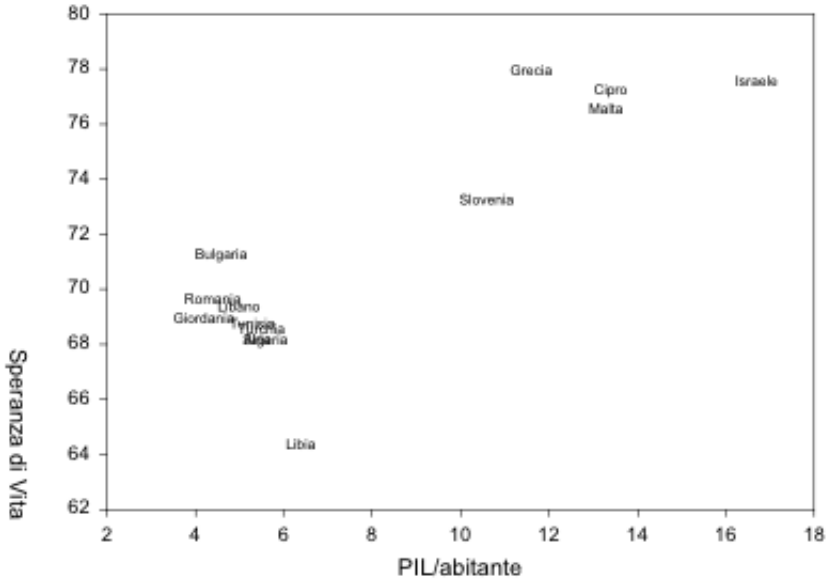
Osservando la tabella si puo' notare come le unita' statistiche rappresentino le osservazioni che definiscono il piano di base dell'analisi: eventuali regolarita', similitudini, correlazioni derivabili dai dati avranno come livello elementare il livello delle unita' statistiche, la tabella potra' insomma dirci qualcosa sull' esistenza di paesi poveri e paesi ricchi e su come questo si lega al grado di alfabetizzazione e all'attesa di vita.

Essenzialmente si tratta di individuare all'interno del nostro campo di dati delle aggregazioni notevoli (cluster) di oggetti che diano una base geometrica alle nostre categorizzazioni. Per fare cio' e' importante ritornare su un concetto molto fondamentale: la distanza.

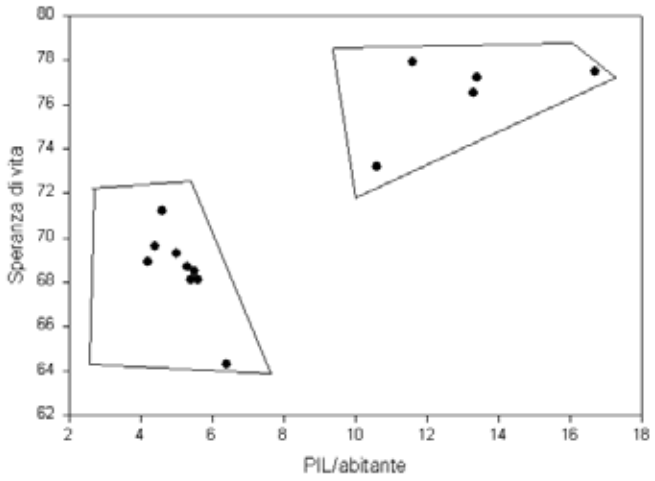
La distanza ha un significato intuitivo, quello del segmento che unisce due punti all'interno di uno spazio di riferimento. Per comprenderlo trasportiamo le colonne "Speranza di vita" e "PIL/abitante" della tabella in un riferimento cartesiano, sostituendo posizioni lungo le due direzioni dello spazio ai valori numerici delle variabili: a questo punto le unita' statistiche diverranno i punti di un grafico come quello in figura 2.1.

Dalla semplice osservazione della figura possiamo trarre indicazioni interessanti basandoci sul concetto intuitivo di distanza: in basso a sinistra nel grafico c'e' un gruppo di nazioni con bassa speranza di vita e basso prodotto interno lordo costituito da Bulgaria,

Romania, Libano, Giordania,Tunisia, Turchia, Siria, Algeria e, in posizione piu' defilata, la Libia, a destra in alto ci sono le nazioni relativamente piu' ricche (i dati si riferiscono al 1995), con la Slovenia in posizione un po' appartata.



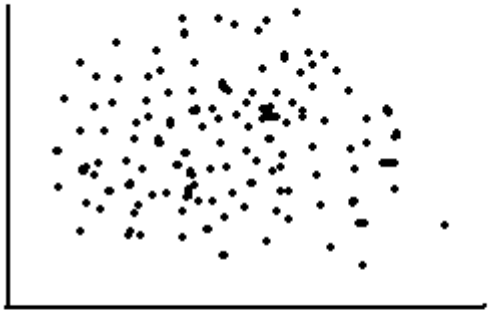
Insomma possiamo sicuramente rispondere con un si' alla domanda sull'esistenza di paesi ricchi e paesi poveri. Cosa abbiamo fatto in maniera intuitiva ? Semplicemente il nostro occhio e' stato attratto da "aggregazioni notevoli" di punti separate da ampi tratti vuoti. Queste aggregazioni notevoli sono state da noi considerate "gruppi omogenei" a cui appare naturale assegnare un'etichetta: il nome "povero" o "ricco" e' stato suggerito dal significato delle variabili che costituiscono gli assi cartesiani del piano. Il vuoto e' stato considerato come la discontinuita' che rendeva legittimo dare alla pura determinazione quantitativa del PIL e della speranza di vita un senso qualitativo in termini di poveri e ricchi. Questo e' ancora piu' evidente se eliminiamo le indicazioni esplicite dei nomi delle nazioni come nella figura 2.2:



Questa operazione e' alla base del nostro concetto intuitivo di forma e si basa sulla segmentazione dell' immagine in aree omogenee cosi' da poter dare un senso "oggettivo" al nome loro assegnato. In qualche modo questa e' la base dell' assegnazione di "senso" alle nostre misure : la presenza di un netto passaggio di scala (la discontinuita' tra le aggregazioni) fornisce la motivazione alla classificazione e quindi alla generazione di categorie.

Ovviamente questa e' una procedura completamente automatica, che puo' essere facilmente eseguita da un personal computer su qualsiasi tabella di dati in modo da offrirne "la migliore classificazione" in termini puramente strutturali.

Immaginiamo allora di dover giudicare di una situazione come quella riportata nella figura 2.3:



A differenza del caso precedente, qui l'assegnazione di classi omogenee, anche se tecnicamente possibile, risulta essere evidentemente arbitraria. L'arbitrarieta' e' legata al fatto che piccoli spostamenti della disposizione dei punti (causati ad esempio da un errore di misura) avrebbero un grande effetto sulla disposizione dei confini tra le classi. In questo caso non ci sono discontinuita' che garantiscano l'esistenza di classi significative, resistenti a piccoli spostamenti degli oggetti, e scompare un punto di vista privilegiato per tracciare i confini: non esiste alcuna scelta univoca di categorizzazione suggerita dai dati.

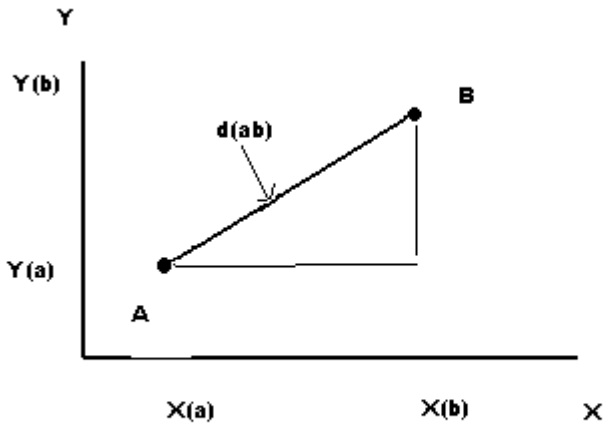
Qual'e' la differenza fondamentale tra questo ultimo esempio ed il caso della distribuzione delle nazioni nello spazio PIL/abitante vs. Speranza di vita ? Semplicemente che in questo caso lo spazio e' assolutamente omogeneo: ogni zona del piano e' piu' o meno identica all'altra, mentre nel caso della distribuzione delle nazioni, la distribuzione dei punti era fortemente asimmetrica, dividendo nettamente il piano in "zone abitate" ed in "zone deserte". Di nuovo, ci accorgiamo come l'asimmetria sia alla base della nostra conoscenza del mondo, della nostra possibilita' di dire qualcosa di sensato sulla realta'. L'asimmetria corrisponde alla possibilita' di comprimere l'informazione: dire che ci sono paesi ricchi e paesi poveri infatti implica una notevole compressione dell' informazione iniziale rappresentata dall' elenco dei valori delle due variabili considerate per i differenti paesi. Una tale compressione sarebbe impossibile nel caso della distribuzione casuale di punti.

.

Ecco allora affacciarsi una prima guida operativa sul tema della complessita': un sistema e' tanto piu' complesso quanto meno e' comprimibile, quanto piu', cioe', rifugge da

semplificazioni, ovvero manca di una forte asimmetria che ci indirizzi verso una vista privilegiata.

Come abbiamo accennato, il sostituto dell'occhio, il quale risulta limitato al caso bidimensionale e non fornisce stime quantitative della separabilita' dei gruppi, e' il calcolo delle distanze reciproche fra le unita' statistiche della matrice di dati. Il concetto di distanza deriva direttamente dall' applicazione del teorema di Pitagora come si puo' apprezzare dal grafico in figura 2.4:



Applicando il teorema di Pitagora possiamo calcolare il valore di distanza tra i due punti come:

$$d(ab) = \sqrt{(X(a)-X(b))^2 + (Y(a)-Y(b))^2} \qquad [2.1]$$

E' immediato accorgersi che per un campo di dati multivariato, caratterizzato invece che da due sole variabili si tratta solo di aggiungere alla 2.1 altri addendi. La [1] semplicemente si arricchira' di termini rimanendo sostanzialmente identica:

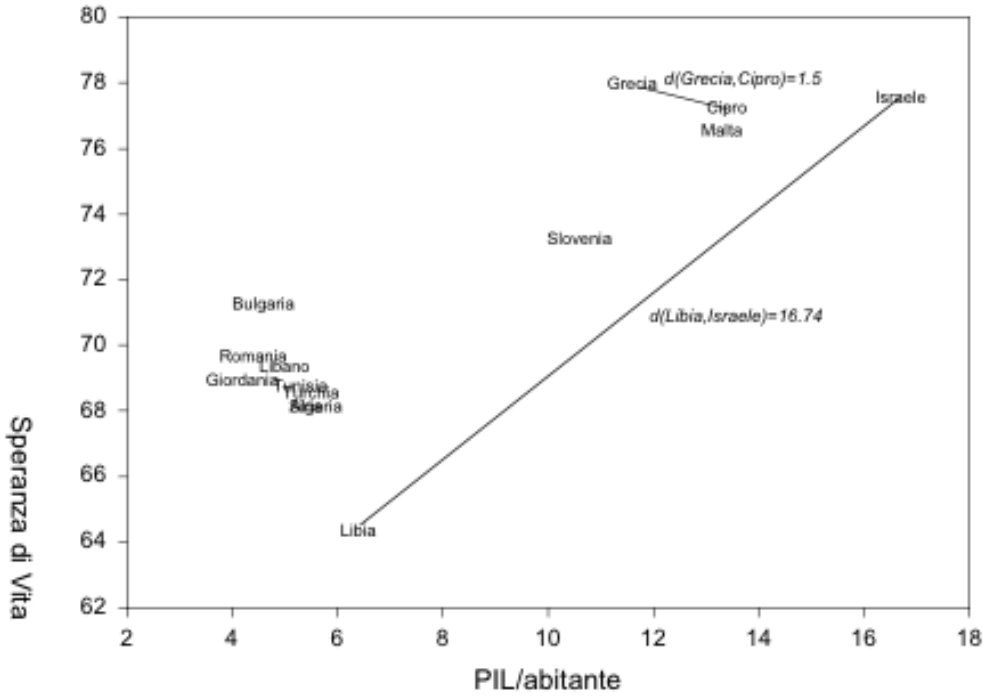
$$d(ab) = \sqrt{(X(a)-X(b))^2 + (Y(a)-Y(b))^2 + (Z(a)-Z(b))^2 + (S(a)-S(b))^2 + (H(a)-H(b))^2 + \ldots} \qquad [2.2]$$

La [2.2] consente quindi di esportare il concetto di categorizzazione per similitudine/discontinuita' a situazioni comunque complesse definite da molte variabili contemporaneamente.

Nel caso delle nazioni, è possibile quindi calcolare la distanza tra i due punti piu' simili tra loro (Grecia e Cipro) e la distanza tra i due punti piu' distanti (Libia ed Israele), mostrate nella figura 2.5, che hanno valore rispettivamente 1.51 e 16.74. Questi due valori derivano dall' applicazione limitata alle prime due variabili della [2.2] ai dati della tabella 2.5, per cui:

$d(\text{Grecia,Cipro}) = \sqrt{(77.9 - 77.2)^2 + (11.6-13.4)^2}$;

e $d(\text{Libia,Israele}) = \sqrt{(64.3 - 77.5)^2 + (6.4-16.7)^2}$.



Il lettore potrà, se vuole, eseguire i calcoli delle distanze tra altri stati per verificare la congruita' della classificazione ad occhio con quella suggerita dal criterio: gruppi con minime distanze interne e massimamente separati tra di loro.

Siamo ora in grado di apprezzare il concetto di "forma" di un sistema costituito da più elementi (nel caso discusso, un insieme di nazioni) come distribuzione di distanze tra gli elementi costituenti (unita' statistiche).

Questa operazione di ricerca di aggregazioni notevoli e' alla base della cosiddetta analisi dei cluster e si puo' immaginare applicata ad ogni situazione in cui si voglia identificare una struttura caratteristica del campo di dati in esame.

Cap.3. *Misurare Correlazioni*

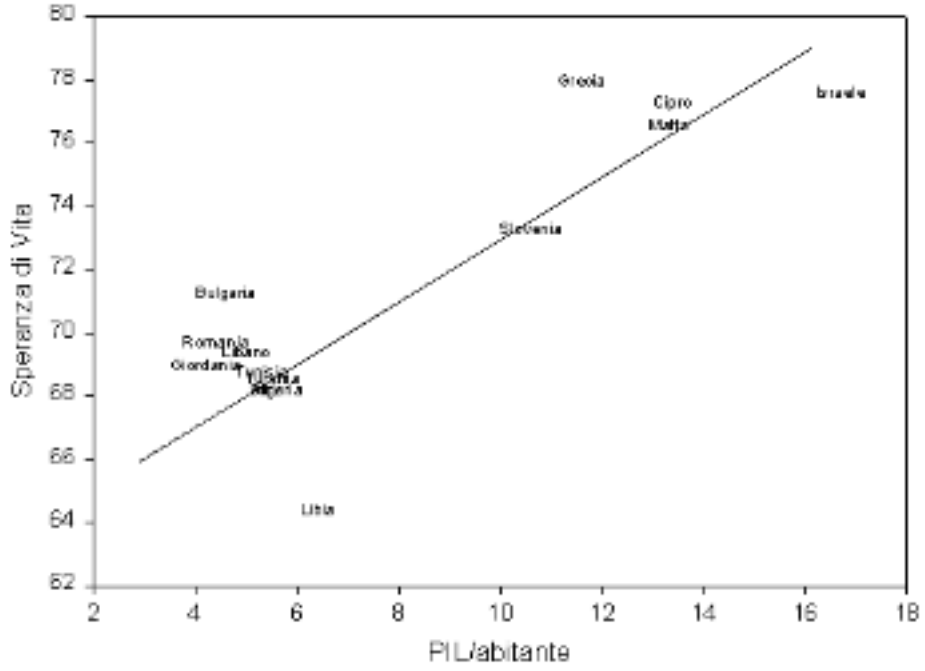
Nel precedente Capitolo, proiettando i dati della tabella nello spazio caratterizzato dagli assi: Speranza di vita e PIL/abitante ed osservando due aggregazioni (cluster) rispettivamente a sinistra in basso e a destra in alto abbiamo immediatamente assegnato alle due aggregazioni il significato di "poveri" e "ricchi". Cosi' facendo abbiamo implicitamente assunto: 1) che la Speranza di vita e la ricchezza prodotta per abitante fossero due facce della stessa medaglia, e 2) che misure in partenza indipendenti e riferentesi a diversi aspetti della realta', vengano a rappresentare una sola dimensione effettiva su cui stimare qualcosa che riflette direttamente il benessere

generale di una nazione. Cio' che correla le due misure non e' pero' qualcosa di intrinseco alle misure *in se'*: un ricco ed un povero entrambi in buona salute, naufraghi su un'isola deserta hanno *a priori* le stesse probabilita' di sopravvivere. Allo stesso modo, due ipotetiche nazioni A e B con lo stesso PIL/abitante ma con due condizioni sociali differenti (ad esempio A con una ricchezza equamente distribuita e B con forti sperequazioni fra i suoi abitanti) avrebbero due speranze di vita medie molto differenti con A nettamente superiore a B.
Insomma la presenza di una correlazione fra due variabili, cioe' il fatto empirico che esse varino insieme, e' semplicemente la spia di un sistema che, all'interno di determinate condizioni al contorno, si organizza secondo una certa struttura, comprimendo i suoi gradi di liberta' e diminuendo la propria potenziale complessita'. Dal punto di vista della metodologia scientifica ci dovremmo allora chiedere: i nostri dati ci autorizzano a fare questa semplificazione e ad assumere che speranza di vita e PIL siano nel nostro ambito di analisi effettivamente correlati ed esprimibili collettivamente col solo concetto unificante di "benessere" ?

La risposta a questa domanda corrisponde alla ricerca di un'altra forma di asimmetria , speculare a quella da noi osservata nel capitolo precedente (creazione di classi omogenee) e che si puo' osservare nella figura 3.1. La linea retta nella figura non e' altro che l'immagine grafica dell'equazione lineare che meglio approssima la relazione empirica osservata tra PIL e Speranza di vita ed e' stata costruita imponendo una ragionevole condizione di ottimalita' basata sul concetto di distanza, cioè che il modello migliore e' quello che rende minime le distanze al quadrato tra i punti e la retta che ne approssima la distribuzione. Insomma, si tratta di trovare quale e' la retta che passa "piu' vicino" ai punti sperimentali e che quindi corrisponde alla descrizione piu' fedele della situazione. L'uso del quadrato e' un artifizio matematico dovuto al fatto che altrimenti differenze negative (punti sotto la retta) si annullerebbero con le differenze positive (punti sopra la retta). Comunque, applicando questo semplice criterio (denominato criterio dei "minimi quadrati", in inglese "least-squares") si ottiene un' equazione che puo' essere scritta come:

$$\text{Speranza di vita} = 64.21 + 0.89*(\text{PIL}) \qquad [3.1]$$

L' applicazione della [3.1] consente di prevedere la Speranza di vita di un paese a partire dal suo prodotto interno lordo con un margine di errore legato alla non perfetta disposizione di punti lungo la retta. La "legge" [3.1] e' la migliore ottenibile dai dati sperimentali a disposizione secondo il criterio dei minimi quadrati e fornisce - lo ripetiamo - la descrizione piu' fedele delle osservazioni raccolte.



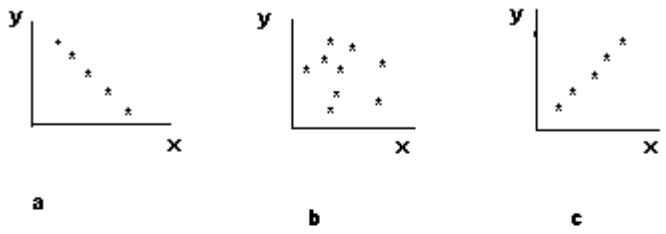
Cosa comporta (da un punto di vista squisitamente metodologico) sapere che una buona approssimazione della Speranza di vita degli abitanti di un paese si puo' ottenere moltiplicando il PIL per 0.89 ed aggiungendo al risultato il valore 64.21 ? Significa semplicemente che con un solo pezzo di informazione (il prodotto interno lordo), possiamo ricostruirne un altro (la Speranza di vita) attraverso l'applicazione di una semplice procedura di calcolo . Abbiamo cioe' attuato una compressione dei dati, di altro tipo rispetto a quella di raccogliere i paesi in ricchi e poveri ma che anche qui conduce al non aver piu' bisogno di due coordinate per localizzare un punto nello spazio perché ne basta una (da cui l'altra e' direttamente ricavabile).

A questo punto la domanda cruciale diventa: "Quanto e' affidabile il mio modello?", o meglio: "Quanto perdo dell' originale ricchezza di informazione attraverso la compressione?" .

Osservando la figura si nota che con la Libia e la Bulgaria il mio errore sarebbe abbastanza rilevante, la Slovenia viene invece predetta molto bene cosi' come la Turchia, la Tunisia, la Siria, l' Algeria e Malta. Queste osservazioni isolate pero' non sono del tutto soddisfacenti. Quello di cui abbiamo bisogno e' una misura generale, per tutto il campo di dati, della sua "compressibilita'" . Abbiamo cioè bisogno di una misura globale di correlazione che stimi quanto fedelmente una variabile sia ricostruibile dall' altra, ovvero quanto lo spazio bidimensionale sia comprimibile su una sola dimensione.

Tra la fine dell' Ottocento e l'inizio del Novecento, lo statistico inglese Karl Pearson:, costrui' lo strumento matematico di gran lunga piu' usato per misurare il grado di correlazione di due variabili. Vale la pena entrare nel dettaglio della costruzione di questa misura, il "coefficiente di correlazione di Pearson:" (o "product-moment correlation coefficient") indicato con la lettera *r*, in quanto rappresenta un esempio geniale di quella semplicita' artigiana che per noi dovrebbe essere la dimensione principale della scienza.

Come detto in precedenza, la prima cosa da chiedere ad una buona misura e' quella di essere facilmente identificabile rispetto al suo ordine di grandezza: dobbiamo insomma a prima vista capire se un valore della misura e' "grande" oppure "piccolo". Nel caso della correlazione dovremmo allora immaginare quale sia il massimo della correlazione possibile (ed equivalentemente il minimo) e far corrispondere un valore finito della nostra misura a questo massimo, cosi' da poter calibrare i valori effettivamente assunti dalla nostra misura nei casi reali sulla base della relativa distanza dal massimo (minimo) di correlazione.
Graficamente possiamo immaginare i "casi limite" della correlazione come descritto nella figura 3.2.



Una buona misura dovra' allora riconoscere come massimi (di segno opposto) le situazioni **a** e **c** ed avere come minimo la situazione **b**.
La soluzione di Karl Pearson: e' stata :

$$r(x,y) = \Sigma (x_i - Mx)(y_i - My) / \sqrt{\Sigma (x_i - Mx)^2 (y_i - My)^2} \qquad [3.2]$$

La [3.2] e' l'espressione analitica del coefficiente di correlazione tra le variabili x ed y.

Iniziamo a capire in che modo e' stato risolto il problema, e quindi la logica di questa misura, a partire dal numeratore e cioe' dall' espressione :

$$\Sigma (x_i - Mx)(y_i - My) \qquad [3.2b]$$

che e' anche detta covarianza. Il simbolo Σ indica la sommatoria, estesa a tutti gli elementi $(x_i - Mx)(y_i - My)$ presenti. I termini x_i ed y_i indicano i valori delle variabili x ed y relativi all'individuo i, quindi la somma va estesa a tutte le coppie x ed y ordinate delle unita' statistiche costituenti il campo di dati, ed i singoli prodotti sono calcolati singolarmente per ogni unita' statistica i. Con Mx ed My si indicano rispettivamente il valor medio della variabile x ed il valor medio della variabile y. Quindi la somma e' relativa ai prodotti tra gli scarti (differenze) di ogni singolo valore dalla sua media per le due variabili x ed y.

E qui è il punto: dalle scuole medie sappiamo che il prodotto di segni concordi da' un risultato positivo ed il prodotto di segni discordi fornisce un risultato negativo ("piu' per piu' fa' piu'", "meno per meno fa' piu'" , "piu' per meno fa' meno", "meno per piu' fa' meno").
La definizione di media (che approfondiremo ulteriormente nei prossimi capitoli) implica l'eguale frequenza nel campo di dati di scarti positivi (valori piu' grandi della media) e negativi (valori piu' piccoli della media). Quindi, tornando al nostro prodotto degli scarti, cio' implica che, con la stessa frequenza, i due termini a prodotto $(x_i - Mx)$ ed $(y_i - My)$ avranno valori positivi e negativi.
Nel caso di mancanza di correlazione fra le due variabili x ed y (situazione **b** della figura), tutte le possibili coppie +,+ ; -,- ; +,-; e -,+ dei due termini compariranno con la stessa frequenza, ci sara' insomma una distribuzione simmetrica delle combinazioni di segni con un equivalersi di prodotti positivi e negativi ed una conseguente loro somma uguale a zero. In altri termini il numeratore della (3.2) avra' valore nullo (o quasi nullo per eventuali fluttuazioni casuali) e il valore di *r* tendera' conseguentemente a zero. Nel caso **a** della figura 3.2, invece, non tutte le combinazioni di segni sono equamente rappresentate: valori grandi (scarti positivi) della x andranno insieme a valori piccoli (scarti negativi) della y ,con un conseguente eccesso di coppie con segni discordi (+,- e -,+) ovvero di prodotti negativi. Questa asimmetria nella distribuzione dei termini del prodotto porta ad una somma di termini negativi e quindi ad un valore inferiore a zero del numeratore e quindi di *r*. Seguendo la stessa linea di ragionamento, il caso **c** della figura, con il suo eccesso di segni concordi dei termini del prodotto (valori grandi con valori grandi, valori piccoli con valori piccoli) portera' a punteggi decisamente positivi del coefficiente di correlazione.

Finora abbiamo esaminato il numeratore della (3.2) ed abbiamo potuto apprezzare come ancora una volta il significato della misura si basi sulla ricerca di asimmetrie, o meglio di rotture di simmetria indotte dalla correlazione di situazioni inizialmente simmetriche: la numerosita' di scarti positivi e negativi presi singolarmente per la x e la y e' identica, la presenza di correlazione rompe questa simmetria verso accoppiamenti preferenziali fra segni, l'assenza di correlazione mantiene tra le coppie

la simmetria presente singolarmente negli scarti della x e della y. Il denominatore della (3.2) ha lo scopo di "contestualizzare" l'indice inserendolo all'interno di un intervallo finito di valori. Risulta chiaro che il massimo di correlazione positiva tra x ed y si raggiunge con l'identita', quando cioe' x = y , mentre il massimo di correlazione negativa corrisponde ad x = -y.
Risulta cosi' stabilita una metrica che consente di apprezzare il grado di correlazione di due variabili e di giudicare quanto forte sia il legame che le unisce.

In base a quanto detto, il coefficiente di correlazione sembrerebbe confinato ai legami di tipo lineare tra due variabili. In realtà, la sua generalità e semplicità ne consentono l'estensione immediata a situazioni ben più complicate , a patto di operare qualche trasformazione sulla struttura dei dati.

Una situazione come quella rappresentata nella prima delle due figure qui di seguito, ad esempio, fornirebbe la falsa impressione di una mancanza di correlazione fra le variabili X ed Y dovuta al semplice fatto che sia valori grandi che piccoli di X corrispondono a valori elevati di Y. Nel caso in figura la situazione ritorna ad essere lineare e la correlazione quindi viene registrata semplicemente trasformando la X nel suo quadrato e quindi correlando la Y non piu' con X ma con la sua trasformata X^2. Per altre situazioni, la scelta di un'adeguata trasformata per mettere in luce la

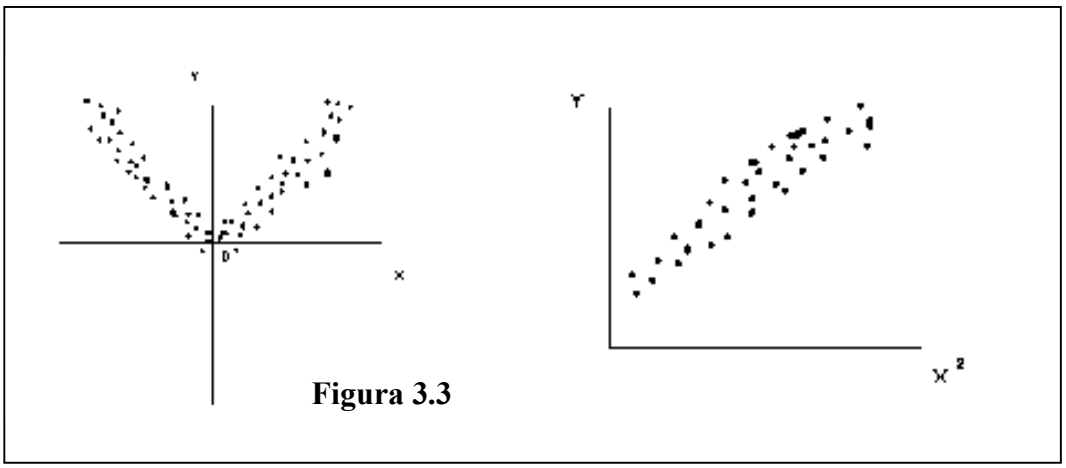


Figura 3.3

presenza di una correlazione e' piu' ardua, ma comunque il concetto e' che qualsiasi situazione di legame, anche non lineare, tra variabili puo' essere valutato in termini di correlazione di Pearson: in un adeguato spazio di riferimento.

Una trasformata particolarmente interessante per la sua versatilita' e' quella che porta a correlare tra di loro invece che le due variabili X ed Y tal quali, le distanze fra le unita' statistiche calcolate lungo l'asse X e le distanze tra le stesse unita' calcolate lungo l'asse Y. Dopo questa trasformazione, qualsiasi tipo di relazione continua tra le due variabili fornira' un coefficiente di correlazione positivo, infatti quello che si chiede per la correlazione e' che due oggetti vicini lungo l'asse X (piccole distanze in X) siano anche vicini in Y (piccole distanze in Y) e analogamente che oggetti lontani in X siano anche lontani in Y. Chiaramente non potremo piu' recuperare il tipo di relazione funzionale esistente tra le due variabili ma, se il nostro scopo era quello di dimostrare una congruenza fra le due rappresentazioni X ed Y, qualsiasi sia il modello alla base di questa congruenza, il nostro scopo sara' raggiunto.

Una generalizzazione di estrema importanza è relativa alla situazione in cui le variabili osservate sono più di due. Questa generalizzazione e' detta analisi delle componenti principali e verra' trattata in seguito.

Cap. 4: *"Modellizzare" Correlazioni (Componenti Principali)*

Nel terzo capitolo abbiamo esaminato in dettaglio il concetto di correlazione tra due o più variabili ed abbiamo sottolineato come la presenza di un legame tra due variabili diminuisca lo "spazio a disposizione" del sistema analizzato costringendo le unita' statistiche, inizialmente descritte in due dimensioni (le variabili X ed Y), a disporsi lungo una sola dimensione corrispondente approssimativamente al vincolo che lega le due variabili. Per mezzo di questa compressione, il campo dei valori della Y si restringe notevolmente rispetto al campo iniziale, la conoscenza della X non e' 'neutra' rispetto alla previsione della Y che si ritrova costretta a variare in un campo piu' piccolo dal vincolo costituito dall'altra variabile.

Intuitivamente quindi, per misurare la compressibilita' di un sistema descritto da più di due variabili avremmo bisogno di una qualche sorta di "media" dei valori della matrice di correlazione tra le variabili. Tale "media" dovrebbe avere a che fare col grado di ricostruibilita' dell'informazione con modelli di regressione di qualche tipo. Oltre a fornirci questa misura di generale di ricostruibilita', la nostra tecnica dovrebbe poi consentirci veramente di ricostruire l'informazione mancante con un numero inferiore di pezzi, ottenendo poi, quasi come un corollario, una descrizione a tutti gli effetti "scientifica" del nostro campo di indagine in termini di relazioni fra gli elementi del fenomeno in studio. Questo e' esattamente lo scopo della tecnica detta Analisi delle Componenti Principali (Principal Component Analysis, PCA).

A differenza del coefficiente di correlazione di Pearson:, la disamina puntuale di PCA (che, analogamente all' indice *r,* costituisce un vero capolavoro metodologico) richiede di entrare in un minimo di tecnicismo algebrico, per cui abbiamo scelto di muoverci su due piani: dapprima forniremo una breve spiegazione formale della tecnica e successivamente un esempio intuitivo del suo uso tentando di renderne più palpabile essenza e potenzialità .

Un teorema di base dell' algebra lineare stabilisce che ogni generica matrice M x N (nel nostro caso la matrice di dati con M righe ed N colonne) , **X**, puo' essere espressa come:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \qquad [4.1]$$

Nella [4.1] le matrici **U** e **V** sono rispettivamente di dimensioni MxK ed NxK e soddisfano alle relazioni $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = 1$. La matrice **S** (matrice di covarianza o di correlazione) di dimensioni KxK e' diagonale simmetrica ed ha i suoi elementi diagonali (valori singolari) ordinabili in ordine discendente $s1 > s2 > s3, \ldots > sk > 0$. In termini intuitivi questo significa che i dati originali possono essere proiettati su un nuovo insieme di coordinate **US** (dette componenti principali o autofunzioni

(eigenfunctions)) in maniera che nessuna informazione sia persa e che quindi ogni elemento della matrice originale dei dati **X** sia ricostruibile dall' equazione:

$\mathbf{X}ij = \Sigma\ \mathbf{U}ik\mathbf{S}k\mathbf{V}jk$, con k che va da 1 ad N [4.2]

Il formalismo non deve trarre in inganno: la [4.2] e' una forma appena piu' sofisticata della [4.1], e ci dice che la posizione esatta di ogni singolo punto del campo di dati puo' essere esattamente ricostruita da una formula del tipo: X = aPC1 + bPC2 + cPC3 + dPC4… cioe' da un' equazione detta "combinazione lineare" (essenzialmente una somma pesata) che ricostruisce la posizione nello spazio delle variabili X tramite nuove variabili (dette componenti principali) che, in questa formulazione, sono in numero uguale al numero di variabili originali (cioe' K = N).
Fin qui quindi non c' e' alcuna compressione, solo un cambio di coordinate che, invece di esprimere i dati con le variabili X, li esprime con le variabili PC dette componenti principali.Geometricamente tutto questo corrisponde ad una rotazione nello spazio del campo di dati.
Diciamo insomma le stesse cose in un altro modo, questo nuovo modo ha pero' una proprieta' interessante: le "parole" sono "ben scandite": le componenti principali sono reciprocamente ortogonali , hanno cioe' tra loro correlazione zero, ogni componente identifica un aspetto indipendente del fenomeno in studio.

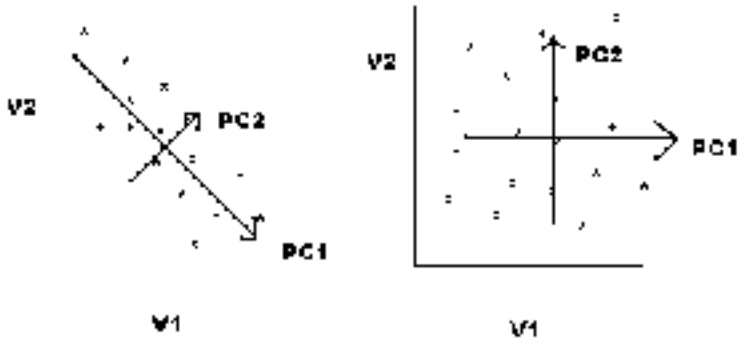
Ma il bello deve ancora venire: se invece di sviluppare tutta la somma definita nell' equazione [4.2] , ci fermiamo ad un numero di termini (A) inferiore ad N otterremo qualcosa come:

$\mathbf{X}ij = \Sigma\ \mathbf{U}ik\mathbf{S}k\mathbf{V}jk + \mathbf{E}ij$,
[4.3]
con k che va da 1 ad A, (A < N), e $\Sigma\ \mathbf{E}ij^2$ = minimo

La [4.3] differisce dalla [4.2] per l'aggiunta del termine **E**ij che e' il termine di errore, quanto cioe' ancora ci manca per ottenere la ricostruzione totale descritta dall'equazione [6.2] se, invece di usare tutti gli N pezzi a nostra disposizione, ne usiamo solo A. La condizione che la sommatoria degli errori quadrati sia un minimo e' la condizione che abbiamo incontrato quando discutevamo di regressione. Il fatto che la somma degli errori sia un minimo implica che la proiezione dell'insieme di dati sul nuovo sistema di riferimento caratterizzato da un minore numero di assi (ricordiamoci che A e' inferiore ad N) e' la proiezione ottimale in termini di fedelta' di descrizione o, analogamente, e' la migliore compressione ottenibile riducendo il numero di dimensioni da N ad A. In altre parole se A = 1, l'unica componente principale estratta (PC1) sara' la piu' fedele compressione monodimensionale del campo di dati in questione, se A=2 la soluzione in due componenti principali (PC1,PC2) e' la migliore rappresentazione bidimensionale e cosi' via…
Quindi le componenti vengono estratte in ordine di potenza di spiegazione : la prima sara' la componente che spiega piu' informazione di tutte mentre le successiva spiegheranno quote via via decrescenti dell' informazione iniziale.
Ecco allora che, tramite questa proprieta', possiamo immediatamente giudicare la compressibilita' (e quindi la complessita') di un campo di dati multivariato semplicemente in termini di fedelta' (il complemento a 100 dell' errore percentuale) della proiezione del campo di dati sullo spazio delle componenti.

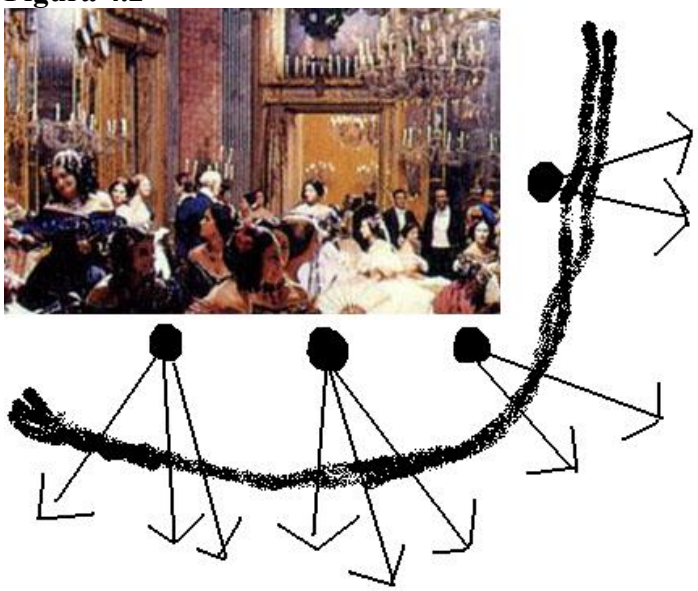
La seguente coppia di grafici ci consente di capire meglio questo concetto:

Figura 4.1



Nel grafico di sinistra si riporta una situazione compressibile: originariamente il sistema e' bidimensionale e descritto dagli assi V1 e V2, la correlazione fra questi assi permette di individuare una direzione preferenziale di variazione corrispondente alla prima componente principale (PC1). Cambiando sistema di riferimento , e cioe' trasportando il piano dalle direzioni V1 e V2 alle coordinate PC1 e PC2, avremo una situazione fortemente asimmetrica in cui PC1 spiega gran parte dell' informazione (questo fenomeno e' rappresentato dalla lunghezza relativa delle due frecce indicanti le direzioni delle componenti principali), mentre PC2 spiega una porzione notevolmente minore della variabilita' fra i punti. Semplicemente la nuvola di punti e' piu' allungata in direzione di PC1, che quindi costituisce un punto di vista privilegiato sul campo di dati: conoscere solo PC1 mi dice di piu' sulla posizione di un punto nel piano che conoscere solo PC2.
Nel grafico a destra viene invece riportata una situazione simmetrica: non c'e' alcuna possibilita' di comprimere il sistema, le due componenti principali spiegano la stessa percentuale di informazione, trasportare il sistema di riferimento dalle coordinate originali V1,V2 al piano delle componenti principali non ci fa fare alcun passo avanti nella compressione (e quindi nella spiegazione) del sistema .
Alla base della tecnica delle componenti principali è l'idea che la natura si presenti ai nostri occhi "velata": noi siamo aldila' di una cortina dietro alla quale si svolge una festa a cui non siamo invitati. Cio' che trapela da dietro alla cortina, voci, musica, risate, tintinnio di bicchieri ha a che vedere con quello che avviene alla festa ma l'informazione e' disturbata e confusa. Se vogliamo conoscere quello che succede dall'altra parte la nostra unica possibilita' e' quella di "incrociare" (confrontare, correlare) le informazioni smozzicate di cui disponiamo e trovarne l'intima coerenza: cio' che unisce i differenti elementi a nostra disposizione costituisce la rappresentazione piu' fedele possibile di cio' che avviene nel salone delle feste.

Figura 4.2



Nella figura [4.2] si esemplifica questa situazione un po' frustrante: le variabili (suoni registrati da orecchi o da microfoni) da noi percepite sono rappresentate dalle numerose frecce che escono dalla cortina, partendo da varie sorgenti sonore al di là di essa (esemplificate dai pallini neri). Da ciascun microfono a noi sembra di percepire informazioni differenti ma in realta', se potessimo partecipare alla festa, ci renderemmo facilmente conto di come molto di cio' che appare diverso proviene da sorgenti comuni (ad esempio una particolare conversazione che avvenga in un angolo del salone, o il fruscio dell'abito di una dama al suo incedere maestoso, etc.).
Individuare ciò che unifica informazioni apparentemente diverse per mezzo della PCA, ci consente allora di immaginare lo svolgimento complessivo della festa.

Cerchiamo di rendere operativa questa metafora. Lo faremo servendoci di una simulazione numerica, un metodo di utilizzo assolutamente generale e divenuto sempre più popolare al crescere della potenza e facilità d'uso dei computer.

Attraverso la simulazione possiamo verificare cosa accadrebbe applicando la PCA in una situazione ideale, simile a quella descritta (certamente semplificata rispetto alla fascinosa scena del ballo del Gattopardo, ma di cui conosciamo a menadito gli ingredienti in quanto decisi interamente da noi) e in cui possiamo controllare la correttezza della soluzione prodotta.

Quello che faremo e' produrre due serie indipendenti di dati (x_0 ed y_0) corrispondenti alle estrazioni da due distribuzioni Gaussiane a media zero e deviazione standard uno, fra le quali non esiste alcun legame (r (x,y) = 0). Queste due variabili sono l'immagine (a noi preclusa) di quello che osserveremmo se fossimo dall'altra parte della cortina. Simuleremo quindi delle "copie distorte" di x_0 ed y_0 semplicemente "sporcando" le variabili originali con rumore, e chiameremo queste copie "sporche" x_1,x_2,x_3,x_4 e y_1,y_2,y_3,y_4.. Le copie sporche costituiscono le nostre misure.

Tornando al nostro ballo possiamo immaginare x_0 come un'originaria sorgente sonora al di là della

cortina (ad esempio, le parole dette da un partecipante alla festa) e x_1, - ,x_4 come i segnali captati da microfoni posti all'esterno della cortina in posizioni via via piu' lontane dal parlante (le variabili sono ordinate in proporzione all'entita' del rumore aggiunto). Ovviamente noi che siamo fuori non sappiamo dove sia la persona spiata e quindi piazziamo microfoni qui e la' e speriamo di capirci qualcosa. Lo stesso vale per l'altra sorgente sonora, y_0, e per y_1 - ,y_4. Anche questo pero' noi non lo sappiamo in quanto per noi che siamo fuori si tratta solo di un altro gruppo di microfoni che, per avventura, capteranno un'altra voce. In altre parole, noi non possiamo conoscere direttamente la 'cosa vera' (x_0 ed y_0), ma solo immagini distorte (X1-X4 ed Y1-Y4) e, soprattutto, non sappiamo neanche se esse siano o meno le immagini distorte di una sola cosa o di tante, o addirittura solo rumore. Nella fattispecie il rumore aggiunto e' pari a: 0.8, 1.2, 1.8 (pari cioe' all'80%, al 120% ed al 180% dell'informazione iniziale), le nostre insomma sono impressioni molto sfumate della realta'. Operando una PCA sull'insieme x_1 - ,x_4 e, separatamente, sull'insieme y_1 - ,y_4 ecco cosa otteniamo:

Tabella 4.1 (a e b)

Caso X

	PC1	PC2	PC3
X1	0.90	0.07	-0.18
X2	0.82	-0.29	-0.37
X3	0.75	-0.20	0.63
X4	0.55	0.83	0.03
% var. sp.	58.8	20.5	14.3

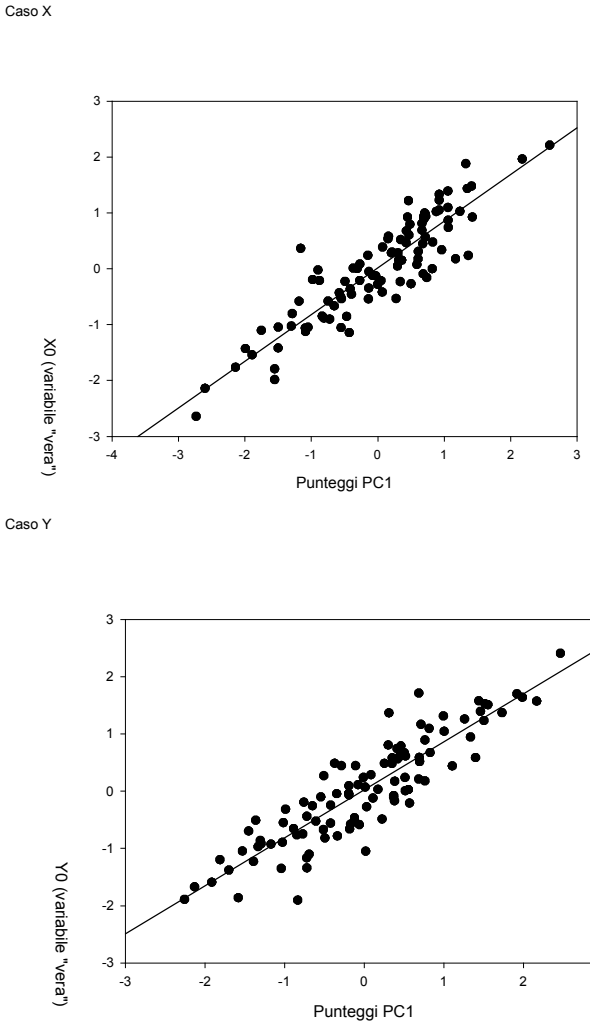
Caso Y

	PC1	PC2	PC3
Y1	0.86	-0.23	-0.22
Y2	0.85	-0.22	-0.28
Y3	0.79	-0.09	0.60
Y4	0.57	0.81	-0.07
% var. sp.	61.0	19.5	12.4

Le tabelle riportano, per entrambi i casi, i coefficienti di correlazione fra le variabili originali X ed Y e le componenti principali (PC#) (component loadings). Nell'ultima riga di ciascuna tabella è indicata la percentuale di informazione (variabilita') spiegata da ogni componente.

Come si vede, nonostante l'elevato grado di rumore aggiunto, in entrambi i casi le componenti principali ci forniscono l'informazione essenziale: la prima componente principale e' di gran lunga la piu' importante in termini di variabilita' spiegata sia nel caso delle X che delle Y il che ci fa supporre che, in entrambi i casi, l'apparente separazione dell'informazione in quattro variabili sia appunto apparente in quanto tutte e quattro le misure sono in realta' riconducibili ad un'unica sorgente primaria (cioè la "vera" variabile X_0 e la "vera" variabile Y_0, che pero' a noi sono precluse).

Osservando i coefficienti di correlazione (component loadings) delle variabili sulle componenti, ci accorgiamo che questo valore scende dalla prima alla quarta: cioè, X_1 e' piu' correlata con PC1 di X_2, che a sua volta e' piu' correlata di X_3.. fino ad arrivare ad X_4 (e rispettivamente Y_4) che sono talmente lontane da PC1 da essere piu' correlate con la seconda componente (PC2) che con la prima. In effetti noi sappiamo che X_4 ed Y_4 incorporano un rumore talmente elevato (circa il doppio dell'informazione significativa) da risultare ormai delle copie quasi irriconoscibili dell' informazione primaria. In ogni caso PC1, calcolata si badi bene SOLO con l'informazione "corrotta", riesce a mantenere il "ricordo" di qualcosa di implicito nelle nostre misure ma mai direttamente misurato, e cioe' il segnale originale X_0 (Y_0)



Nel pannello superiore (a) è riportata la relazione fra la componente principale osservata e la supposta variabile 'vera', nel caso denominato 'x'. Il pannello inferiore (b) riporta la stessa

che e' presente come addendo in ognuna delle copie corrotte Y# ma non e' mai direttamente accessibile.

Pero' qui siamo in una simulazione e, a differenza che nei casi reali, possiamo tirar via di botto la cortina, fare la nostra entrata alla festa (chi si ricorda la sceneggiata napoletana "o zappatore" ?) , prendere i valori di X_0 e di Y_0 e correlarli con le componenti principali, cioe' con la loro immagine ricostruita da cio' che hanno in comune le "variabili copia", ed ecco allora che scopriremo che il coefficiente di correlazione di X_0 con la prima componente principale e' pari a 0.89 e quello di Y_0 con PC1 e' uguale a 0.90, le componenti minori sono invece indipendenti dalle

variabili "reali" X_0 ed Y_0. Insomma con un'ottima approssimazione (i coefficienti di correlazione di PC1 con le variabili generatrici sono vicinissimi all'unita') siamo riusciti a scoprire "cosa c'era sotto" alle nostre osservazioni, mentre il rumore aggiunto si e' distribuito sulle componenti minori ed e' stato "filtrato via" dall'analisi. Le figure 4.3 A e B riportano la relazione fra la variabile "vera" e la sua stima tramite PC1, assumendo cioe' che la "verita' oltrecortina" sia cio' che unisce tra di loro le copie imperfette

Ma immaginiamo un compito ancora piu' difficile, prendiamo in esame contemporaneamente sei variabili (x1-x3 ed y1-y3) e cerchiamo di ricostruire piu particolari possibili della scena.
L' analisi del caso congiunto ci fornisce questi risultati:

Tabella 4.2

Caso XY

	PC1	PC2	PC3
X1	0.31	0.81	-0.11
X2	0.43	0.77	-0.28
X3	0.39	0.81	0.41
Y1	0.79	-0.50	0.07
Y2	0.83	-0.41	0.12
Y3	0.77	-0.47	-0.13
% var.sp.	43.8	43.2	5.0

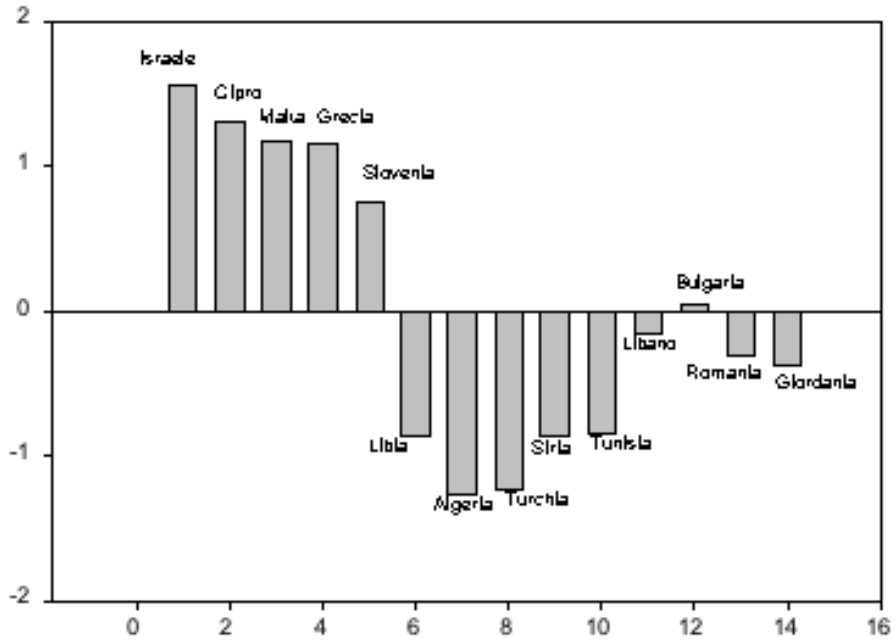
Qui abbiamo mescolato due generi differenti di informazioni, quelle provenienti dalla "sorgente" X e quelle provenienti dalla "sorgente" Y, e l'analisi ci segnala subito che le dimensioni "attive" in questo caso sono due e non piu' una sola: infatti le prime due componenti principali spiegano un quantitativo piu' o meno equivalente di variabilita' (43.8% e 43.2%). La complessita' del sistema e' raddoppiata, ed ho bisogno di due ingredienti per cucinare questa realta', uno solo non basta.
Osservando i coefficienti di correlazione tra le variabili originali ed i fattori, mi accorgo che le variabili X si situano preferenzialmente sulla seconda componente, mentre le variabili Y sulla prima. Questo ci consente di discriminare i due "elementi fondanti" del sistema in studio.

Origliando da dietro la tenda con i nostri rozzi microfoni e analizzando il loro segnale è possibile capire molto di quello che succedeva dall'altra parte con assunzioni puramente strutturali del tipo "quello che e' comune tra le misure e' cio' che piu' interessa", oppure "la correlazione e' guidata dalla struttura dell'oggetto misurato". Si noti che tali assunzioni non hanno niente a che vedere con le caratteristiche peculiari di questo specifico esempio, e possono quindi essere applicate a qualsiasi campo di indagine.

Abbiamo costruito, insomma, una buona stima di qualcosa che non abbiamo misurato direttamente estraendo "il succo" da diversi aspetti parziali e distorti della grandezza, incognita, di interesse. Non abbiamo assegnato arbitrariamente le importanze relative dei suoi vari costituenti, abbiamo lasciato che la loro importanza si definisse "spontaneamente" dalle correlazioni presenti nei dati. Questo stile di lavoro, detto "data-driven" (guidato dai dati) e' particolarmente efficiente nello studio dei sistemi

complessi dove non ci si puo' appoggiare a teorie forti e caratterizza una fase iniziale esplorativa dello sviluppo di un campo scientifico.
PCA e' forse la tecnica di piu' vasto utilizzo: si va dalla psicologia alla meccanica quantistica ed il campo di studi ecologici e' molto ben rappresentato. La ragione di questo enorme successo e' probabilmente da ricercare nel fatto che PCA raccoglie tutti gli aspetti del fare scientifico: raccolta di informazioni, ricerca di correlazioni, generazione di spiegazioni.
Torniamo allora al nostro esempio 'geografico' per apprezzare con mano il tipo di informazioni fornite da questa tecnica. Tornando alla nostra tabella del primo capitolo possiamo vedere come il sistema e' inizialmente disposto su 4 dimensioni (popolazione, Speranza di vita, alfabetizzazione, PIL/abitante), quindi la localizzazione esatta di ogni nazione in questo spazio necessita di quattro pezzi di informazione. Quando sottoposto a PCA, il nostro campo di dati ci fornisce quattro componenti che ovviamente ci permettono di ricostruire esattamente la posizione delle singole unita' statistiche (confronta equazione 4.4). Queste quattro componenti sono pero' gerarchizzate,
con un grado di "potenza esplicativa" pari rispettivamente a:

PC1 = 63.12 %; PC2 = 19.25 %; PC3 = 15.27 %; PC4 = 2.3 %



Questo profilo e' decisamente asimmetrico (nel caso di assoluta mancanza di correlazione tutte le 4 componenti avrebbero spiegato il 25% dell' informazione) e ci indica come esista un concetto ordinatore dei dati, corrispondente alla prima componente, che e' di gran lunga piu' importante degli altri in termini di descrizione delle nostre osservazioni. E' insomma "la lezione piu' evidente" che si puo' trarre dalle nostre misure. Ordiniamo allora le nazioni secondo questo asse corrispondente ai punteggi della prima componente principale e riportato nella figura precedente.

L'altezza delle barre rappresenta il valore di PC1 per ogni unita' statistica. Allo scopo di facilitare l'interpretazione della misura, le componenti hanno media pari a zero, cosi' possiamo immediatamente rilevare chi ha valori grandi (barre verso l'alto) e chi valori piccoli (barre verso il basso). Osservando il grafico, ci accorgiamo che PC1 e' una misura complessiva del benessere delle nazioni con i paesi ricchi con valori alti della componente, paesi poveri con valori negativi e qualche situazione al limite. Il carattere quantitativo della componente ci consente di misurare una grandezza che non era originariamente presente nei nostri dati ma che e' "emersa" da essi tramite le correlazioni tra le diverse misure. Come possiamo esprimere questo concetto nuovo che risulta dalla composizione delle singole informazioni che avevamo sulle unita' statistiche esaminate?
Il modo piu' diretto e' quello di calcolare i coefficienti di correlazione di Pearson (detti "component loadings") delle variabili originali con le nuove variabili (componenti principali) che costituiscono la nuova vista sul sistema. Le variabili con i coefficienti di correlazione piu' elevati con le componenti ci serviranno da guida per l'interpretazione. Ecco allora le correlazioni tra le variabili originali e la prima componente principale:

	PC1
Popolazione	-0.58
Speranza di vita	0.94
Alfabetizzazione	0.74
PIL/abitante	0.87

Ricordandoci il significato del coefficiente di correlazione, e' evidente come il valore di PC1 sia essenzialmente guidato dalla Speranza di vita e dal PIL/abitante che rappresentano le due variabili piu' importanti per la definizione della componente con delle correlazioni vicine alla correlazione massima. L' alfabetizzazione, con una correlazione di 0.74 contribuisce anch'essa (seppur in maniera piu' limitata) alla definizione della componente, mentre il numero di abitanti ha un ruolo minore (con una lieve tendenza ad avere valori piu' alti di PC1 nei paesi meno popolosi). Ricordiamoci come Speranza di vita, Alfabetizzazione e PIL gia' ci apparivano come una triade fortemente connessa dal semplice esame della tabella di correlazione, qui ne abbiamo una misura quantitativa esplicita e integrata in un punteggio sintetico.
Ecco allora che PC1 emerge come una misura dello "sviluppo" di una nazione , di "benessere" insomma, con i "component loading" che pesano l'importanza dei vari costituenti questo costrutto generale.Abbiamo in altre parole costruito una misura globale di benessere, derivante dalla combinazione di vari aspetti parziali di questo concetto.
L' analisi in componenti principali ha un campo di applicazione che va dalla meccanica quantistica alla psicologia ed all'ecologia e in qualche modo riassume tutti gli elementi dell'impresa scientifica dalla raccolta di dati alla costruzione di un'ipotesi di spiegazione.

Cap. 5*: Popolazioni e Campioni*

Il lavoro scientifico inizia sempre con una domanda di cui in parte si immagina gia' di che tipo sia la risposta : la fase iniziale di scelta delle misure da eseguire, della strategia di raccolta dei dati, della definizione delle modalita' sperimentali che precede la fase dell'analisi vera e propria e' orientata a creare un contesto in cui la risposta (resa intelligibile dai metodi descritti nei precedenti capitoli) sia il piu'

possibile esauriente e non ambiguo. In altre parole, bisognerà sistemare le cose in maniera tale che, se l'esperimento da' i risultati attesi, l'unica spiegazione possibile (o almeno quella di gran lunga piu' ragionevole) sia una di quelle previste. Possiamo sicuramente dire che e' in questa fase di progettazione dell'esperimento che si giocano gran parte delle possibilita' di successo dell' impresa.

L'impressione che il lavoro dello scienziato sia molto simile ad un artifizio retorico o a quello di un avvocato che prepari un' arringa e' a questo punto non del tutto infondata. La retorica scientifica (quella buona beninteso, non le tronfie dichiarazioni ai mass media di certi scienziati a caccia di finanziamenti ...) ha pero' una caratteristica che la rende particolarmente affascinante e che costituisce buona parte del gusto del lavoro dello scienziato: tutti i suoi "raggiri" devono essere fatti prima che si inizi a giocare, a "bocce ferme". Una volta apparecchiato con la massima cura il quadro di riferimento, si da' il via all'imponderabile, al gioco vero e proprio, all'osservazione, all'esperimento, e la palla passa alla "Natura" o comunque ad un insieme di circostanze su cui noi non si ha controllo diretto e con cui non si può (e non bisogna assolutamente) interferire. A questo punto "si va a vedere" cosa e' successo: se la configurazione e' quella che ci da' ragione abbiamo avuto successo, altrimenti pazienza, avevamo torto (.. questo e' il risultato di gran lunga piu' frequente) oppure non avevamo organizzato le cose in maniera corretta, ci eravamo dimenticati qualcosa ,,,ma in ogni caso la particolare partita e' persa e, se pensiamo ne valga la pena, dovremo progettare una 'rivincita' indipendente.

In effetti il gioco del biliardo e' forse la cosa che piu' si avvicina al mestiere dello scienziato: le combinazioni di leggi di natura e pura stocasticita' sono piu' o meno nella stessa proporzione in cui si ritrovano nella ricerca scientifica e, soprattutto, i giocatori di biliardo professionisti sono obbligati a dichiarare in anticipo l'effetto del colpo: se per avventura si verificasse un colpo ancora piu' sensazionale ed astruso di quello dichiarato, non avrebbe lo stesso valore della realizzazione della predizione. Insomma, il bravo scienziato non si misura dal numero di volte che ci azzecca (cosa che dipende per larga parte dalla fortuna) ma da come dispone gli elementi del suo quadro cosi' che le sue risposte appaiano chiare (sia in positivo che in negativo). Il cammino e' molte volte piu' importante della meta e sicuramente piu' interessante, come qualsiasi decente escursionista può confermare, e come sapeva bene Martin Heidegger, che premetteva alle sue opere il detto "Wege nicht Werke" (sentieri, non opere compiute).

Il meccanismo retorico alla base di un qualsiasi decente pezzo di scienza puo' essere sintetizzato a grandi linee come segue: " Se non ci fosse stato il fenomeno X (*che io dico di aver scoperto*) le cose sarebbero andate nella maniera A; invece gli avvenimenti sono andati nella maniera B, dimostrando l'esistenza di X".
Per mettere su un' arringa del genere abbiamo bisogno di tre elementi fondamentali:

1) avere un modello condiviso (A) di come andrebbero le cose se lo stato del mondo fosse lo stato reale "diminuito" dell' ente X Questa e' quella che i metodologi chiamano ipotesi "controfattuale" ;
2) dimostrare in modo convincente che le cose sono andate secondo lo schema B, sensibilmente diverso da A;
3) rendere necessario il legame tra il non avverarsi di A e l'esistenza di X.

Per motivi di chiarezza espositiva inizieremo dal secondo punto, che e' poi il meno difficile dei tre ed il piu' facilmente trattabile rimanendo sulle generali e senza entrare nel merito di un particolare problema. Per risolvere il punto 2) dobbiamo dimostrare

che il particolare risultato osservato, se le cose stanno nella maniera A, e' veramente improbabile e, a questo fine, ci serviamo di due concetti estremamente utili: il concetto di *popolazione* ed il concetto di *campione*. Per *popolazione* intendiamo l'intero insieme di misure teoricamente possibili di un certo fenomeno; per *campione* le misure effettivamente realizzate. Il punto 2) allora si risolve stabilendo una misura di "plausibilita'" dell' estrazione di un sottinsieme di misure B dalla popolazione di riferimento A.

Nell'esercizio proposto in questo capitolo vogliamo esaminare la plausibilità della congettura che il ricorso alla psicoterapia sia legato al livello di cultura del soggetto. A questo scopo, viene scelta la popolazione romana di sesso femminile e d'età compresa fra i trenta ed i cinquanta anni. Questa popolazione, anche se non praticamente contattabile nella sua totalita', e' perfettamente definita: in un qualunque momento (diciamo il 21 Gennaio 2005): si puo' cioè stabilire con assoluta certezza se una persona ha i requisiti per far parte della popolazione controllando i registri dell' anagrafe. La popolazione e' costituita da donne con diverso livello di cultura, che noi abbiamo bisogno di esprimere con una stima il piu' possibile fedele. Scegliamo come misura il numero di anni di studio (STUD), anche se tutti sappiamo che non sempre la cultura di una persona e' proporzionale al numero di anni passato sui banchi di scuola. Tuttavia, decidiamo che "grossolanamente" la variabile STUD possa essere un' approssimazione di questo concetto, tanto piu' che ci basiamo su un campione piuttosto vasto in cui eventuali eccezioni si dovrebbero diluire o quantomeno bilanciare.

Una volta definita la popolazione di riferimento, dobbiamo stabilire se la "pesca casuale" di un sottoinsieme molto piu' limitato di questa popolazione fornisca dei valori della variabile STUD sensibilmente diversi da una "pesca mirata", cioè da una pesca in cui si accettano solo gli individui che hanno in piu' la caratteristica "aver partecipato a qualche forma di psicoterapia". Lo scopo è verificare:

- Se il "campione mirato" e' a tutti gli effetti considerabile un estratto qualsiasi della popolazione (e quindi ci fornisce dati coerenti con i dati generali della popolazione stessa), oppure è "distorto" dalla richiesta aggiuntiva relativa alla psicoterapia;
- Se questa eventuale distorsione si riflette in un valore della variabile STUD non compatibile con una pura estrazione casuale.

In altre parole, la nostra domanda iniziale: 'E' vero che la psicoterapia e' seguita da donne di estrazione culturale piu' elevata rispetto a chi non la segue ?' ora viene espressa come: 'Una selezione di donne che frequentano la psicoterapia tende a farci osservare un'istruzione media superiore rispetto ad una selezione di donne che non ha mai frequentato questo tipo di terapia?'.

Per fare questo occorre prima di tutto avere una chiara idea dell'entità delle differenze tra un campione di donne che hanno seguito la psicoterapia ed un campione che non l'ha mai seguita. Questa entita' puo' essere apprezzata solo confrontando la differenza osservata con l'entita' dell' effetto del caso sulle grandezze sintetiche utilizzate nel confronto. Se e solo se la differenza osservata tra i due campioni supera 'significativamente' (preciseremo nel seguito il senso di questo avverbio) la differenza attesa per effetto del caso saremo autorizzati a prendere in considerazione la nostra ipotesi. Per avere una misura relativa delle differenze tra i due campioni, la maniera piu' diretta di procedere e' quella di stimare la media aritmetica della variabile STUD nei due campioni e confrontarne la differenza con la variabilita' attesa nella popolazione. La media aritmetica, cioè la somma di tutti i valori della variabile STUD diviso per il numero N degli individui, e' semplicemente esprimibile dalla formula:

$E(STUD) = \Sigma_i \, stud_i \, /N$ [5.1]

Immaginiamo di aver fatto delle telefonate a caso e di aver raccolto i dati delle persone eligibili per la sperimentazione per eta', sesso e residenza nei due campioni NOPSIC e PSIC riferentesi rispettivamente a chi non ha mai sentito il bisogno di alcuna forma di psicoterapia e a chi l'abbia sentito (sorvoliamo per ora sulla propensita' ad ammetterlo ed immaginiamoci delle risposte veritiere). Alla fine della fase di raccolta dati avremmo allora una matrice di dati cosi' formata (le righe corrispondono alle singole intervistate, le colonne, come al solito alle variabili di interesse, piu' un identificativo costituito dal nome):

Nome	Classe	STUD	Nome	Classe	STUD
Anna	NOPSIC	8			
Laura	NOPSIC	18	Ambra	PSIC	12
Renata	NOPSIC	8	Elisabetta	PSIC	8
Maria	NOPSIC	13	Emanuela	PSIC	8
Antonietta	NOPSIC	5	Grazia	PSIC	17
Giovanna	NOPSIC	13	Maddalena	PSIC	13
Luisa	NOPSIC	8	Cristina	PSIC	13
Benedetta	NOPSIC	18	Maria	PSIC	5
Marianna	NOPSIC	13	Rosa Maria	PSIC	20
Felicita	NOPSIC	13	Rita	PSIC	8
Annamaria	NOPSIC	8	Valentina	PSIC	3
Maria Luce	NOPSIC	8	Valeria	PSIC	5
Immacolata	NOPSIC	13	Ornella	PSIC	10
Ida	NOPSIC	5	Cristiana	PSIC	13
Maria	NOPSIC	17	Silvana	PSIC	18
Olga	NOPSIC	12	Bruna	PSIC	18
Irene	NOPSIC	13	Lavinia	PSIC	18
Giuseppina	NOPSIC	13	Patrizia	PSIC	16
Loredana	NOPSIC	8	Paola	PSIC	13
Laura	NOPSIC	5	Marina	PSIC	5
Veronica	NOPSIC	18	Tiziana	PSI C	13
Ilaria	NOPSIC	8	Ilde	PSIC	17
Fiorenza	NOPSIC	8	Beatrice	PSIC	17
Silvia	NOPSIC	8	Manuela	PSIC	17
Karen	NOPSIC	8	Dora	PSIC	17
Teresa	NOPSIC	18	Alberta	PSIC	14
Marina	NOPSIC	5	Giulia	PSIC	18
Elena	NOPSIC	20	Gina	PSIC	8
Rosa	NOPSIC	5	Federica	PSIC	13
Barbara	PSIC	18	Cleo	PSIC	13
Rossana	PSIC	17	Celeste	PSIC	13
Ida Maria	PSIC	13	Deborah	PSIC	13
Letizia	PSIC	13	Irma	PSIC	18

Finestra 5,1 -- Dati raccolti in un'indagine psico-sociologica

In cui la lettera E sta' per "expectation" cioe' "valore atteso" della variabile STUD. La formula [5.1] e' stata applicata al campo di dati riportato nella Finestra 1, che contiene il nostro materiale di studio.

Si noti che la media dei 31 individui del gruppo NOPSIC e' pari a 10.87 anni di studio, e quella dei 36 individui del gruppo PSIC e' pari a 13.19 anni di studio, con una differenza di 2.32 anni. Sono sufficienti circa 2 anni di differenza nelle medie a dire che il grado di cultura ha un effetto sensibile sul ricorso alla psicoterapia? Per rispondere, occorre confrontare questo differenziale di "circa due anni di studio" con la variabilita' casuale. Abbiamo quindi bisogno di una stima di tale variabilita' nella popolazione, ovvero di un indice che ci dica quanto, in media, ci aspettiamo che un'osservazione presa a caso differisca dalla media della popolazione. Si tratta ancora di un valore atteso, ma non piu' della grandezza tal quale, bensì degli scarti (differenze, distanze) dei singoli accadimenti rispetto al valor medio. Questo indice e' la cosiddetta deviazione standard:

$$\text{Dev. Std. (STUD)} = \sqrt{\Sigma\,(\text{stud(i)}) - \text{E(STUD)})^2 / N} \qquad [5.2]$$

A ben vedere, non e' altro che la media degli scarti delle singole osservazioni dal proprio centro con il solito trucco del quadrato seguito dalla radice per evitare la somma zero. La deviazione standard insomma ci da' un'idea di quale sia l'entita' di una differenza "normalmente attesa" tra un elemento della popolazione e il valor medio della popolazione stessa. E' insomma una misura dell' incertezza della nostra media o, se si preferisce della sua rappresentativita' come indice sintetico dell'intera popolazione.

La figura 5.1 ci fornisce un breve riassunto di questi concetti.

Nel pannello superiore vengono rappresentate due popolazioni **a** e **b** con la stessa varianza e diversa media, nel pannello inferiore due distribuzioni **c** e **d** con la stessa

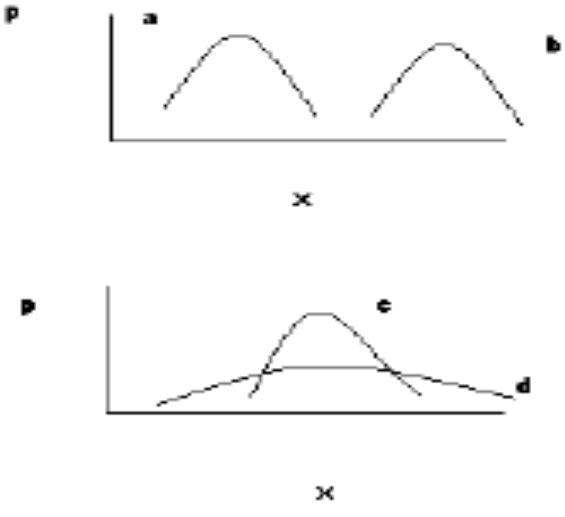


Figura 5.1
Il significato degli assi , in entrambi i pannelli della figura e' identico: x sta per un valore riscontrato in diversi individui (come la nostra variabile STUD); p indica una frequenza, e misura quanto (percentualmente)spesso si osserva il corrispondente valore della x dopo estrazione casuale da una certa popolazione. Il pannello superiore mostra due distribuzioni di diversa media e simile variabilita'. Il pannello inferiore due distribuzioni (c e d) di eguale media e differente variabilita'.

media e differente varianza. E' allora chiaro che la conoscenza del "valore atteso" (media) di **c** ci dara' un'informazione piu' rilevante per la conoscenza della popolazione **c** di quanto non ci fornisca la media di **d**. Ed ecco allora l'uso che potremmo fare dell' informazione sulla dispersione (varianza) per giudicare dell' entita' delle differenze tra due popolazioni: semplicemente misurare quanto le due distribuzioni di frequenze sono sovrapposte tra di loro. Nella figura 5.2, la presenza di una notevole area di incertezza (colorata in nero) in cui un'estrazione casuale potrebbe provenire da entrambe le distribuzioni rende la mia decisione se una certa

osservazione venga dalla popolazione di destra o da quella di sinistra molto piu' incerta nel caso descritto sopra rispetto al caso descritto sotto.

Torniamo alla nostra psicoterapia, la deviazione standard dei due campioni e' praticamente coincidente , essendo pari a 4.55 per la classe NOPSIC e a 4.49 per la classe PSIC.

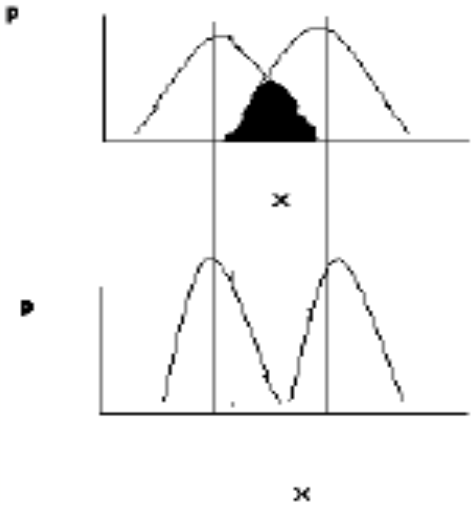


Figura 5.2

Nel pannello superiore le due distribuzioni hanno esattamente la stessa differenza media fra di loro delle distribuzioni del pannello inferiore (la media e' rappresentata dalle linee verticali). Ma la maggiore varianza delle distribuzioni del pannello superiore diminuisce la separabilita' delle due distribuzioni rispetto alle distribuzioni del piano inferiore.

Questo ci dice che la differenza osservata di 2.32, dovuta o meno al caso, e' comunque piuttosto piccola, essendo ampiamente all'interno della variabilita' naturale. Possiamo farla finita qui? Potremmo, e sicuramente non ci saremmo persi un risultato eclatante, ma ci stiamo dimenticando di una sottigliezza legata ad un insidioso concetto che la scienza ha mutuato da quella potente quanto pericolosa alleata che e' la matematica e cioe' il concetto di "valore vero".

Immaginiamo di poter effettivamente chiedere a TUTTE le donne di Roma di eta' compresa tra trenta e cinquanta anni se hanno mai frequentato sedute di psicoterapia e per quanti anni sono andate a scuola. Questa e' un'operazione lunga e tediosa ma che non presenta impossibilita' intrinseche. Immaginiamo che alla fine della rilevazione le due popolazioni PSIC e NOPSIC abbiano come valor medio e deviazione standard proprio gli stessi valori che noi abbiamo osservato nei nostri miseri campioni di 36 e 31 individui. A questo punto noi SAPREMMO CON CERTEZZA che le due popolazioni SONO DIVERSE in quanto PSIC ha media 13.19 e NOPSIC 10.87.
Qui il caso non sembrerebbe aver piu' ragion d'essere in quanto ho una vista completa ed esauriente sull'intera popolazione.
Allora il nostro ragionamento di buon senso sull'entita' delle differenze si dimostrerebbe un'imperdonabile leggerezza che ci celerebbe la verita'. Ricordiamoci che alla base dell' inferenza statistica (come di qualsiasi tipo di inferenza) c'e' l'ambizione di derivare "verita' generali" da osservazioni parziali e cioe' di generalizzare da un campione alla popolazione: noi stiamo preparando un'arringa per dimostrare che la psicoterapia c'entra con lo studio IN GENERALE e non per le nostre 67 intervistate. Pero' di nuovo qualcosa sembra mancare, proviamo a fare ordine…

Allora, se il nostro scopo e' quello di inferire, da un campione ridotto, qualcosa sulla popolazione generale di riferimento, il punto e' quello di definire la rappresentativita' del campione rispetto alla popolazione generale. In altri termini il problema diventa : "con che verosimiglianza il valor medio della popolazione di riferimento e' sovrapponibile a quello osservato nel campione. In questo caso la variabilita' che ci

interessa non e' piu' la 'variabilita' naturale della popolazione' che giustifica il fatto che una persona coltissima non abbia mai sentito il bisogno di intraprendere un percorso di psicoterapia, oppure che una persona di bassa istruzione, per un suo disagio psicologico, si sia indirizzata verso quel tipo di aiuto. Questa variabilita' e' ineliminabile ed e' un dato di fatto del problema. Nel caso della rappresentativita' del campione pero' la variabilita' e' qualcosa di completamente diverso ed ha a che vedere semplicemente con quanto bene una media derivata da un campione ridotto approssima la media della popolazione. Qui insomma stiamo parlando di medie e non di valori singoli, questo secondo genere di variabilita' non e' ineliminabile anzi, puo' essere resa piccola a piacere fino a sparire del tutto quando le dimensioni del campione divengono uguali a quelle della popolazione di riferimento.
Possiamo allora definire una misura di variabilita' che, a differenza della deviazione standard, non misura piu' lo scarto medio di una osservazione dal suo insieme di riferimento, bensì lo scarto della media di un campione dalla media della popolazione da cui e' tratto. Questa misura di variabilita' si chiama Errore Standard (ES) e corrisponde alla Deviazione Standard diviso la radice della numerosita' del campione:

$$ES = Dev.Std./ \sqrt{N} \qquad [5.3]$$

Comprendiamo facilmente come l'errore standard, al crescere del valore di N (numerosita' del campione), diventi sempre piu' piccolo.
Nel nostro caso le deviazioni standard dei due gruppi PSIC e NOPSIC che si situavano attorno a 4.5 corrispondono ad un errore standard attorno a 0.8, decisamente inferiore alla differenza di 2.32 osservata tra i valori medi della variabile STUD nei due gruppi. Infatti l'applicazione di un classico test inferenziale che associa alle distribuzioni osservate una densita' di probabilita' denominata Gaussiana porta a stimare una probabilita' del 4% che la differenza osservata fra i due gruppi sia dovuta al caso. Questo valore del 4% ci consente di dire che l' effetto degli anni di studio sul ricorso alla psicoterapia e' statisticamente significativo, anche se in modo marginale.

Quanto sopra e' assolutamente rigoroso ed in linea con l'ortodossia statistica, ma all'autore (e cosi' speriamo anche a qualcuno dei lettori) sembra di sentire un vago sentore di imbroglio, ancora piu' insidioso perche' supportato dall' evidenza matematica. Insomma, siamo ammirati della sottigliezza del ragionamento, ma tutto sommato ci era piu' simpatico quel rozzo ma schietto confronto fra 2.3 e 4.5. Torniamo allora all' equazione [5.3] e, come fanno i matematici, facciamo un discorso "al limite", e vediamo cosa succede stiracchiando oltre il ragionevole le condizioni al contorno. In particolare, se la numerosita' del campione tende ad infinito (ovvero quando N e' molto grande), è facile comprendere che, dato che la Deviazione Standard e' un numero finito, aumentando il valore di N che e' al denominatore della [5.3], ES tendera' a zero. Tutto giusto no ? Al crescere della numerosita', il campione diventera' sempre piu' rappresentativo della popolazione con cui tende a coincidere. Questo risultato pero' porta con se' una conseguenza perversa: qualsiasi differenza, comunque piccola, rispetto a qualsiasi cosa, con un campionamento sufficientemente grande, diventa statisticamente significativa, cioe' 'vera'. E con questo la matematica seppellisce definitivamente la scienza sotto una montagna di ridicolo dimostrando come si possa rigorosamente affermare qualsiasi bestialita' se si hanno abbastanza mezzi e tempo a disposizione per affrontare una sperimentazione sufficientemente vasta e quindi (secondo la vulgata corrente anche tra la maggioranza degli scienziati) rigorosa.

Cerchiamo di stare calmi e, mentre traiamo da questo risultato un' altra occasione di meditazione su quanto la sensibilita' scientifica differisca da quella matematica, ma anche, cosa molto piu' importante, di comprendere l'imbroglio potenzialmente celato dietro frasi come "…suffragato da una vastissima sperimentazione", e vediamo di salvare la possibilita' di un discorso razionale. Per prima cosa separiamo il concetto di "statisticamente significativo" da quello di "rilevante". La significativita' statistica si riferisce alla probabilita' che due misure siano o no assolutamente identiche. Assolutamente identiche e' molto diverso da ragionevolmente identiche, tant'e' che mentre e' praticamente impossibile che, diciamo, l'altezza di due persone misurata con la precisione del decimo di millimetro sia identica, il mondo e' pieno di persone alte circa un metro e ottanta.
Questo implica che la significativita' statistica, come d'altronde qualsiasi altro strumento, e' un concetto utilissimo ad un certo dettaglio di misura mentre, oltrepassato questo dettaglio, diventa addirittura controproducente.
Ecco allora che il punto non e' quello di dimostrare la significativita' di una differenza per quanto piccola essa sia ma di definire, prima di iniziare i calcoli, quale sia la minima differenza che noi consideriamo come rilevante per i nostri scopi.

Finestra 5.2 . *SIGNIFICATIVO non coincide con RILEVANTE*

Alla base del raggiungimento della significativita' statistica della differenza fra due campioni stanno tre ingredienti che intervengono alla pari nel raggiungimento di un certo valore di significativita':

1) Il valore effettivo della differenza osservata (Differenze grandi hanno piu' possibilita' di essere significative).
2) L' entita' della variabilita' naturale della popolazione stimata dalla Deviazione Standard dei campioni (popolazioni piu' eterogenee saranno caratterizzate da un grado maggiore di incertezza rispetto alla loro media e quindi permetteranno con piu' difficolta' di osservare effetti significativi).
3) Numerosita' del campione per cui campioni piu' numerosi, essendo immagini piu' verosimili della popolazione di riferimento aumentano la potenza del test statistico.

Il bilancio di questi tre ingredienti deve essere accortamente dosato dallo sperimentatore per evitare risultati assurdi, per cui una numerosita' molto alta del campione ed una variabilita' relativamente bassa della popolazione consente di osservare come significative anche differenze molto piccole tra le medie. Viceversa un campione poco numeroso in presenza di una notevole variabilita' naturale non permettera' di dimostrare come significative anche differenze piuttosto elevate delle medie. Esistono delle formule facilmente accessibili per via Internet (ad esempio all' indirizzo (http://department.obg.cuhk.edu.hk/ResearchSupport/Sample_size_EstMean.asp o anche http://www.raosoft.com/samplesize.html per test sulle proporzioni e http://www.health.ucalgary.ca/~rollin/stats/ssize/n2.html per test su variabili continue) che consentono di simulare i possibili esiti di un esperimento con domande del tipo " In presenza di una Dev. Std. attorno a 10 e due campioni da 50 individui ciascuno posso individuare un' eventuale differenza di 20 punti dei miei due gruppi come significativa ?" o viceversa "quale e' la differenza minima osservabile come significativa a partire da una Dev. Std. pari a 5 e due campioni da venti individui ?" e cosi' via….

Questa attivita', da svolgersi prima dell' ottenimento dei risultati, consente di avere un'idea del tipo di esperimento da programmare sulla base delle risorse disponibili e, principalmente di cosa io ritengo una differenza degna di nota. Insomma uno sperimentatore che sia interessato ad un farmaco antiipertensivo inizia a ritenere un risultato rilevante se il farmaco abbassa la pressione del 20%, mentre non ha alcun interesse ad organizzare un test che dimostri in modo inequivocabile che VERAMENTE il suo farmaco abbassa la pressione in media dello 0.1% in quanto questo abbassamento di pressione ancorche' verissimo, certificato da una impressionante e rigorosa sperimentazione, non cambia di nulla la condizione del paziente.

Insomma mentre in matematica esiste un bordo netto tra il VERO ed il FALSO, nelle scienze sperimentali il confine cruciale e' tra il rilevante e l'irrilevante e quest'ultimo, anche se e' vero, resta sconsolatamente irrilevante. Ed ecco allora che comprendiamo due punti chiave in parte contrari al nostro consueto modo lineare di vedere:
1) Che gli strumenti piu' sensibili non sono necessariamente i migliori.
2) Che il concetto di verita' di un'informazione e' in una posizione subordinata rispetto al concetto di rilevanza.

Dimostrare che un farmaco abbassa significativamente la pressione arteriosa di 0.2 mm di mercurio, ancorche' vero e rigorosamente dimostrabile, e' del tutto irrilevante da un punto di vista clinico. Prima di entrare nel merito della significativita' statistica di un risultato dobbiamo allora decidere della rilevanza conoscitiva di un risultato e per farlo dobbiamo ricorrere a quella che si definisce la 'potenza del test' e calibrare la numerosita' del campione in funzione della rilevanza del risultato. Immaginiamo che, sempre nel caso della pressione arteriosa, si decida che la minima differenza rilevante sia di 10 mm di mercurio, allora selezioneremo un campione di numerosita' N tale per cui, data una certa variabilita' naturale (varianza) della popolazione, un livello del 5% di significativita' statistica si raggiunga con un valore di differenza tra le medie attorno a 10. Il 'vero' insomma non confina con il 'falso' ma entrambi sono immersi in una vasta regione di 'irrilevante'. E' importante tenere presente questo dato quando leggiamo sulla stampa di ricerche che ci dimostrano le cose piu' straordinarie (correlazioni fra abbondanza della prima colazione e rischi di impotenza, consumo di vino rosso e tendenza a delinquere e via di questo passo..).
Il caso della psicoterapia era rimasto in sospeso: del nostro risultato, possiamo dire che vale la pena approfondire o possiamo cavarcela facendo spallucce ? Nel prossimo capitolo tenteremo di mostrare come un punto di vista apparentemente piu' grossolano consenta di ricavare una risposta sensata per questo caso, per ora un po' disastrato.

Cap. 6: *Descrivere contingenze*

La scelta della misurazione della "cultura" in termini di anni di studio nell' esempio trattato nel capitolo precedente e' stata presentata come un ' approssimazione un po' brutale ma necessaria, dettata dall' esigenza (peculiare del lavoro scientifico) di basarsi su osservabili quantitative. Cosa significa pero', in dettaglio, "osservabile quantitativa" ? A tutta prima ci viene in testa un numero che ci indica il livello raggiunto da una certa quantita' di interesse : "ho comprato due litri di vino" e' un' affermazione che fa' uso di un'osservabile quantitativa; "ho comprato un po' di vino" no. La differente potenza euristica delle due affermazioni salta agli occhi: se ho comprato due litri di vino so che ne ho comprato meno di Giovanni (il quale invece ne ha comprati tre) e piu' di Maria (che ne ha preso solo uno). So anche che, con ragionevole approssimazione, posso invitare per cena fino a sei amici senza fare brutte figure, e via speculando….. Tutte queste conoscenze sarebbero ovviamente per sempre celate dal vago "ho comprato un po' di vino". Non e' pero' SEMPRE cosi' e capire come mai ci porta al centro esatto del concetto di nonlinearita' e dello stile delle descrizioni in scienza.

Torniamo al caso della psicoterapia, e approfondiamo la natura del nostro osservabile STUD (numero di anni di studio). Nelle societa' moderne lo studio e' organizzato secondo diversi livelli gerarchici (scuole elementari, medie, medie superiori, Universita', etc.) che, avendo durate standard, provocano un'ovvia organizzazione (con conseguente diminuzione di complessita') della nostra misura. In altre parole chi ha fatto solo le scuole elementari avra' STUD = 5; chi e' arrivata fino alle medie avra' STUD = 8, chi fino alle superiori si attestera' su STUD = 13, chi fino all' Universita', a seconda delle facolta', si sistemera' su un valore di STUD da 17 a 19. Insomma, per usare un po' di gergo, STUD non e' libera di assumere tutti i valori possibili entro certi estremi ma risulta "quantizzata". Ovviamente ci saranno delle eccezioni: chi ha fatto solo i primi due anni delle superiori e poi ha abbandonato, chi un breve corso professionale di un anno dopo le medie e cosi' via. Nonostante questo, la particolare

organizzazione degli studi in Italia nel periodo relativo alla vita scolastica del nostro campione tra i trenta e i cinquanta anni, avra' sicuramente lasciato un' impronta sulla distribuzione del nostro osservabile. Questa impronta sara' visibile in termini di cluster: cioe' di aggregazioni notevoli dei valori attorno ai valori "guida" corrispondenti ai livelli di istruzione.

Quando sottoponiamo le nostre 67 (31 NOPSIC + 36 PSIC) osservazioni ad un' analisi dei cluster osserviamo come la classificazione che spiega la maggior parte della variabilita' dell' insieme di dati e' una classificazione in 4 gruppi (cluster) aventi come media e deviazione standard della variabile STUD i valori riportati qui sotto insieme alla numerosita' delle diverse classi:

Tabella 6.1

	E (STUD)	Std. Dev. (STUD)	N
Cluster A :	4.78	0.67	9
Cluster B :	8.12	0.50	16
Cluster C :	12.91	0.43	22
Cluster D :	17.75	0.97	20

Osservando i valori medi dei cluster ci accorgiamo come l'insieme studiato si distribuisca secondo i livelli gerarchici dell' ordinamento scolastico italiano, con il cluster A corrispondente all'istruzione elementare, il cluster B alle medie inferiori e i clusters C e D, rispettivamente, alle medie superiori ed all'universita'. Questa "clusterizzazione" spiega cira il 94% dell' intera variabilita' del campione, il che corrisponde a dire che, se sostituiamo al valore reale degli anni di studio di ogni unita' statistica il valor medio del suo cluster di appartenenza, perdiamo solo il 6% dell'informazione, e precisamente quello legato alle singolarita' personali di studi interrotti, corsi professionali ecc.

Finora abbiamo scoperto soltanto che il sistema e' fortemente compressibile (poco complesso) a causa della rigida struttura dei curriculum scolastici. Abbiamo pero' anche la sicurezza che, se sostituiamo ad un valore numerico un valore qualitativo corrispondente al livello gerarchico non perdiamo molto. Anzi, possiamo esagerare in grossolanita' e raggruppare le nostre osservazioni in due grandi classi: la prima, che indicheremo come "istruzione inferiore", raccoglierà le appartenenti alle classi A e B; la seconda che indicheremo come "istruzione superiore", sarà formata dalle unita' statistiche dei clusters C e D. A questo punto il test sulla significativita' dell'effetto del grado di cultura sull'uso della psicoterapia si risolve nel controllo dell'asimmetria della distribuzione dei due gruppi (*a priori*) PSIC e NOPSIC nei due clusters (*a posteriori*) "istruzione inferiore" (INF) e "istruzione superiore" (SUP), dopo aver raggruppato le osservazioni in una tabella a due entrate, detta *tabella di contingenza*, organizzata in questo modo:

Tabella 6.2

	SUP	INF	**Totali riga**
NOPSIC	15	16	31 (46.27%)
PSIC	27	9	36 (53.73%)
Totali colonna	42 (62.69%)	25 (37.31%)	

Come si vede le somme dei totali riga e dei totali colonna corrispondono entrambe a 67 (il numero delle osservazioni), ma la distribuzione tra SUP ed INF non e' la stessa per le due classi PSIC e NOPSIC. Infatti, NOPSIC ha il 48.4% delle osservazioni (15/31 * 100) nella classe ad istruzione superiore ed il restante 51.6% (16/31 * 100) nella classe ad istruzione inferiore, laddove per l'intero insieme di dati (PSIC+NOPSIC, vedi totali colonna) le corrispondenti percentuali sono notevolmente diverse e precisamente pari al 62.7% nella classe SUP e 37.3 % in INF. La classe PSIC e' invece "sbilanciata" verso un eccesso di istruzione superiore con il 75% delle osservazioni (27/36*100) nel gruppo SUP ed il restante 25% (9/36*100) nella classe INF.

La significativita' statistica di questa asimmetria si puo' misurare, analogamente a quanto visto con le differenze tra le medie nel precedente capitolo, facendo riferimento ad un modello di distribuzione di probabilita' noto (in questo caso la distribuzione detta del chi-quadro) che consente di calcolare la probabilita' che l'asimmetria osservata sia dovuta al caso. Il valore di probabilita' raggiunto e' in questo caso pari al 2%. Ci sono insomma solo due possibilita' su cento che lo squilibrio osservato sia dovuto all'assortimento casuale del campione.
Se confrontiamo questo valore del 2% con la significativita' statistica del 4% osservata considerando il dato quantitativo tal quale (media della variabile STUD nei due gruppi), ci accorgiamo dell' apparentemente strabiliante risultato di come una misura piu' grossolana (classi generali di istruzione piuttosto che "effettivi" anni di studio) ci abbia consentito di raggiungere un livello di sicurezza della decisione sull' esistenza di un effetto dell'istruzione sulla propensione alla psicoterapia doppio rispetto all'informazione "completa".

Dov'e' il trucco? Come la mettiamo con i sensatissimi discorsi sui litri di vino? Intanto, consideriamo con piu' attenzione la tabella di contingenza: al lettore accorto non sara' sfuggita la forte somiglianza con il coefficiente di correlazione di Pearson. Di fatto, per spiegare il coefficiente di correlazione avevamo proprio fatto ricorso ad una classificazione discreta (nominale) degli scarti dalla media in "positivi" corrispondenti a valori grandi della variabile e "negativi" (valori piccoli). Vale la pena approfondire l'analogia descrivendo l'indice che si chiama appunto "Pearson: chi-square statistic" e corrisponde, anche formalmente, all'indice r (cfr capitolo 3) trasportato al caso delle variabili nominali.

Immaginiamo allora una tabella di contingenza tra due variabili X ed Y che possono assumere due valori (1/0, si/no, +/- sono notazioni del tutto equivalenti):

Tabella 6.3

	Y = « + »	**Y = « - »**
X = « + »	15 (N_{11})	3 (N_{12})
X = « - »	2 (N_{21})	16 (N_{22})

La tabella riporta la distribuzione degli elementi per le varie combinazioni delle determinazioni delle variabili X ed Y, per cui ci saranno 15 unita' che avranno entrambe le denominazioni + (qualsiasi cosa significhi) per le due variabili X ed Y, 3 unita' che avranno la determinazione + per X e – per Y e cosi' via. La numerosita' delle varie combinazioni ++; +- ; -+; --; sara' indicata come N_{11} , N_{12} , N_{21} , N_{22} rispettivamente. Anche qui, come nel caso dell'indice r, si tratta di misurare una distanza dalla simmetria, cioe' dalla distribuzione equa di tutte le unita' nelle caselle.

Qui la simmetria corrisponde ad un numero di elementi in ogni cella uguale al prodotto della frequenza relativa di ogni singola alternativa per il numero totale di elementi N, cioe' al prodotto, per ogni cella del totale di riga per il totale di colonna. Allora per ogni cella ij (con i e j che possono essere 1 o 2) il numero atteso di elementi sotto l'ipotesi simmetrica (nessuna relazione tra X ed Y) e' pari a:

$$Mij = (f_i)*(f_j)*N \qquad [6.1]$$

Nella [6.1] i termini f corrispondono alla frequenza relativa di ogni alternativa SINGOLA , per cui ad esempio $f_{i=1}$ = (elementi con X = "+" diviso il totale degli elementi) = ((15+2) / 36 = 0,5), mentre $f_{j=1}$ = (elementi con Y = "+" diviso il totale degli elementi) = ((15+3) / 36 = 0,47), per cui $M_{11} = 0.5*0.47*36 = 8.46$. Ci aspettiamo cioe', per puro effetto del caso, se non esiste nessun legame tra la variabile X e la variabile Y circa 9 individui nella casella +,+. Invece ce ne troviamo 15. Come e' questo scostamento, grande o piccolo ? Comunque sia, lo dobbiamo valutare per tutte le caselle e, come abbiamo visto in altre occasioni lo misuriamo sotto forma di distanza quadratica e cioe' :

$$Qp \text{ (Pearson: chi-square statistic)} = \Sigma\Sigma (N_{ij} - M_{ij})^2 / M_{ij} \qquad [6.2]$$

Nella [6.2] il doppio segno di sommatoria indica che il calcolo va fatto per tutti i valori i e per tutti i valori j (tutte le celle), il quadrato serve a pesare in maniera positiva sia gli scarti in una direzione che nell'altra (che altrimenti si annullerebbero) ed il denominatore ci consente di essere indipendenti dalla 'scala' del fenomeno. La [6.2] e' insomma a tutti gli effetti una 'distanza dal caso di pura simmetria'.

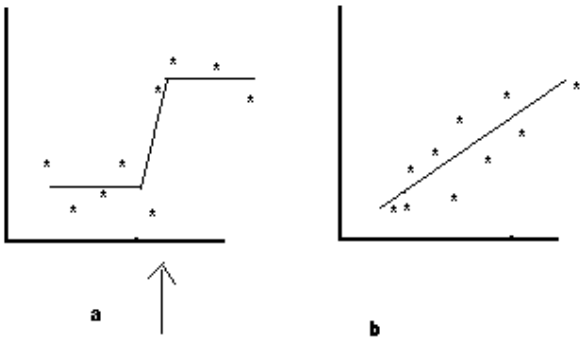
Nel caso visto in Tabella 6.3, che risulta fortemente asimmetrico (le alternative ++ e – sono molto piu' popolate delle altre), Qp vale 18.84 . Ovviamente, come abbiamo visto per il caso delle variabili continue, per giudicare dell'entita' della misura devo normalizzare l'indice mettendolo in rapporto con la massima asimmetria ottenibile che, analogamente al caso di r corrispondono al valore di Qp nel caso X = Y ed X = -Y (cioe' la Y identica alla X con valori scambiati + quando X vale – e viceversa). Nel caso in tabella il valore dell' indice normalizzato risulta essere pari a 0.73. Facciamo caso al fatto che qui stiamo ragionando in termini pienamente numerici (18.84, 0.73 sono numeri a tutti gli effetti) anche se eravamo partiti da misure semiquantitative (istruzione superiore, istruzione inferiore). Cio' che ci ha fatto passare dalla relativa vaghezza della misura singola alla precisione del dato quantitativo e' stato l'operazione del contare, cioe' l'allargamento della nostra visuale ad un collettivo statistico.

Chiarito come di fatto, a livello di collettivo statistico, una determinazione qualitativa sia a tutti gli effetti dotata di un contenuto operativo equivalente a quello di una misura "intrinsecamente" quantitativa, ci risulta piu' chiaro capire come l'apparente grossolanita' della classificazione possa risultare in una maggiore sensibilita' effettiva. La questione e' che il dato di partenza e' per natura discreto e non continuo, dal momento che la gerarchizzazione del sistema scolastico fa si' che i cluster siano le "classi naturali" su cui si dispongono le osservazioni. In altre parole, quando effettivamente esistono delle classi naturali dovute ad un certo "parametro d'ordine", l'operazione di classificazione (e quindi il passaggio da una determinazione continua ad una discreta) corrisponde ad una sorta di "filtraggio" del segnale dal rumore di

fondo, e quindi produce un aumento di sensibilita' dello strumento di misura. Con le nostre classi SUP ed INF siamo insomma piu' vicini alla realta' della strutturazione della distribuzione della 'cultura' nella popolazione che con l'apparentemente piu' precisa definizione attraverso gli anni di studio.

Ovviamente, la determinazione della statistica di Pearson e, in generale, delle analisi di contingenza, non e' ristretta al caso di rappresentazioni binarie tipo + e – . Possiamo costruire classificazioni comunque complesse, con variabili nominali coinvolgenti numeri comunque elevati di categorie. Un esempio tipico ci e' fornito dall' ecologia: la caratterizzazione di un certo ambiente in ecologia vegetale si basa sulla conta degli individui appartenenti alle diverse specie incontrati in una certa porzione di territorio. Immaginiamo allora di avere 50 specie vegetali che possiamo incontrare in due ambienti A e B e di essere interessati alla determinazione di una similarita' significativa tra i due ambienti (i.e. l'appartenenza di due aree lontane geograficamente allo stesso tipo di ambiente). Quello che faremo sara' riempire due serie da 50 caselle l'una per i due ambienti studiati, inserendo in ogni casella il corrispettivo numero di individui di quella specie trovati nell' ambiente A e nell' ambiente B rispettivamente. Questa raccolta di dati generera' una tabella 50*2 che tratteremo esattamente allo stesso modo di come abbiamo trattato la tabella 2*2 dell'esempio precedente.

A questo punto il lettore smaliziato potrebbe sospettare: "Qui c'e' puzza di Achilli e Tartarughe! Qui ci state prendendo in giro: se posso fare il numero di classi che mi pare, alla fine tornero' al caso continuo o meglio, visto che in ogni caso le quantita' continue vengono rappresentate con un numero finito di cifre, a ben vedere esiste solo il caso discreto ed il continuo e' numericamente non rappresentabile". Il lettore ha sicuramente studiato ed e' anche piuttosto acuto, qui pero' noi non facciamo "discorsi al limite" (dovrebbe essere chiaro, l'abbiamo ripetuto un sacco di volte) e, soprattutto siamo interessati alla maggiore o minore "convenienza" delle diverse descrizioni per mettere in luce cio' che ci interessa e non alla loro possibilita' teorica. In ogni caso tagliuzzare arbitrariamente una distribuzione continua in una pedante descrizione discreta con tante classi diminuisce di molto l'efficienza della descrizione e conseguentemente la sensibilita' dei test statistici che e' strettamente legata ad un numero di osservazioni congruo per ogni classe. Quindi, a parita' di osservazioni, molte classi poco rappresentate forniscono descrizioni poco efficaci, e l'approccio continuo e' sicuramente da preferire. La filosofia generale e' insomma quella di violentare il meno possibile la natura dei dati che ci si presentano (indipendentemente dal loro significato), cercando di aderire il piu' possibile alla loro struttura matematica.



Nell' esempio precedente abbiamo visto come possa essere importante la scelta della descrizione, nel caso dell'analisi dei sistemi complessi, in cui non abbiamo a disposizione teorie sufficientemente solide da indirizzare la nostra analisi su binari noti, l'attenta valutazione del tipo di descrizione che aderisca maggiormente al carattere dei dati e' l'ingrediente piu' importante per il successo del lavoro.
La convenienza di "Buscar el Levante para el Ponente" (come avrebbe detto Colombo) e quindi di "aumentare la sensibilita' con misure piu' grossolane" presuppone uno spazio non lineare (il globo terracqueo di Colombo opposto al mondo planare). La figura 6.1 ce ne mostra le ragioni: la diversa distribuzione che si osserva nei due pannelli fa si' che mentre nel pannello a) esiste la possibilita' di una scelta "naturale" di una soglia per dividere i punti sperimentali in due classi (a sinistra e a destra della freccia), nel pannello b) ogni divisione risulterebbe arbitraria e non giustificata dai dati, ma solo da considerazione esterne alla distribuzione. La nonlinearita' insomma e' strettamente legata alla generazione di divisioni qualitative, e quindi in qualche modo di classificazioni suggerite autonomamente dalle distribuzioni, e consente di passare da una descrizione continua dei dati ad una descrizione discreta secondo "classi omogenee" e separabili tra loro.

Cap. 7 Conclusioni

Nei capitoli precedenti abbiamo cercato di dare un'idea, forzatamente sommaria, del senso profondo della metodologia statistica, facendoci guidare da alcuni concetti basilari come quello di simmetria/asimmetria, di distanza e di correlazione.
Queste dispense dovrebbero servire ad impostare in maniera soddisfacente una ricerca, non importa se 'sul campo' od in 'laboratori' od anche semplicemente sfruttando dati gia' presenti in Internet.
Non esiste nessuna 'statistica particolare' per l'ecologia, la meccanica quantistica o la psicologia, semplicemente certe classi di problemi ricorreranno con frequenze relative diverse nei differenti ambiti sperimentali, certi test statistici saranno piu' popolari presso certi campi di ricerca piuttosto che altri, qui non siamo potuti ovviamente entrare nel dettaglio delle singole tecniche, esistono comunque ottimi supporti sia nei libri di testo che direttamente sulla rete ed i programmi applicativi per eseguire i differenti algoritmi statistici sono per larga parte gratuiti.
Rimango comunque convinto che saper costruire una buona matrice unita'/variabile dopo aver messo su delle misure coerenti con quanto si vuole osservare costituisca il 90% dell'impresa.