Medie e regressioni

Nelle precedenti lezioni ho cercato di descrivere per grandi linee le caratteristiche più importanti del sistema demografico europeo. Per tracciare i contorni di questo importante oggetto storico ho più volte fatto ricorso, in modo surrettizio, al concetto di media; ho, ad esempio, fatto riferimento a concetti come quello di età media al primo matrimonio, numero medio di figli avuti da una donna, età media al parto, età media alla morte (o speranza di vita), popolazione media, ecc. Queste misure su cui si è costruita l'analisi demografica tradizionale, e altre ancora che incontreremo nelle prossime lezioni, sono tutte «interpretazioni» particolari del concetto astratto e generale di «media». Comprendere la logica e i fondamenti teorici che sono alla base di questo tipo di misura risulta dunque un passaggio obbligato per entrare nel mondo delle analisi storico-sociali (e, se si vuole, per iniziarne una critica). Questa lezione sarà allora dedicata a comprendere l'insieme di procedure che hanno portato alla formazione di un concetto d'uso così frequente. Ciò che ci interesserà, contrariamente alla prassi abituale, sarà non tanto la procedura di calcolo che ci permette di passare da un insieme di osservazioni quantitative alla costruzione di una media (questo genere di operazione è molto semplice e non merita che gli venga dedicato lo spazio di una lezione), quanto piuttosto la teoria matematica che sta dietro al concetto di media. Conoscere da dove vengono le medie, capirne la logica e la filosofia, consentirà, infatti, di acquisire distacco e consapevolezza, rispetto ad un concetto intorno al quale storicamente sono venute costruendosi molte discipline che studiano l'uomo e il suo vivere in società. Questa presa di distanza, questo ripensamento critico di uno dei concetti intrinseci all'analisi sociale io credo siano oggi fondamentali per chi voglia tentare una soluzione alla crisi epistemologica che sempre più va diffondendosi nelle scienze umane.

Medie semplici e medie ponderate

Per introdurre in forma generale il concetto di media è sufficiente avere una strumentazione matematica limitata. Per descrivere completamente questo concetto sono infatti sufficienti quattro soli simboli matematici:

1) Il simbolo «$\in$» che indica la relazione di appartenenza ad un insieme ($o_i \in O$ indica che l'«elemento» o_i appartiene all'«insieme» O)

2) Il simbolo «$O \equiv \{o_1, o_2, \ldots, o_n\}$» che indica che gli elementi da o_1 a o_n sono «tutti e soli» gli elementi dell'insieme O.

3) La notazione «#O = n» che indica la «cardinalità» dell'insieme O, ovvero che il numero di elementi che lo compongono è uguale a *n*.

4) Infine la notazione «$\sum_{i=1}^{n} o_i$ » indica la sommatoria di tutti i valori dell'insieme *O*, ovvero:

$$\sum_{i=1}^{n} o_i = o_1 + o_2 +,...,+o_n$$

Una volta che si sia in possesso della notazione appena descritta si può definire il concetto di media come segue:

1) Sia dato un insieme *O* di osservazioni (misurazioni) per un dato carattere reale (ad esempio la statura)
2) Sia *n* la cardinalità di *O* (cioè il numero complessivo di osservazioni-misurazioni compiute su una data popolazione):

$$\#O = n$$

3) Sia o_i una singola osservazione in *O*:

$$o_i \in O \qquad \text{con } i \in \mathrm{I} \equiv \{1,\ldots,n\}$$

Si definisce allora «media» il valore:

1) $$\bar{o} = \frac{1}{n}\sum_{i=1}^{n} o_i$$

Se la popolazione che stiamo studiando è composta da tre soli individui (dunque P ≡ {a, b, c} dove a,b,c sono i tre individui che formano la popolazione P), e il carattere che stiamo misurando è la statura, potrà accadere che l'insieme S di tutte le misure compiute sulla popolazione P abbia la stessa cardinalità di P, dunque:

$$\#S = \#P = 3$$

Assumiamo ora che i tre individui di P siano alti 160, 170, 180 cm, potremo dunque scrivere:

$$S \equiv \{160, 170, 180\}$$

Allo stesso modo avremmo potuto scrivere:

$$S \equiv \{170, 160, 180\}$$

o anche:

$$S \equiv \{180, 170, 160\}$$

Senza per questo cambiare il contenuto delle nostre conoscenze sull'insieme S. L'ordine con cui vengono dichiarati in forma «estensiva» (cioè specificandoli uno per uno) gli elementi di un insieme non è rilevante per la sua definizione. Conosciamo dunque la composizione dell'insieme S, ne conosciamo la cardinalità, possiamo dunque calcolare la media:

$$\bar{s} = \frac{1}{3}\sum_{i=1}^{3} s_i \qquad (\text{con } s_i \in S)$$

Il che significa semplicemente:

$$\bar{s} = \frac{1}{3}(s_1 + s_2 + s_3) \qquad (\text{con } s_1, s_2, s_3 \in S)$$

che ci porta infine al risultato:

$$\bar{s} = \frac{1}{3}(160 + 170 + 180) = 170$$

Dunque la statura media degli individui della popolazione P è 170 cm. Si è in questo modo calcolata una media semplice. La "semplicità" di questo tipo di calcolo deriva dal fatto che l'insieme S di tutte e sole le misure della statura rilevate sulla popolazione P ha esattamente la stessa cardinalità di quest'ultimo insieme. I calcoli diventano appena un po' più complicati nel momento in cui si lasci cadere quest'ultima condizione. Per rappresentare anche questo secondo caso possiamo pensare di avere una popolazione composta da quattro individui - P ≡ {a, b, c, d} - per cui si sappia che gli elementi (individui) *a* e *b* abbiano statura di 160 cm, *c* di 170 e *d* di 180. Ciò che accadrà in tale situazione sarà che la cardinalità dell'insieme S delle osservazioni relative alla statura, e l'insieme P composto dagli individui della nostra popolazione, non avranno stessa cardinalità, infatti:

$$S \equiv \{160, 170, 180\} \rightarrow \#S = 3$$

$$P \equiv \{a, b, c, d\} \rightarrow \#P = 4$$

dunque:

$$\#S \neq \#P$$ (si legge: la cardinalità di S è diversa da quella di P)

Per giungere, in questo secondo caso alla determinazione della statura media degli elementi di P occorre inserire nel nostro ragionamento un nuovo simbolo matematico corrispondente ad un nuovo concetto insiemistico. Selezioniamo dunque all'interno di P tutti quegli elementi che abbiano statura 160 cm; questi elementi formeranno un nuovo insieme che chiameremo $P_1 \equiv \{a,b\}$. L'insieme P_1, gode di un'importate proprietà che lo lega indissolubilmente all'insieme P; risulta infatti sempre verificata l'affermazione secondo cui «*se* un dato elemento appartiene a P_1 *allora* tale elemento appartiene anche a P», il che può essere altrimenti espresso attraverso l'affermazione «per un elemento l'essere incluso in P_1 *è condizione sufficiente* per essere incluso anche in P», o ancora «P_1 è un *sottoinsieme* di P». Per esprimere il rapporto di inclusione di un dato insieme P_1 in un altro insieme P si userà il simbolo $\subseteq$:

$$P_1 \subseteq P \qquad (P_1 \text{ è un sottoinsieme di P})$$

Possiamo ora notare come l'operazione di misura della statura compiuta sugli individui della popolazione P identifica tre sottoinsiemi distinti (due sottoinsiemi si dicono distinti quando nessuno degli elementi del primo appartiene al secondo e viceversa). Possiamo allora chiamare con P_1 l'insieme degli elementi di P che hanno statura 160, con P_2 l'insieme di coloro che hanno statura 170, con P_3 coloro che hanno statura 180:

$$P_1 \equiv \{a,b\} \rightarrow \#P_1 = 2$$
$$P_2 \equiv \{c\} \rightarrow \#P_2 = 1$$
$$P_3 \equiv \{d\} \rightarrow \#P_3 = 1$$

Nel calcolare la statura media dei nostri quattro individui dobbiamo tenere conto del fatto che i sottoinsiemi di P_2 e P_3 - associati rispettivamente alle stature 170 e 180 - hanno "peso" (cardinalità) 1 mentre il sottoinsieme P_1 -associato alla statura 160 - ha "peso" (cardinalità) 2. Per giungere al calcolo della media in questa nuova condizione si procede "pesando" ogni singola determinazione del carattere S con il peso del sottoinsieme di P associato a tale determinazione:

2) $$\bar{s} = \frac{1}{\#P}\sum_{i=1}^{n} s_i \cdot \#P_i$$

Nel caso specifico della nostra popolazione composta da quattro individui il calcolo della media diventa allora:

$$\bar{s} = \frac{1}{\#P}\sum_{i=1}^{n} s_i \cdot \#P_i = \frac{1}{4}\left[(160\cdot 2)+(170\cdot 1)+(180\cdot 1)\right] = 167{,}5$$

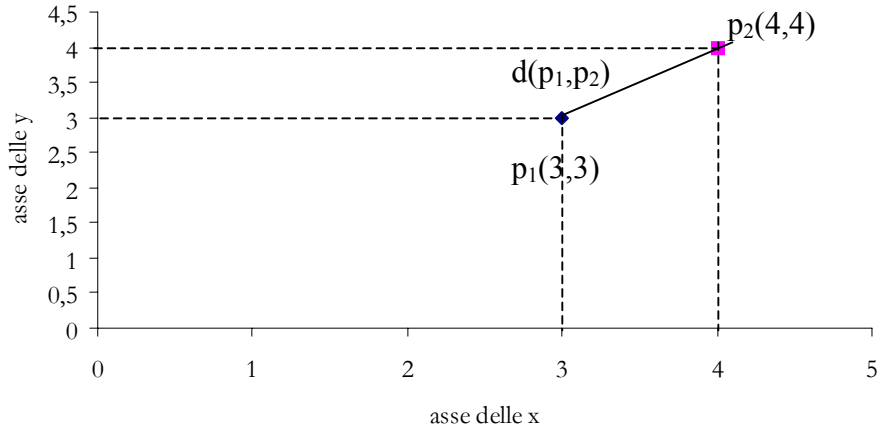
Le formula 1) con cui si è presentato il modo di calcolo di una media semplice, diviene dunque un caso particolare di ciò che abbiamo chiamato «media ponderata»; la media semplice è una media ponderata in cui tutti i pesi siano posti uguali a 1. Altrimenti detto la media semplice è quel caso particolare di media ponderata in cui i sottoinsiemi di P associati alle diverse determinazioni del carattere S siano composti dai singoli elementi di P.

Ci troviamo ora nell'imbarazzante situazione di sapere procedere al calcolo di una media attraverso la formula 2), ma di non sapere ancora esattamente cosa rappresenti il concetto di media. Perché, dunque, le medie vengono calcolate attraverso la formula 2), e come si è arrivati ad utilizzare tale formula e non un'altra per il calcolo delle medie? perché le medie hanno un ruolo così importante nell'analisi delle popolazioni? Per rispondere a tali domande cercherò, nelle prossime pagine, di fornire un'interpretazione "geometrica" del concetto di media.

L'interpretazione geometrica del concetto di media

Le medie mostrano una prossima parentela con il concetto di «distanza euclidea» fra due punti, per tale ragione inizierò definendo tale concetto, e poi passerò a mostrare la relazione che esiste fra il concetto di distanza e quello di media.

Graf. 1 Distanze Eudidee

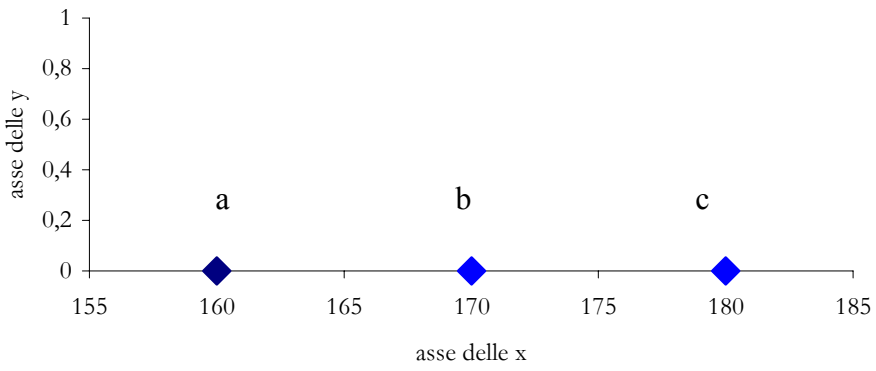


Il grafico 1 rappresenta su un piano cartesiano 2 punti: p_1 di coordinate (3, 3) e p_2 di coordinate (4, 4). Per calcolare la distanza che separa questi due punti - $d(p_1, p_2)$ - si utilizza una formula che riconduce tale problema al celebre teorema di Pitagora:

3) $$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{(3-4)^2 + (3-4)^2} = \sqrt{(1)^2 + (1)^2} = \sqrt{2} = 1{,}414$$

Dopo questa fondamentale scoperta torniamo a considerare il problema del calcolo della statura media di una popolazione. Torniamo, per comodità, al caso semplificato di una popolazione P composta da tre soli individui e classificata secondo tre differenti determinazioni del carattere (insieme) S. Supponiamo dunque di avere tre individui associati alle stature 160, 170 e 180 cm. Tali tre punti possono essere rappresentati attraverso un piano cartesiano, nello stesso modo appena visto per i punti p_1 e p_2 del grafico 1. Cosa accade, dunque, quando riportiamo le nostre tre altezze in un piano cartesiano?

Graf. 2 Rappresentazione delle stature di tre individui sull'asse dei numeri reali



Accade, come mostrato in figura 2, che tali punti vadano a collocarsi in luoghi differenti dell'asse delle x. Ciò accade, in effetti, poiché gli individui della nostra popolazione sono individui "monodimensionali", individui, cioè, che sono stati descritti attraverso un solo tipo di misura, la statura appunto. Essi sono associati ad un valore delle x, ma non sono associati ad alcun valore delle y. Si può allora considerare il valore delle y sempre uguale a zero, e ciò costringe i nostri tre individui a spostarsi "avanti e indietro" sul solo asse delle x fino a che essi non si fermino in corrispondenza del valore che rappresenta la loro statura. Poiché siamo in possesso del concetto di distanza euclidea fra due punti possiamo allora calcolare la distanza che separa l'individuo *a* dall'individuo *b*. Utilizziamo la formula impiegata per calcolare la distanza fra p_1 e p_2, ottenendo quanto segue:

$$d(a,b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{(160-170)^2 + (0-0)^2} = \sqrt{(10)^2 + (0)^2} = \sqrt{100} = 10$$

Qualcuno potrà aver notato che per trovare la distanza fra *a* e *b* sarebbe stato più facile compiere semplicemente la differenza fra le stature di *a* e di *b*:

170-160=10

Il problema è che, sebbene il numero così ottenuto sembri uguale a quello precedentemente ottenuto (in effetti lo è, rassicuro i lettori), esso tuttavia non è una distanza. Per definizione la distanza fra un punto *a* e un punto *b* è uguale alla distanza fra *b* e *a* (si dice che le distanze, tutte, anche quelle non euclidee, godono della proprietà di simmetria d(a, b) = d(b, a)). E' sufficiente dunque invertire i termini della differenza 170-160=10 e compiere l'operazione 160 - 170 = -10 per vedere che la distanza euclidea non può essere definita, nemmeno nel caso di osservazioni monodimensionali, come la differenza fra due valori della x. Anche nel caso di osservazioni monodimensionali converrà dunque utilizzare la formula 3) per giungere al calcolo delle distanze fra i nostri punti.

Che c'entra tutto ciò con le medie? La parentela fra il concetto di distanza euclidea e quello di media e data, nell'interpretazione geometrica che stiamo seguendo, dal fatto che una media altro non è che quel punto che risulta "globalmente" essere "più vicino" ai valori delle nostre osservazioni. La media μ è dunque quel punto dell'asse x che «minimizza» la somma delle distanze dai tre punti *a*, *b* e *c*. Più analiticamente:

4) $$\mu : \min \sum_{i=1}^{n} \sqrt{(\mu - s_i)^2}$$

E' evidente che tale quantità è minima quando la quantità $\sum_i(\mu - s_i)^2$ è minima, dunque, per semplificare i calcoli posso riscrivere la 4) come segue:

5) $$\mu : \min \sum_{i=1}^{n} (\mu - s_i)^2$$

La definizione data dalla proposizione 5) (μ è tale che la somma delle distanze al quadrato calcolate fra μ e le diverse osservazioni risulti minima) è in effetti la definizione generale di media.

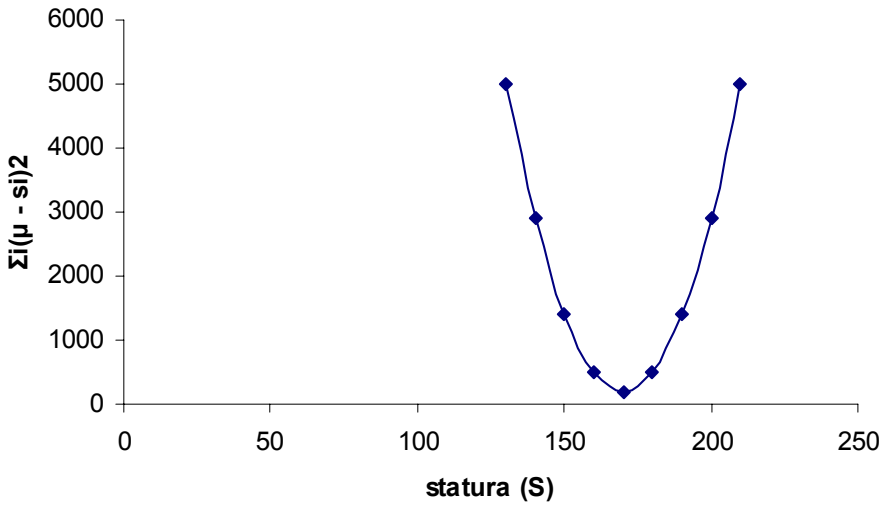
Vediamo ora come dalla definizione generale di media data dalla proposizione 5) si possa giungere alla formula impiegata nel primo paragrafo di questa dispensa. Il vero problema della definizione data attraverso la proposizione 5) è legato al fatto che non possediamo (ancora) degli strumenti che ci permettano di trovare il valore minimo per la funzione $y = \sum_i(\mu - s_i)^2$. Tenteremo allora di giungere al valore minimo di tale funzioni per tentativi (o per approssimazione, come dicono i matematici); sceglieremo quindi a caso dei valori, li inseriremo nella formula $y = \sum_i(\mu - s_i)^2$ e verificheremo passo dopo passo quale sia il valore che minimizzi tale funzione. Cominciamo allora assumendo che la media per le tre osservazioni 160, 170, 180 sia μ = 130 (già sappiamo che tale valore non è la media, ma faremo finta, per il momento, di non saperlo). Introduciamo tale valore nella nostra formula ottenendo quanto segue:

$$\sum_{1}^{3}(130-s_i)^2 = (130-160)^2 + (130-170)^2 + (130-180)^2 = 5.000$$

Introducendo il valore 130 nella formula abbiamo dunque ottenuto il risultato 5.000; ripetiamo ora tale tipo di procedura per i valori 140, 150, ..., 210 fino ad ottenere la tabella seguente:

μ	$y = \sum_i(\mu - s_i)^2$
130	5.000
140	2.900
150	1.400
160	500
170	200
180	500
190	1.400
200	2.900
210	5.000

Proviamo ora a rappresentare su un piano cartesiano l'insieme di coppie di valori così ottenuti:



Dalla formula $y = \sum_i(\mu - s_i)^2$ è uscita fuori, come dal cilindro magico, una parabola, e in effetti è facile verificare come questa formula sia nient'altro che un modo inabituale di presentare un'equazione di secondo grado. Per dimostrarlo è sufficiente effettuare pochi semplici calcoli:

6) $$y = \sum_{i=1}^{n}(\mu - s_i)^2 = \sum_{i=1}^{n}(\mu^2 - 2\mu s_i + s_i^2) = n\mu^2 - 2\mu\sum_{i=1}^{n} s_i + \sum_{i=1}^{n} s_i^2$$

si osservi ora come la quantità $2\mu\sum_i s_i$ sia una costante (nel caso del nostro problema tale quantità è uguale a 2x3x(160+170+180)= 3.060) Posso dunque sostituire tutti questi complicati simboli con il simbolo «b», che per noi sarà qui uguale al valore 3.060. Allo stesso modo anche il simbolo $\sum_i s_i^2$ che rappresenta il terzo termine della nostra formula è una costante: tale valore è uguale a $160^2+170^2+180^2$=86.900. Anche in questo secondo caso possiamo sostituire al simbolo più complesso il più semplice «c». Infine anche il simbolo *n* della formula è una costante che esprime la cardinalità dell'insieme S, rappresentiamolo, in ossequio alla tradizione, anche se non ce ne sarebbe bisogno, con il simbolo «a». Effettuiamo queste sostituzioni nella formula 6) ottenendo:

7) $$y = a\mu^2 - b\mu + c$$

in cui sarà facile distinguere un'equazione di secondo grado (una parabola) con variabile μ.

Torniamo ora al nostro problema principale. Poiché sappiamo che per definizione una media è quel valore che minimizza la funzione $y = \sum_i(\mu - s_i)^2$, poiché inoltre sappiamo che tale funzione rappresenta una parabola convessa, allora potremo affermare che il punto di minimo di tale funzione coincide con il vertice della parabola. Poiché l'ascissa del vertice per una generica parabola è data dalla formula:

8) $$\mu = \frac{-b}{2a}$$

inserendo i parametri dell'equazione 7) (- *b* e *a*) nella 8) ottengo

9) $$\mu = \frac{b}{2a}$$

Ricordando poi che i simboli *a* e *b* dell'equazione 7) stanno rispettivamente per *n* e $2\mu\sum_i s_i$ si ottiene:

10) $$\mu = \frac{2\sum_{i=1}^{n} s_i}{2n} = \frac{1}{n}\sum_{i=1}^{n} s_i$$

otteniamo dunque la definizione di media dalla quale eravamo partiti con la formula 1. Abbiamo dunque dimostrato che la formula 1 con cui abbiamo calcolato le medie del paragrafo precedente

esprime quel valore che minimizza la formula 5). Una media è quel punto che minimizza la distanza globale dalle diverse osservazioni compiute su una popolazione. Nell'interpretazione geometrica che stiamo seguendo una media è anche chiamata baricentro

Una generalizzazione del concetto di media: le regressioni lineari

Finora abbiamo condotto sulla popolazione P un tipo di osservazione molto semplice; per ogni suo individuo ci siamo limitati a osservarne la statura. Non esiste tuttavia alcun limite al numero di osservazioni che si possono compiere su una data popolazione. Ogni individuo di questa popolazione, in altri termini, potrebbe essere descritto attraverso una serie virtualmente illimitata di numeri ciascuno dei quali rappresenti una singola misura per un dato carattere reale. Oltre alla statura, potremmo, in uno slancio positivistico, misurare il peso, la circonferenza toracica, il volume cranico, la pressione sanguigna, l'età ecc. Naturalmente per ciascuno di tali valori singolarmente preso possiamo calcolare una media seguendo la procedura che si è definita nel precedente paragrafo. Tuttavia, in qualche modo, è possibile anche procedere al calcolo di particolari tipi di medie che si applichino alle misure prese a due alla volta, a tre alla volta ecc. Invece di immaginare l'esistenza di due caratteri separati di cui l'uno abbia nome «statura» e l'altro «peso», possiamo immaginare di "fondere" insieme tali due caratteri e di ottenere il carattere "combinato" statura-peso, oppure quello peso-statura (che sono due differenti caratteri combinati). In queste dispense ci limiteremo a considerare il caso ancora relativamente semplice dell'osservazione di due caratteri su una popolazione. Supponiamo dunque di avere una popolazione di tre soli individui - $P \equiv \{a, b, c\}$ - a cui siano stati misurati la statura, ottenendo rispettivamente i valori di 160, 170 e 180 cm, e il peso, ottenendo come risultato 50, 60 e 80 kg. Possiamo ora rappresentare tali risultati attraverso una serie di «coppie ordinate» del tipo: (160,50), (170, 60), (180, 80). Con la prima di tali coppie ordinate - (160, 50) - si rappresenta il fatto elementare che l'individuo *a* della popolazione P sia alto 160 cm e abbia peso di 50 kg. Ciò che accade al nostro insieme O di tutte e sole le osservazioni ottenute attraverso la misurazione degli individui di P è che esso non è più formato da singoli valori, bensì da coppie ordinate del tipo di quelle appena viste, risulterà dunque:

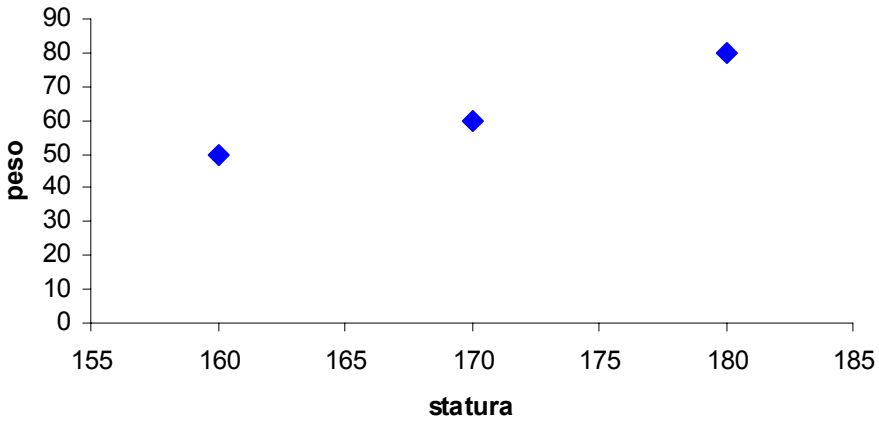
$$O \equiv \{(160, 50), (170, 60), (180, 80)\}$$

Un insieme che sia composto da coppie ordinate e non da singoli elementi nella terminologia insiemistica viene detto «relazione». Una relazione, in senso generale e astratto, è dunque un insieme di coppie ordinate. La relazione specificata dall'insieme O è la relazione che sussiste fra la statura e il peso degli individui della popolazione P. E in effetti, ciò che cercheremo di quantificare nelle prossime pagine è la forza di questa relazione; in quale misura, cioè, a partire dalla conoscenza della statura di un

individuo, saremo in grado di "prevederne" il peso. La logica che seguiremo in questo tentativo, come si vedrà, è in tutto simile a quella che abbiamo seguito nel caso del calcolo delle medie.

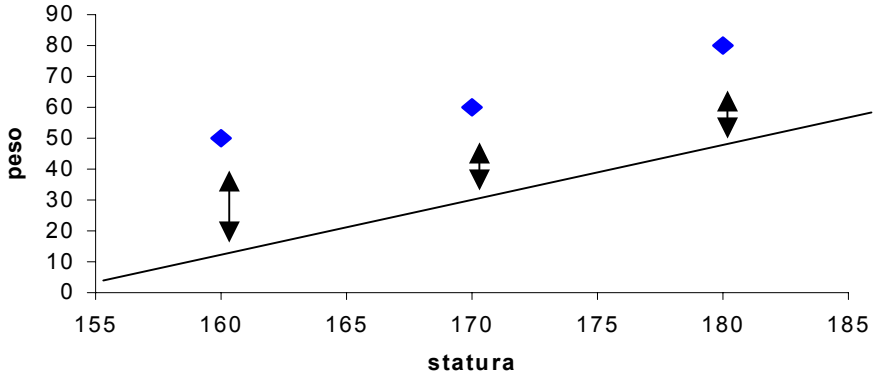
Iniziamo rappresentando, come a sua volta facemmo per i valori dell'insieme S, i valori dell'insieme O sul piano cartesiano:

Graf. 4 Relazione fra statura e peso



Nel caso di osservazioni monodimensionali come quelle legate alla misurazione della sola statura il problema era costituito da trovare quel punto che minimizzasse globalmente la distanza con le diverse osservazioni. Ora che siamo di fronte ad un caso bidimensionale (ogni individuo è associato alla misurazione di due differenti caratteri) il problema che ci porremo sarà quello di trovare la retta che permetta di rendere minimo lo scarto con le osservazioni doppie compiute nella nostra popolazione.

Graf. 5 Scarti fra la retta di regressione e le osservazioni compiute nella popolazione



Per trovare lo scarto fra uno dei punti del piano cartesiano e la retta di regressione ci avvarremo ancora una volta del concetto euclideo di distanza fra due punti. Se dunque la retta che cerchiamo di far passare il più vicino possibile alle osservazioni ha equazione generica:

$$r = mx + q$$

dove *m* indica il «coefficiente angolare» della retta (la sua inclinazione) e *q* il suo «termine noto», il punto cioè in cui la retta interseca l'asse delle y, allora la distanza del nostro primo punto (160, 50) dalla retta sarà[1]:

11) $$d(p_1, r) = \sqrt{[50 - (m \cdot 160 + q)]^2}$$

La retta *r* che minimizzerà la somma di tutte le distanze che è possibile calcolare fra questa retta e i diversi punti che rappresentano sul piano cartesiano le nostre osservazioni doppie sarà allora:

12) $$\text{r: } \min \sum_{i=1}^{n} d(p_i, r) = \min \sum_{i=1}^{n} \sqrt{[y_i - (m \cdot x_i + q)]^2}$$

Come nel caso delle medie si comprende immediatamente come la quantità $\sqrt{[y_i - (m \cdot x_i + q)]^2}$ sarà minima quando lo sarà $[x_i - (m \cdot x_i + q)]^2$. Possiamo dunque semplificare la 12) scrivendo:

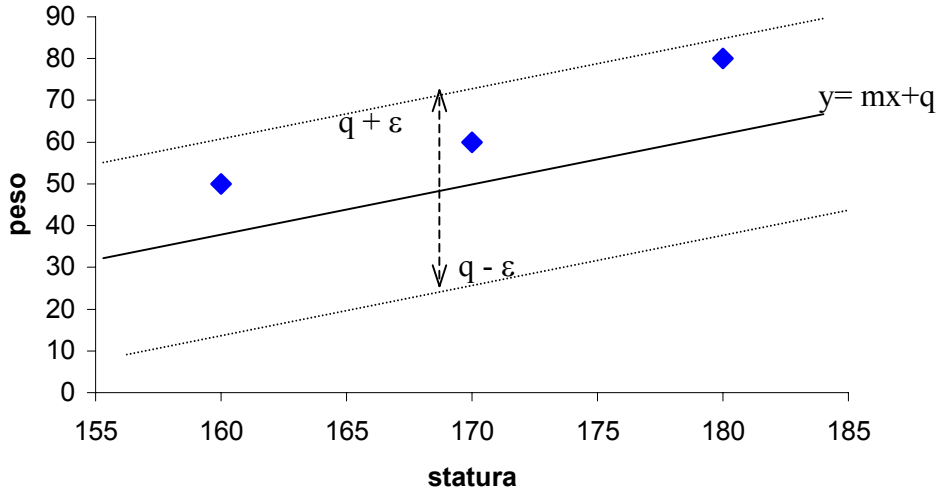
13) $$\text{r: } \min \sum_{i=1}^{n} d(p_i, r) = \min \sum_{i=1}^{n} [y_i - (m \cdot x_i + q)]^2$$

che è in effetti la definizione generale di retta di regressione. Trovare il valore che minimizzi l'equazione descritta nella 13) non è altrettanto agevole del caso incontrato per le medie. Allora fummo facilitati dalla possibilità di riconoscere nella funzione da minimizzare una parabola convessa per cui sapevamo che il punto di minimo andava a collocarsi nel vertice della parabola. In questo caso il problema è più complicato, perché sebbene l'equazione $y = \sum_i [y_i - (mx_1 + q)]^2$ sia in effetti ancora l'equazione di una

[1] A rigor di termine non è esatto affermare che la distanza $d(p_1, r)$, così come essa viene calcolata dal equazione 11, sia la distanza fra il punto p_1 e la retta r. La distanza fra un punto e una retta è, infatti, data dal più breve tragitto che separa il punto dalla retta, e che nel caso di una geometria euclidea è quello che si trova sulla retta normale (perpendicolare) a r passante per p_1. Nel caso descritto dalla 11 stiamo invece calcolando la distanza fra il punto p_1 e quel punto della retta r che ha stessa ascissa di p_1. Tale distanza non necessariamente coincide con il concetto di distanza di un punto da una retta. Nonostante questa imprecisione, nel seguito di questo testo continuerò a utilizzare in senso improprio l'espressione «distanza del punto p_1 dalla retta r» per indicare la distanza fra questo punto e quello con medesima ascissa giacente sulla retta r, perché ciò semplifica il linguaggio rendendolo più piano.

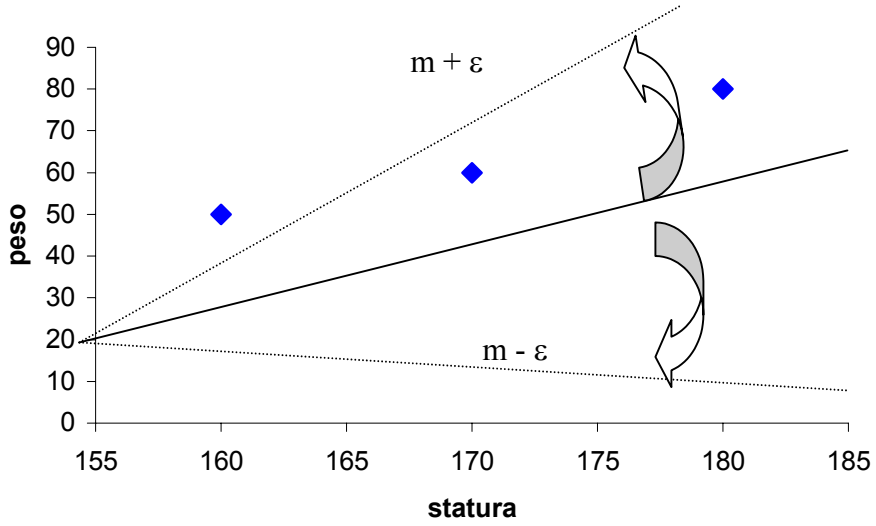
parabola, essa è però una parabola in uno spazio a tre dimensioni. Questa equazione ha, in effetti, due variabili (*m* e *q*) ciascuna delle quali può variare indipendentemente dall'altra. Se teniamo ferma, ad esempio, la variabile *m* e facciamo variare il parametro *q* questo indurrà nella retta di regressione una traslazione in "alto" o in "basso", come mostrato nel grafico 6:

Graf. 6 variazioni di *q* tenuto fermo *m*



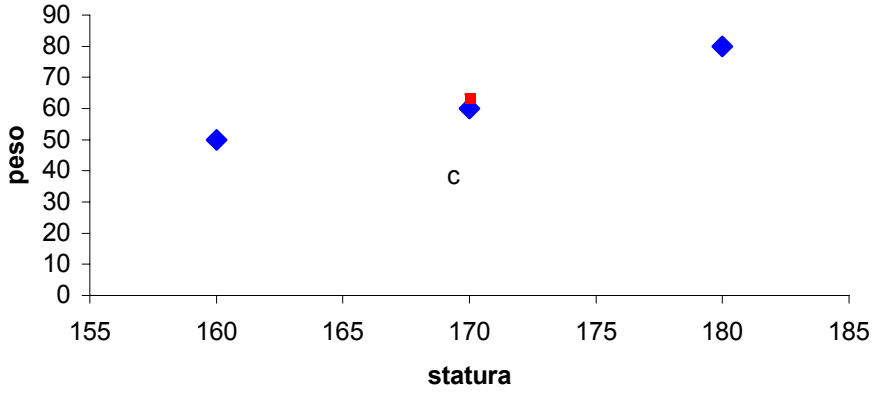
Se al contrario facciamo variare la variabile *m* tendo ferma la variabile *q*, otteniamo una rotazione della retta *r*:

Graf. 7 variazioni di *m* tenuto fermo *q*



Ciò vuol dire che per minimizzare la funzione 13) dobbiamo agire simultaneamente sia sul parametro *m* sia su *q*. Per giungere a questo risultato si impiegano di norma semplici metodi dell'analisi differenziale (si risolve il sistema delle derivate parziali dell'equazione 13) poste uguali a zero). Eviteremo tuttavia questo metodo. Per aggirare l'ostacolo e metterci comunque in grado di risolvere il nostro problema di trovare quella retta che passi il più vicino possibile all'insieme di punti che rappresentano le nostre osservazioni, ci serviremo di un piccolo trucco. Tale piccolo trucco si risolve, in primo luogo, nel trovare il baricentro per i nostri tre punti. Ricordo che i tre individui della popolazione P che stiamo analizzando hanno altezza 160, 170, 180 e altezza media 170, mentre essi hanno peso rispettivamente di 50, 60 e 80 kg il che ci porta ad un peso medio di 63,33 kg. Dunque l'individuo medio con cui possiamo rappresentare la nostra popolazione ha statura 170 cm e peso 63,33 kg. Rappresentiamo ora tale individuo fittizio sul nostro grafico:

Graf. 8 Relazione fra statura e peso



Il quadratino rosso del grafico 8 rappresenta il nostro individuo medio (come si vede l'individuo reale *b* ha caratteristiche molto prossime a quelle dell'individuo medio della popolazione). Come si vede l'individuo medio non è altro che quel punto del piano che riporta in ascissa il valore della statura media, e in ordinata il valore del peso medio. Tale punto, che viene detto baricentro o centroide delle nostre osservazioni, gode di un'importante caratteristica. La retta di regressione passa necessariamente per tale punto. Questo ci permette di scartare dal numero di tutte le rette che attraversano il nostro piano, tutte quelle che non hanno la caratteristica di passare per il centroide delle nostre osservazioni. La nostra retta perde uno dei suoi gradi di libertà; essa potrà continuare a ruotare intorno al centroide, ma non potrà più essere traslata verso l'alto o verso il basso. Se ora chiamiamo con il simbolo $\bar{x}$ la statura media dei nostri individui, e con $\bar{y}$ il loro peso medio, potremo affermare che se la retta di regressione passa per il centroide, allora varrà la seguente equazione:

$$\bar{y} = m\bar{x} + q$$

Da cui in modo molto semplice stabiliamo che:

14) $$q = \bar{y} - m\bar{x}$$

Siamo dunque riusciti a determinare il termine noto della retta di regressione a patto di riuscire a determinarne il coefficiente angolare. Possiamo dunque ora sostituire il valore di *q* appena ottenuto nell'equazione 13) ottenendo:

15) $$r : \min \sum_{i=1}^{n} [y_i - (m \cdot x_i + \bar{y} - m\bar{x})]^2$$

Siamo così riusciti ad eliminare dall'equazione 13) la variabile *q* cosicché una complicata parabola in tre dimensioni si è trasformata d'incanto in una mite parabola convessa a due dimensioni per cui trovare il punto di minimo significa semplicemente essere in grado di determinarne il vertice. Procedo a dimostrare le due affermazioni appena fatte:

$$y = \sum_{i=1}^{n} [y_i - (m \cdot x_i + \bar{y} - m\bar{x})]^2 =$$

$$\sum_{i=1}^{n} [y_i - m \cdot x_i - \bar{y} + m\bar{x})]^2 =$$

$$\sum_{i=1}^{n} [(y_i - \bar{y}) - m(x_i - \bar{x})]^2 =$$

$$\sum_{i=1}^{n} m^2 (x_i - \bar{x})^2 - 2m(y_i - \bar{y})(x_i - \bar{x}) + (y_i - \bar{y})^2 =$$

16) $$m^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2m \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Come nel caso precedentemente visto per le medie otteniamo un'equazione di secondo grado con variabile *m*, come si verifica facilmente ponendo il termine $\sum_{i=1}^{n} (x_i - \bar{x})^2 =$ a, il termine $-2\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) =$ b, e ponendo infine il termine $\sum_{i=1}^{n} (y_i - \bar{y})^2 =$ c. Inserendo nell'equazione 15) le sostituzioni appena descritte otteniamo, infine

17) $$y = am^2 + bm + c$$

cioè l'equazione di una parabola. Poiché sappiamo che il punto di minimo di tale funzione si trova nel vertice della parabola, e poiché sappiamo che l'ascissa di tale punto è dato dalla formula m = -b/2a, ricordando che nel caso presente a = $\sum_{i=1}^{n}(x_i-\bar{x})^2$ e b = $-2\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$, otteniamo:

$$m=\frac{2\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{2\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Da cui semplificando al numeratore e al denominatore si ottiene:

18) $$m=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Sappiamo dunque ora dalla 18) quale debba essere il coefficiente angolare della retta di regressione e dalla 14) quale sia il suo termine noto. Disponiamo dunque di tutte le informazioni necessarie e sufficienti per giungere ad individuare la retta di regressione per le nostre osservazioni. Sappiamo dunque che per la nostra popolazione la statura è data da $x_1 = 160$, $x_2 = 170$, $x_3 = 180$ e che la statura media è $\bar{x}=170$. Allo stesso modo sappiamo che il peso è dato da $y_1 = 50$, $y_2 = 60$, $y_3 = 80$ e che il peso medio è $\bar{y}=63{,}33$. Inserendo tali valori nell'equazione 18) otteniamo allora:

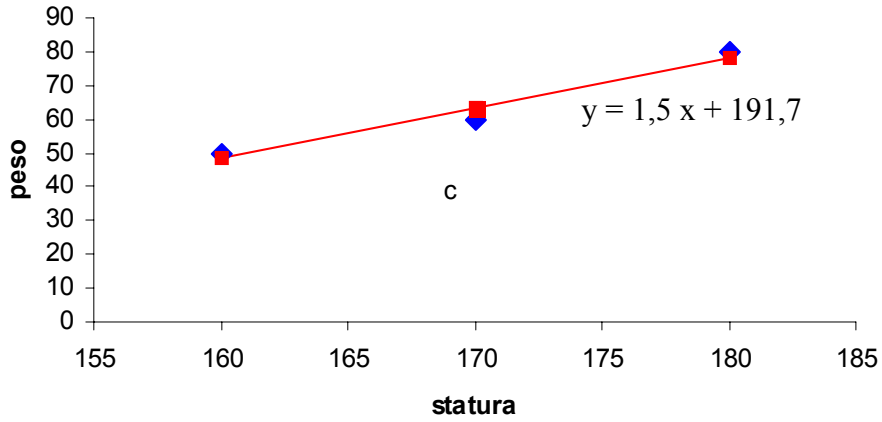
$$m=\frac{(50\cdot160-170\cdot63{,}33)+(170\cdot60-170\cdot63{,}33)+(180\cdot80-170\cdot63{,}33)}{(160-170)^2+(170-170)^2+(180-170)^2}=1{,}5$$

Dunque la nostra retta di regressione avrà coefficiente angolare pari a 1,5; risulterà dunque y=1,5x+q. Per determinare il valore del parametro *q* ci serviamo ora dell'equazione 14), ottenendo:

$$q=63{,}33-1{,}5(170)=191{,}67$$

Dunque la nostra retta di regressione avrà equazione y=1,5x+191,7. Rappresentando tale retta attraverso un piano cartesiano si ottiene:

Graf. 9 Relazione fra statura e peso



La retta di regressione è di potente ausilio nella ricerca sociale poiché, in qualche modo, permette di compiere delle previsioni sulle determinazioni assunte da un dato carattere (ad esempio il peso) in una popolazione a partire dalle conoscenze che si hanno su un altro carattere. Per esempio diviene possibile prevedere quanto peserà un individuo alto 167 cm. Sarà sufficiente inserire tale valore nell'equazione che rappresenta la retta di regressione (y=1,5x+191,7) e osservare che risultato viene fuori:

$$y=1,5(167)+191,7 = 58,83$$

Dunque un individuo alto 167 cm pesa all'incirca 59 kg. Tale tipo di previsione è reso possibile dal fatto che, in effetti, i due caratteri che stiamo esaminando sono fra loro legati da una stretta relazione che vuole, in generale, che un individuo alto pesi di più di uno basso. Non sempre tuttavia accade che la conoscenza di un certo carattere permetta di prevedere l'andamento di un secondo carattere. Se tale forma di "previsione" risulta impossibile allora si dice che i due caratteri sono fra loro «indipendenti» (è questo uno dei concetti più importanti del calcolo delle probabilità e della statistica matematica), come potrebbe accadere se scegliessimo come variabili, che so, la statura e il colore degli occhi. Per misurare l'efficienza dell'operazione di previsione della determinazione di un dato carattere a partire dalla conoscenza del valore assunto da un altro carattere si utilizza una misura che si chiama «correlazione» che ci permette di misurare quanta della "variabilità" di un dato carattere è stata "catturata" dall'operazione di regressione. Nel caso appena considerato tale valore giunge fino al livello del 98 per cento, il che ci permette di affermare che il 98 per cento della variabilità osservata nel peso degli individui della nostra popolazione immaginaria può essere spiegata a partire dalla variabilità della statura dei nostri individui.

E' importante sottolineare, per concludere questa sezione legata alla definizione generale dei concetti di media e di regressione lineare, che lo stesso tipo di ragionamento che si è seguito sia nel calcolo della media, sia in quello della retta di regressione, può essere esteso al caso di un numero virtualmente illimitato di variabili. Nel caso di tre variabili si parlerà, ad esempio di un piano di regressione rispetto al quale minimizzare la distanza di punti disposti in uno spazio a tre dimensioni. Insomma, anche se il numero di variabili cresce e passiamo progressivamente da uno spazio monodimensionale, come per la media, ad uno spazio a due dimensioni, come nel caso della retta di regressione, o a uno spazio a tre dimensioni qualora si considerino tre carratteri differenti di una popolazione, il tipo di ragionamento che si segue è sempre lo stesso, quello cioè di minimizzare le distanze euclidee rispetto di volta in volta al punto, alla retta o al piano che si sta considerando.Questo fa di questa parte della statistica, in fondo, un approccio un po' monotono.

Il problema delle medie

Le medie sono insetti ubiqui nel mondo delle discipline storiche e sociali; sono blatte con cui conviviamo quotidianamente, e alle quali abbiamo finito per abituarci. Ormai le medie sono di uso talmente frequente che il loro utilizzo è, per così dire, divenuto meccanico, un riflesso condizionato. Si ottengono dei dati su una certa popolazione, ed ecco che il primo compito che ci si propone è quello di misurare la media di questo e di quel carattere, di regredire questa variabile rispetto a quell'altra variabile, e così via. Davvero il lavoro di ricerca nel campo delle discipline storico-sociali sembra ormai coincidere con quello del calcolo di medie più o meno complicate, in situazioni più o meno intricate, al fine di riassumere, sintetizzare, derivare i caratteri di una data popolazione. Nel fare questo, si è cercato di mostrarlo nei precedenti paragrafi, è all'opera, continuamente e monotonamente, sempre lo stesso tipo di procedura concettuale. Questa procedura è una procedura di minimizzazione che ci consente di collocare a seconda dei casi, un punto, una retta, un piano il più vicino possibile all'insieme delle nostre osservazioni, nella convinzione che così facendo sia possibile sintetizzare in pochi valori (la media appunto, o i parametri di regressione) un insieme eterogeneo di dati compiendo l'errore più piccolo possibile. Errori, comunque, se ne commetteranno sempre: solo in casi degeneri una media o una retta di regressione andranno a coincidere esattamente con l'insieme delle osservazioni. Ed ecco, gradualmente, emergere la filosofia, la struttura paradigmatica, che si nasconde dietro tali concetti; l'idea di fondo è allora che l'errore appartenga alle cose umane, che nessuna comprensione è perfetta, che necessariamente la realtà debba essere approssimata, avvicinata, ma che essa, in fondo, alla fine, riesca sempre a sfuggire alla nostra analisi. L'unica cosa che si possa fare è limitare, per quanto possibile, l'entità dell'errore, minimizzandolo attraverso i metodi che si sono visti. Si finisce dunque per separare due diverse forme di comportamento: la media, la norma, da una parte, i comportamenti devianti, dall'altra. La media è ciò che ci permette di cogliere ciò che avviene «per lo più» - come scrive Aristotele

nel De Partibus - nelle popolazioni. E' ciò che ci consente di separare ciò che è normale da ciò che non lo è riducendo al minimo i costi di una tale separazione. E' dunque legittimo chiedersi se all'interno delle popolazioni, per così dire *in re*, esistano effettivamente delle medie, dei comportamenti normali e dei comportamenti devianti. Se sia lecito, cioè appiattire un certo aggregato demografico, una certa società, su pochi valori caratteristici, oppure se tale operazione non finisca per nascondere la parte più intima e più reale di una popolazione, ovvero la propria diversificazione interna.
Dare una risposta semplice a tali questioni non è, per me, cosa semplice. Ritengo però che chiunque decida di dedicarsi a degli studi storici-sociali debba porsi questa questione e tentare di darne una soluzione. Oggi, si può dire, esistono due differenti forme di risposta a questo tipo di domanda. La prima risposta, per semplificare, è di tipo "positivitico" e ritiene che non si può porre ordine all'interno della matassa intricata di una popolazione senza estrarre degli indici, come le medie appunto, che permettano di condensare l'enorme volume di informazione in alcune sue caratteristiche essenziali. La seconda risposta è di tipo invece "post-modernista" o "post-strutturalista" e afferma, in modo in verità sottile, che tutte le volte che poniamo, attraverso medie e cose del genere, una distinzione fra comportamenti "normali" e "devianti" stiamo in realtà tentando di creare individui "normali" e individui "devianti"; cioè stiamo tentando di esercitare un controllo sulla popolazione, stiamo tentando di darle una forma, e in tal senso esercitiamo un potere guidati da una finalità politica. Entrambi queste risposte mi appaiono insoddisfacenti, incomplete, in fondo, troppo elementari. Io tenterò qui di dare una mia personale (e salomonica) risposta a questo quesito, ricordando tuttavia che si tratta di una mia interpretazione e che ciascuno deve sentirsi libero di dare la sua personale risposta al problema delle medie.

Sono esistiti comportamenti normali e devianti nelle popolazioni del passato, in quelle stesse popolazioni delle quali abbiamo tentato di delineare la storia nelle passate lezioni? beh... in qualche modo sembra di sì. Quando abbiamo analizzato il microcosmo parrocchiale abbiamo visto come in effetti il parroco fosse in grado di controllare il comportamento matrimoniale dei propri parrocchiani in modo tale da far rispettare le disposizioni tridentine che stabilivano che il matrimonio dovesse avvenire fra individui che si trovassero ad una distanza di almeno 4 gradi canonici di parentela. Il comportamento normale, in tale situazione, è dato dallo sposarsi con individui che siano imparentate oltre il quarto grado, il comportamento deviante è invece quello che porta il matrimonio all'interno della sfera del quarto grado canonico. Ciò cui si deve tuttavia fare attenzione è il fatto che la possibilità di distinguere i matrimoni in normali e devianti, non è un fatto assoluto. La norma è stabilita da una struttura sociale che ha stabilito delle regole e ha poi costruito una serie di strumenti di controllo che permettono di verificare che le regole non vengano troppo spesso infrante. Nella seconda dispensa di questo corso abbiamo imparato a chiamare tale tipo sistemi di controllo con il nome di «sistemi

omeostatici», ciò di cui ci accorgiamo ora è che la possibilità di calcolare una media sembra essere legata alla possibilità di riconoscere tale tipo di sistema in una popolazione. Sembra dunque avere un senso parlare di medie quando si voglia conoscere quale sia il punto intorno al quale il sistema oscilla, e quale la sua capacità di mantenere le oscillazioni subite vicine al punto medio. Sembra, in altri termini, avere senso calcolare il grado medio di parentela degli individui che si sposano in una data comunità, perché tale valore descrive l'entità del controllo esercitato dal parroco sulla sua popolazione. Sembra allo stesso modo avere senso confrontare l'età media al primo matrimonio per un'entità demografica come il sistema demografico europeo e per il sistema demografico indiano, perché lo scarto fra tali misure ci consente di individuare l'azione opposta di due diversi sistemi di controllo di cui il primo tende a ritardare il matrimonio, mentre il secondo tende ad anticiparlo. Vedete, abbiamo cambiato prospettiva; le medie non servono più per descrivere in sintesi i comportamenti degli individui, quanto piuttosto per misurare il grado di controllo esercitato sul comportamento di quegli stessi individui da qualche istituzione sociale. Le prospettive positivista e post-strutturalista si sono fuse.

Cosa accade tuttavia se il carattere, o i caratteri sociali che stiamo studiando non sono sottoposti ad alcun controllo in una data popolazione. Sappiamo, che date tali condizioni, il sistema comincerà ad evolvere e a diversificarsi rispetto alla sua condizione iniziale. In tale situazione, ciò che appare rilevante per il processo, non è più la misura del controllo esercitato sulla popolazione, quanto piuttosto le singole variazioni che il sistema comincia a evidenziare rispetto alla sua condizione iniziale. Ciò che diviene interessante, in altri termini, è proprio l'apparizione di quei comportamenti devianti che l'uso delle medie e delle regressione tende a voler ridurre, nascondere, minimizzare. In tale condizione, nella condizione cioè in cui si stia studiando un fenomeno evolutivo, sembra illecito calcolare dei valori medi per i caratteri che si stanno considerando. Non appare tuttavia illecito calcolare delle medie sul disordine progressivo che si diffonde nella popolazione. L'entropia, così come essa è stata definita dalla termodinamica o dalla teoria dell'informazione, e che abbiamo detto essere lo strumento con cui si misura la velocità evolutiva di una popolazione, altro non è che una strana forma di media; è una misura attraverso cui si calcola il valore medio di disordine in una popolazione, la media dei comportamenti devianti. Per fare un esempio torniamo al problema dei matrimoni. Si è detto che il controllo esercitato dal parroco, per molti secoli è riuscito a mantenere il matrimonio fra due individui oltre la soglia del quarto grado di parentela. Ad un dato punto della storia europea - collocabile approssimativamente fra metà Settecento e l'inizio dell'Ottocento - tale controllo sembra perdere progressivamente di forza. Il numero dei "processetti matrimoniali" in tale epoca comincia a crescere rapidamente mostrando come quote sempre crescenti di popolazione (comunque sempre basse in valori assoluti) si orientino verso matrimoni endogamici. E' questo uno degli indizi più importanti che ci permettono di identificare la rottura dell'antico sistema demografico europeo. In tale condizione saremo interessati non tanto a conoscere quale sia il comportamento normale dei nostri individui (perché tale espressione perde

progressivamente di significato in una situazione senza controlli esterni), quanto la rapidità con cui i comportamenti devianti prendono piede progressivamente nella popolazione; saremo interessati a conoscere con quale rapidità il disordine si diffonde nella popolazione.

Ecco allora che nella mia interpretazione del problema delle medie, la soluzione viene data dal tornare ai due concetti di omeostasi e di evoluzione: a partire da questi due concetti, che sono concetti dell'analisi sociale, si può far discendere la prassi del lavoro di ricerca stabilendo di volta in volta se sia opportuno utilizzare le medie oppure no.