BIBLIOTECA DOCUMENTAZIONE RELAZIONI INTERNAZIONALI

quaderni di discussione

85.04

Analisi esplorativa dei dati "Esposizione di alcune tecniche per la investigazione dei dati"

Pierpaolo Napolitano *

istat

I quaderni di discussione sono a circolazione ristretta e non impegnano la responsabilità dell'ISTAT ma riflettono solo il punto di vista degli autori. Non possono, quindi, essere citati e fatti circolare senza il permesso degli autori.

Le richieste vanno indirizzate a :
«ISTAT - Centro Documentazione - Dr.ssa Borgnino-Valenzano
Via Balbo, 16 - 00100 - ROMA

85.04

Analisi esplorativa dei dati
"Esposizione di alcune tecniche per la investigazione dei dati"

Pierpaolo Napolitano *

Maggio 85

* Reparto Studi

I N D I C E

S O M M A R I O

Il seguente lavoro è rivolto alla esposizione di alcune tecniche di rappresentazione e manipolazione di dati numerici per l'investigazione preliminare, potremmo dire di tipo indiziario, delle informazioni in essi contenute.

Per dati intendiamo insiemi di numeri che sono misure risultanti da operazioni identiche effettuate su un fissato fenomeno. Essi possono essere non strutturati o già dotati di una data struttura fin dall'inizio dell'analisi. Nel primo caso l'insieme di numeri è una semplice collezione di misure, fra i cui elementi non riteniamo esistano discriminazioni o ordinamenti preliminari. Nel secondo caso si possono verificare strutture diverse. Qui ci occupiamo di quelle che:

a) ipotizzano una relazione funzionale fra coppie di misure e in cui i dati si presentano come insieme

di coppie di valori (il fattore e la risposta, rispettivamente).

b) associano a ciascuna misura livelli prefissati di una coppia di fattori, per cui i dati si presentano naturalmente sotto forma di matrice (tabelle a doppia entrata).

Non ci occuperemo di dati di tipo sequenziale, in cui lo ordine di apparizione delle misure, di tipo ad esempio temporale o spaziale, giuoca un ruolo essenziale nella interpretazione.

Brevemente il contenuto del lavoro è il seguente. Nel capitolo 1° diamo una descrizione generale dell'analisi esplorativa dei dati, del suo significato e delle sue motivazioni indicando, infine, brevemente, le idee che sono alla base della scelta delle tecniche in essa usate.

Nei capitoli 2° e 3° descriviamo, per il caso di un insieme di dati non strutturato, le tecniche per la loro rappresentazione sintetica e l'uso

di trasformazioni numeriche per semplificare le rappresentazioni e facilitare i loro confronti.

Nel capitolo 4° si descrive un metodo per interpolare con una retta nel caso di un insieme di dati costituito da coppie di valori.

Nel capitolo 5° si propone una tecnica per semplificare la lettura di una tabella a doppia entrata, che utilizza valori mediani invece che i classici valori medi.

Seguono un capitolo di conclusione e una nota su alcuni programmi meccanografici utilizzati in alcuni degli esempi riportati.

I N T R O D U Z I O N E

1.1 Le due fasi dell'analisi dei dati

Nella maggior parte delle definizioni esistenti un metodo statistico presuppone un modello probabilistico fissato. Ora, capita che esso è raramente riscontrabile nella pratica e si può talvolta arguire che non è fondamentale al suo sviluppo teorico [7].

Seguendo le riflessioni e le conseguenti proposte del Tukey [18], usiamo il nome di "Analisi dei dati" per definire l'attività dello statistico in quanto rivolta al momento applicativo, a sottolineare l'uso elastico dei concetti e degli strumenti della teoria della probabilità, quale in effetti si riscontra nella pratica.

Un'aspettativa importante che il Tukey ripone nella sua definizione è quella di sviluppare una

attitudine di maggiore aderenza ai dati rispetto ad un uso acritico dei modelli, sovente sovrasemplificati, che la statistica teorica propone.

Nell'investigazione dei dati, infatti, ci si trova quasi sempre a dover decidere il trattamento a cui sottoporli, sulla base di un certo tipo di lettura dei dati di partenza o di loro successive trasformazioni. Essa non si riduce a lineare e arido "Data processing", dove scelto il tipo di analisi si arriva meccanicamente ai risultati; richiede, al contrario, l'intervento critico e cosciente del ricercatore.

Secondo il Tukey l'Analisi dei dati si può schematicamente suddividere in due momenti: la fase esplorativa e quella di conferma.
E' la sottolineatura dell'importanza che riveste la prima delle due fasi che costituisce l'elemento di maggiore originalità nella sua proposta [19].

Rispetto alle tecniche classiche dell'inferenza statistica, che hanno un valore effettivo quando i dati

da analizzare rispettano assunzioni piuttosto stringenti, è apparso opportuno al Nostro valorizzare, esplicitare e sviluppare quelle tecniche di analisi, spesso implicitamente già presenti nella attività dello statistico, che ci avvicinano ai dati con il minimo di ipotesi precostituite.

Sono queste le tecniche, apparentemente molto elementari, che meglio ci guidano, nella fase esplorativa, ad una prima lettura dei dati, evidenziando gli aspetti fondamentali della distribuzione empirica, o di una sua parte, se i dati di partenza sono troppi, senza tener conto di problemi prematuri come, ad esempio, la scelta di modelli, l'individuazione di componenti deterministiche e/o stocastiche,etc.. .

La fase di conferma è quella in cui cerchiamo di dare alle evidenze rivelate delle valutazioni di carattere probabilistico, utilizzando le tecniche usuali dell'inferenza statistica, come, ad esempio, test di significatività, intervalli di confidenza... .

Una richiesta di conferma può,ad esempio, essere: con quale accuratezza i lineamenti apparenti delle accezioni numeriche rilevate nella prima fase sono credibili? Una analisi confermatoria può stabilire che tali lineamenti sono ben determinati o, al contrario, che sono così scarsamente definiti da poterli escludere (almeno come evidenze, anche se forse non come indizi).
Può anche accadere che i risultati siano scarsamente significativi e che sia, a questo punto, necessario raccogliere il massimo di informazione disponibile nei nostri dati rispetto al problema specifico in questione, raffinando le nostre tecniche, come anche informazioni contenute in altri aspetti dei dati o addirittura informazioni esterne a questi.

Di questa seconda fase, qui, non ci occuperemo.

1.2 Analisi esplorativa dei dati (EDA)

Quale che sia il gruppo di dati oggetto di indagine, una semplice collezione di numeri priva di struttura, un insieme di coppie di valori, un insieme di dati organizzati secondo una struttura matriciale, per ciascuno di questi contesti la questione di base è riuscire a rilevarne le caratteristiche essenziali nel modo più semplice per la nostra intelligenza, che, raggiunto un primo orientamento, può così esprimere valutazioni iniziali sulla cui base eventualmente proseguire l'analisi a un livello più sofisticato.

Secondo l'EDA conviene partire con semplici operazioni aritmetiche, di conteggio e di rappresentazione grafica, allo scopo di renderli più semplicemente e quindi più efficacemente trattabili.

Una rappresentazione adeguata e tecniche semplici di manipolazione possono infatti aiutarci a evi-

denziare strutture e ad ipotizzare possibili spiegazioni; possono farci modificare il modo di guardare ai dati, spingendoci verso rappresentazioni che ne semplifichino la forma, o a far ciò anche quando la rappresentazione iniziale è già soddisfacente, arricchendo la nostra comprensione del fenomeno; possono agevolarci nel trovare una forma quantitativa a delle intuizioni; possono infine, nei casi più fortunati, evidenziare comportamenti inaspettati.

Ad esempio, poter descrivere l'insieme dei numeri formato dalle altezze dei vulcani della terra, invece che con la distribuzione asimmetrica dei valori espressi in metri, con la distribuzione quasi perfettamente simmetrica della radice quadrata di tali valori, costituisce di per sè una semplificazione nella descrizione del fenomeno "altezza dei vulcani" e può inoltre facilitare il confronto con la distribuzione delle altezze di altro tipo di montagne o fra le altezze dei vulcani nei due emisferi.

Ci si potrebbe anche chiedere se la trasformazione usata riveste un significato puramente formale o sollecita ipotesi esplicative sull'origine della distribuzione.

Così anche la semplice osservazione che il logaritmo della pressione di vapor saturo è in relazione lineare con il reciproco cambiato di segno della temperatura assoluta comporta nell'analisi di dati empirici un guadagno effettivo rispetto alla affermazione, di pari contenuto formale, che la pressione cresce con la temperatura a tassi vieppiù crescenti. In questo caso una semplificazione nella relazione può facilitare una descrizione quantitativa e critica di dati inerenti al fenomeno.

L'EDA usa tecniche non inferenziali, si limita ad una, il più fedele possibile, rappresentazione dei dati raccolti, dando importanza alle ispezioni di tipo visivo. Il suo scopo è favorire un atteggiamento fortemente rilevato in senso empirico ed un uso flessibile delle tecniche, per un confronto il più serrato possi-

bile fra l'utente e i dati.

Stante che qualsiasi interrogativo nasce su un retroterra culturale e talvolta nell'ambito di qualche teoria, possiamo ritenere che la quantità di informazione contenuta nei dati (e che la realtà riversa in essi) possa non essere completamente discernibile nell'ambito teorico che pur ha generato il tipo di misure prescelto.
Ne consegue la necessità di favorire il ricorso al massimo di risorse conoscitive ed intuitive da parte di chi è coinvolto nella ricerca di un'interpretazione dei dati.

Intelligenza e comprensione di questi è, d'altra parte, essenziale quando i nostri interrogativi nascono da esigenze operative e ciò che chiediamo all'analisi è di fornirci spunti per decisioni efficaci.

1.3 Metodi resistenti e analisi dei residui

Accanto all'uso di trasformazioni numeriche e di metodi grafici particolari, ciò che caratterizza l'analisi esplorativa dei dati è l'utilizzazione di metodi cosiddetti "resistenti" per calcolare dai dati grezzi dei valori di sintesi e parametri di modelli descrittivi.

In generale qualsiasi raccolta di misure comporta il verificarsi di due tipi di errore: a) errori grossolani che possano superare l'ordine di grandezza del fenomeno in studio anche parecchie volte; b) errori di arrotondamento nel riportare le misure stesse.
Nella pratica il primo tipo di errore si verifica generalmente un numero limitato di volte e può essere attribuito a cause diverse (errata scelta dell'elemento o comportamento anomalo di questo, contaminazione fra popolazione con distribuzioni diverse, etc.) che solo un'analisi accurata successiva può discriminare.

Molto più frequente, quasi sempre presente, il secondo.

Diciamo che un metodo è resistente al primo tipo di errori se un piccolo sottoinsieme del gruppo dei dati non può avere un effetto sproporzionato nel valore calcolato. Un metodo è invece detto resistente al 2° tipo di errore se esso risponde in modo continuo a piccoli errori e se inoltre il valore calcolato non è determinato dall'arrotondamento o troncamento di una piccola frazione delle osservazioni.

Ordinariamente si temono gli effetti degli errori del primo tipo e qui ci poniamo il problema solo relativamente a questi.

Intenderemo perciò resistente un metodo quando esso mostra notevole insensibilità a comportamenti anomali localizzati nel gruppo di dati,tiene conto soprattutto del corpo principale delle misure e scarsa attenzione ai valori eccezionali. Come esempio di stima resisten-

te che utilizzeremo direttamente e indirettamente nei metodi di analisi proposti, c'è la mediana; essa chiaramente non risulta affetta in maniera significativa se inseriamo nel gruppo di dati anche valori molto elevati, al contrario di quello che accade per la media cui a una variazione non limitata di anche un solo dato, corrisponde una variazione non limitata del suo valore.

Per valutare quantitativamente il livello di resistenza di un metodo si utilizza il limite o punto di "rottura" (Hampel 1968) che misura la più elevata frazione possibile delle osservazioni che sostituita senza restrizione alcuna, comporta variazioni limitate nel valore calcolato.

Per la media tale valore è chiaramente zero.

Tenendo conto che la mediana o coincide con il valore centrale della distribuzione (se il numero n dei dati è dispari) o con le medie dei due valori centrali (se n è pari), per essa il limite di rottura è:

$$\frac{1}{2} - \frac{2 - d(n)}{2\,n}$$

dove n è il numero dei dati nella distribuzione e d(n) è la funzione di parità

$$d(n) = \begin{cases} 0 \text{ se } n \text{ è dispari} \\ 1 \text{ se } n \text{ è pari.} \end{cases}$$

Dalla stessa definizione risulta che essa è invece estremamente sensibile a errori di arrotondamento.

I residui sono ciò che rimane dopo aver sottratto una quantità di sintesi o un dato modello ai dati iniziali.

Simbolicamente

$$\text{residui} = \text{dati} - \text{modello}$$

Uno degli aspetti chiave del modo di procedere dell'EDA è l'analisi successiva dei residui, costituenti a loro volta un gruppo di dati.

Tale analisi ha l'indubbio vantaggio che l'uso di metodi resistenti porta a una netta separazione fra l'andamento predominante dei dati e comportamenti inusuali presenti in questi. Depurati i dati della struttura

predominante più significativa, attraverso la sottrazione del modello iniziale, i residui potrebbero contenere oltre che fluttuazioni derivanti dal caso, strutture più sottili non rilevabili immediatamente, o drastici scarti delle strutture regolari rilevate.

E, in effetti, solo il confronto con un andamento di base, consente alla nostra mente di rilevare comportamenti inusuali: normalmente non pensiamo che un livello zero sia inusuale, anche se certo il fatto che nessuno sia morto al mondo nelle ultime 24 ore o che ci sia un inverno senza neve sul Monte Rosa siano fatti assolutamente straordinari, come il fatto che il cane non abbai nella notte in alcune storie di Sherlock Holmes.

L'analisi dei residui può, quindi, essere utilizzata per migliorare in stadi il modello, includendo, ad esempio, variabili addizionali o un trattamento speciale per i casi anomali.

Quello che vogliamo infine sottolineare è che

mentre nell'analisi statistica classica la struttura dei dati è anticipata nella scelta di un modello e l'analisi dei residui, per fare un esempio, è utilizzata per verificare la correttezza delle assunzioni di tipo probabilistico sugli errori, nell'EDA si è alla ricerca di un modello che sveli ed esalti la struttura dei dati, e il ritorno ai dati sotto forma di residui è nell'ottica di una valutazione critica del modello usato e di suoi successivi adattamenti.

2. PRIME ISPEZIONI VISIVE E CONTROLLI NUMERICI

2.1 Diagramma "albero e foglia" (Stem and leaf)

Avere a che fare con i dati è una situazione che capita abbastanza di frequente non solo a chi si occupa di problemi di natura statistica o di tipo più specifico, nell'ambito di qualche scienza empirica, ma spesso anche nella formazione di giudizi legati alla quotidianità; anche in questo caso passiamo, per lo più implicitamente, attraverso una sommaria analisi di tipo quantitativo.

Le valutazioni dei prezzi di mercato delle autovetture usate di una certa cilindrata, le quantità di metri cubi di gasolio consumati giornalmente per riscaldamento di un edificio, le concentrazioni di biossido di zolfo rilevate da una certa stazione nel corso dell'anno, ognuno di questi fenomeni si presenta all'analisi come un insieme di dati.

Nonostante la grande varietà di situazioni in cui possiamo essere coinvolti è possibile dare degli strumenti di tipo generale, predisposti a guidarci efficacemente alla lettura e ad una prima interpretazione dei dati raccolti.

Queste tecniche si rivolgono verso insiemi di dati che si riferiscono tutti ad uno stesso fenomeno e che sono in numero consistente, riuscendo perciò difficilmente interpretabili ad un primo impatto.

Già nella fase di raccolta e trascrizione dei dati ci si possono porre problemi come la scelta della unità di misura e dell'arrotondamento o troncamento delle cifre significative. Le nostre scelte, che possono avere influenze decisive al proseguo dell'analisi, dipenderanno, da una parte, dal desiderio di renderli più trattabili, ad esempio più facilmente trascrivibili o meglio confrontabili fra loro, e, dall'altra parte, dalla nostra comprensione globale del fenomeno da investi-

gare. Rappresentare una popolazione in centinaia di migliaia di unità può essere una questione di ordine pratico e la relativa perdita di informazione inessenziale alla nostra indagine.

Per analizzare i prezzi di mercato delle auto usate potremo scegliere le decine o le centinaia di migliaia di lire; questo può dipendere da vari fattori, schematizzabili come sopra. In altre circostanze, può essere importante mantenere la massima precisione possibile.

Anche una struttura così povera come quella di una semplice collezione di numeri, può avere delle caratteristiche non facilmente discernibili.

Se il nostro gruppo di dati non è eccessivamente numeroso, diciamo non superiore a 300-400 numeri, può essere vantaggioso per costruirci una visione d'insieme, usare la rappresentazione grafica "albero e foglia" come proposta dal Tukey e che, opportunamente adattata può anche essere utilizzata per memorizzare i dati.

Questo, principalmente, ci facilita la lettura dei diversi aspetti della distribuzione con particolare riferimento a:

- l'ampiezza del campo di variabilità dei dati
- l'esistenza di valori intorno a cui i dati si concentrano
- la simmetria o asimmetria della distribuzione
- salti nell'intervallo dei valori ove non si osservano misurazioni
- valori che appaiono isolati rispetto al nucleo dei dati.

Prendiamo un esempio che ci guiderà nella descrizione. Ricaviamo da una pubblicazione che vengono offerti sul mercato, nel 1981-82, 18 amplificatori HI-FI di potenza RMS pari a 50 W e 17 modelli per quelli di potenza compresa fra 65 e 70 W. Dei 18 modelli da 50 W, 2 risultano sprovvisti di prezzo, i rimanenti 16 hanno i seguenti prezzi espressi in migliaia di lire: 290, 370, 470, 350, 485, 375, 540, 445, 280, 1970, 425, 380, 430, 320, 345, 350.

L'idea base del diagramma "albero e foglia", diversamente che per l'istogramma, che si riduce a delimitare delle aree, è di utilizzare nella rappresentazione le stesse cifre che compongono i dati.

Eliminando eventuali punti decimali, avremo n numeri di k cifre; di queste scegliamo le prime k' più significative, in numero tale che diano dimensioni accettabili al grafico. Queste, scritte le une sotto le altre in modo crescente a percorrere l'intero campo dei valori, costituiscono l'"albero" del nostro diagramma.

Nel nostro esempio, il valore minimo è 280, il massimo 1970; questo appare però isolato dal resto dei valori e conviene perciò rappresentarlo su una riga fuori diagramma preceduto da un simbolo speciale, ad esempio, HI (per high). Il valore massimo che rimane è 540. Siamo praticamente obbligati a scegliere come albero del nostro diagramma i valori delle centinaia di migliaia di lire, la prima delle cifre che compongono i nostri numeri. Una

scelta differente, diluendo eccessivamente il grafico, non ci faciliterebbe il modo apprezzabile la lettura dei dati.

L'"albero" ci apparirà in questo caso

2
3
4
5

HI 197

Accanto a ciascun elemento dell'"albero" e separato da esso con uno spazio, si porrà la prima delle cifre restanti, a costituire una foglia dell'"albero"; ciò viene fatto per ogni numero della collezione, ponendo l'uno accanto all'altro sulla stessa linea le cifre che hanno le prime k' uguali.

L'esempio completato apparirà

Unità = 10.000

2 89
3 1455778
4 23478
5 4

HI 197

In testa si è posta l'unità di misura che ci consente di leggere effettivamente i numeri; si fa osservare che a meno delle migliaia i singoli valori sono esattamente ricostruibili.

In generale, l'unità di misura potrà essere opportunamente modificata, per consentire rappresentazioni più concentrate o diluite secondo le necessità e tenendo conto che incrementarla di un fattore 10 comporta la perdita di una cifra significativa nella rappresentazione grafica.

Come variazione della rappresentazione grafica, a consentire una diluizione parziale delle cifre, senza modificare la quantità di informazione, si possono porre su ciascuna riga dell'albero non tutte le cifre da 0 a 9, ma rappresentarne 5 o 2 col conseguente moltiplicarsi degli elementi dell'albero in 2 o 5 righe differenti che potremo ad esempio rappresentare al modo seguente:

```
5*          5*
5.          5T
            5F
            5S
            5.
```

dove, nel primo caso, accanto all'* faremo seguire le cifre da 0 a 4 e al . da 5 a 9; nel secondo allo * i valori 0 e 1, a T 2 e 3, a F 4 e 5, a S 6 e 7 e a . 8 e 9. (1)

Utilizzando per il nostro esempio la prima delle due possibilità avremo

Unità = 10.000

2. 89
3* 24
3. 55778
4* 234
4. 78
5* 4

HI 197

La maggiore diluizione delle cifre ci consente di ricavare delle informazioni significative in modo probabilmente più immediato. Il grafico ci dice che i prezzi sono distribuiti in modo abbastanza simmetrico intorno ad un valore centrale di circa 37; il valore 197 è chiaramente un valore 'anomalo' che, se lo vogliamo, richiede un'analisi a sè.

(1) T per two e three, F per four e five, S per six e seven.

Presentiamo ora un esempio che utilizza i dati relativi alle temperature medie annuali espresse in ^{o}C , e alle precipitazioni annuali in mm per 39 stazioni termopluviometriche ed osservatori posti nel bacino del Tevere, ricavati dalla pubblicazione ISTAT (1981)

Annuario di statistiche metereologiche.

I rispettivi diagrammi "albero e foglia" sono i seguenti

Grafico 2.1 Temperature medie annuali-bacino del Tevere

```
LO   38                          UNIT =    0.1000
 2    9. 8
 3   10* 2
 5   10. 57
 9   11* 3344
10   11. 6
13   12* 224
17   12. 6899
(3)  13* 013
19   13. 66799
14   14* 0334
10   14. 5578
 6   15* 004
 3   15. 68
 1   16* 0
```

Grafico 2.2 Precipitazioni annuali-bacino del Tevere

```
 1    5  4
 2    6  2                    UNIT =   10.0000
 3    7  4
12    8  011467799
19    9  1234568
(6)  10  333359
14   11  688
11   12  0223
 7   13
 7   14  3
 6   15  244
 3   16  1

     HI  170, 176
```

- Dal grafico 2.1, osservato il caso "anomalo" del monte Terminillo con media annuale di 3,8 °C, possiamo dire che la temperatura sembra mostrare una distribuzione abbastanza uniforme fra gli 11 e 15 gradi.(1)
- Per le precipitazioni si può osservare una forte concentrazione intorno ai valori fra gli 800 e 1000 mm, con zone dai valori più elevati, intorno a 1150, 1500 e 1700 mm.

Ora, calcolati dalla definizione dell'indice di aridità del de Martonne [6] (uno dei vari indici climatici utilizzato soprattutto per cercare di spiegare le ripartizione della vegetazione), data da

$$i = \frac{P}{(T+10)} \quad ,$$

dove P è la piovosità annuale espressa in mm e T la temperatura media annuale in °C, i valori di tale indice per le 39 zone di cui disponiamo dei dati, dal corrispondente

(1) I grafici, elaborati con programma in Fortran, presentano sulla sinistra la colonna della profondità, che dà per ciascuna riga dell'albero il numero dei valori presenti sulla riga e su quelle più vicine al limite più prossimo del gruppo dei dati. Per la riga che contiene il valore centrale viene riportato fra parentesi il numero di elementi della riga stessa.

diagramma "albero e foglia" sembra di poter discriminare meglio le zone che già si potevano individuare nel grafico precedente

Grafico 2.3 Indici di aridità de Martonne - bacino del Tevere

```
                                        UNIT =    1.0000
  3      2. 579
  5      3* 23
 16      3. 55777889999
(7)      4* 0111224
 16      4. 7
 15      5* 0111
 11      5. 667
  8      6* 2
  7      6. 57
  5      7* 1222
  1      7. 5
```

2.2 "Valori lettera"

Se il diagramma albero e foglia ci consente di avere una visione d'insieme della distribuzione empirica dei dati, tuttavia esso non ci dà valutazioni strettamente quantitative, maneggiabili numericamente. Ora il nostro scopo è definire, stabilendo certe convenzioni, delle quantità sommarie, che chiamiamo valori-lettera, utili, nella maggior parte dei casi, a definire sinteticamente le caratteristiche fondamentali della struttura dei dati e a confrontare fra loro diversi gruppi di dati. Non ci si deve certo aspettare che queste quantità sommarie ci rivelino comportamenti inusuali o aspetti di dettaglio, per quanto questi possano essere importanti per una effettiva comprensione dei dati. Esse ci consentono di formarci delle idee che, per la gran parte dei dati che capita di trattare, ci rivelano gli aspetti essenziali della distribuzione.

Siano X_1, X_2, X_3,.,X_j,.,X_n i numeri che costituiscono il no-

stro gruppo di dati. Da essi formiamo il campione ordinato $X_{(1)},\ldots,X_{(i)},\ldots,X_{(n)}$ dove $X_{(i)}$ è la i-esima osservazione più piccola.

Per ciascun elemento definiamo suo rango superiore e inferiore la misura, intesa come conteggio di elementi, della distanza rispettivamente dal minore e dal maggiore degli elementi del campione ordinato, percorrendo i dati, una volta nel verso crescente e l'altra nel verso decrescente. Chiaramente, se l'elemento X_j diventa $X_{(i)}$ nel campione ordinato, il suo rango superiore sarà i e quello inferiore n+1-i.

Per ragioni di simmetria, ovvero per dare uguale importanza ai due estremi del campione ordinato, sembra conveniente definire per ciascun elemento un'altra quantità: la sua profondità; essa è il più piccolo fra i due ranghi associati al dato.

La profondità di un elemento ci dà la distanza di questo dal più vicino dei due estremi del campione. Utiliz-

zando questa nuova quantità possiamo ricavare una serie di quantità chiave nel gruppo dei dati ordinati. A profondità (n+1)/2 abbiamo la mediana, quel valore centrale della distribuzione che ha fra i dati tanti elementi ad essa superiori quanti inferiori. (1)

Indicheremo la mediana con la lettera M e quando scriveremo D(a) intenderemo la profondità dell'elemento a. A profondità 1 troviamo chiaramente i due valori estremi della distribuzione. In questo caso, come per ogni altro, escludendo la mediana, avremo a una certa profondità due valori posti simmetricamente rispetto ad M.

Un modo che riesce utile per caratterizzare il gruppo dei

---

(1) Nel caso di n pari calcoleremo la mediana come media aritmetica dei due valori vicini di rango superiore (n+1)/2 - 1/2, (n+1)/2 + 1/2.

dati è considerare successivamente i punti centrali delle code della distribuzione definite dai valori precedentemente calcolati. Partendo dalla mediana definiamo i quarti della distribuzione, che indicheremo con la lettera F, come i valori mediani delle code della distribuzione a destra e a sinistra di M. Essi avranno profondità pari a

$$D(F) = ([D(M)] + 1) / 2$$

dove con il simbolo [.] si intende la funzione parte intera.

Nel caso che D(F) sia una quantità semintera calcoleremo i quarti facendo la media aritmetica dei due valori vicini posti a profondità -1/2 e +1/2 rispetto a quella ricavata.

Si possono definire successivamente, con le convenzioni stabilite sopra, gli ottavi, i sedicesimi, etc.. della distribuzione partendo dalla loro profondità, definita dalla formula

$$(1) \qquad \frac{[\text{profondità quantità precedente}] + 1}{2}$$

Per indicare questi valori useremo come simboli delle apposite lettere. Chiamiamo, secondo le convenzioni del Tukey, la mediana valore M, i quarti valori F, gli ottavi valori E e successivamente avremo i valori D,C,B,A,Z,Y,X. [19]

Come esempio riportiamo alcuni di questi valori calcolati per le precipitazioni annuali in mm da 39 stazioni termo-pluviometriche ed osservatori posti nel bacino del Tevere, di cui abbiamo già visto il diagramma "albero e foglia".

	DEPTH	LOWER	UPPER	MID	SPREAD
N=	39				
M	20.0	1030.100		1030.100	
F	10.5	886.100	1212.100	1049.100	326.000
E	5.5	815.300	1531.550	1173.425	716.250
D	3.0	749.700	1613.600	1181.650	863.900
C	2.0	626.300	1702.200	1164.500	1075.900
	1	547.500	1766.500	1157.000	1219.000

Accanto a ciascun valore lettera (mediana, quarti, ottavi ed estremi) vengono riportate di seguito la profondità a cui è posizionata, il valore od i valori inferiore e superiore, il valore medio di questi e la loro differenza. Specie queste ultime due quantità ci forniscono uti-

li informazioni sulla struttura della distribuzione, l'andamento del valore della mediana e delle medie dei successivi valori lettera ci dà indicazioni sulle caratteristiche di simmetria o asimmetria della distribuzione; quello delle loro differenze sulla rapidità con cui cresce l'ampiezza del campo dei valori quando ci spostiamo verso le code della distribuzione, anche se non siamo in grado di controllare l'influenza che eventuali valori anomali hanno su tale ampiezza.

I valori lettera e le quantità che da queste potremo definire hanno delle interessanti proprietà statistiche in relazione alla presenza, nel gruppo dei dati, di grossolani errori di misurazione o di valori eccezionali, nel senso che hanno proprietà notevolmente differenti da quelle del nucleo dei dati.
Esse sono chiamate stime resistenti, ad indicare che sono insensibili a queste eventuali anomalie presenti nei dati. Al contrario, la media aritmetica e lo scarto qua-

dratico medio, le classiche misure di localizzazione e di dispersione, vengono perturbate dalla presenza anche di un singolo dato anomalo. La mediana si presenta come una stima di localizzazione estremamente resistente, in quanto, ad esempio nel caso in cui il 50% dei dati viene sostituito con valori eccezionalmente elevati, la sua variazione resta limitata.

Una stima di localizzazione meno resistente della mediana, ma che tiene parzialmente conto delle caratteristiche di dispersione della distribuzione è la cosiddetta Trimean

$$\frac{1}{4}\left(F_l + 2M + F_u\right)$$

dove M è la mediana, F_u e F_l rispettivamente il valore F superiore ed inferiore.

Nel nostro esempio M=1030 Trimean= 1039

Per misurare la dispersione possiamo semplicemente prendere la dispersione sui quarti (1)

$$d_F = F_u - F_l$$

Nel nostro esempio $d_F = 326$

---

(1) Usualmente indicati in statistica come differenza interquartile.

Tramite d_F possiamo ricavare un criterio empirico per la individuazione di valori anomali, che secondo i criteri della EDA, richiedono un'analisi particolareggiata per comprenderne l'origine. Il criterio proposto consiste nel definire dati anomali quei valori che assumono valori all'esterno dell'intervallo detto di troncamento $1{,}5\ d_F$

$$(F_1 - 1{,}5d_F,\ F_u + 1{,}5d_F) \qquad (2).$$

Nel nostro esempio abbiamo come valori anomali 1700 e 1760 delle stazioni rispettivamente di Atina e Casamari.

---

(2) Non c'è nessuna indicazione teorica per tale scelta, pare comunque che funzioni abbastanza bene. Come termine di confronto, nel caso di una distribuzione Gaussiana la percentuale della distribuzione che cade al di fuori dell'intervallo di troncamento $1{,}5\ d_F$ è 0,00698. Per campioni finiti di dimensione n, provenienti da una distribuzione Gaussiana, c'è un lavoro di Hoaglin, Iglewicz e Tukey (1981) [10] che verifica, attraverso simulazioni, la approssimazione di questa percentuale con la formula

$$0{,}00698 + \frac{0{,}4}{n}.$$

Oltre a tale intervallo il Tukey propone di considerare per individuare i dati "molto anomali" anche l'intervallo $(F_1 - 3d_F,\ F_u + 3d_F)$ detto di troncamento $3d_F$

2.3 "Valori lettera": alcune osservazioni

La definizione di mediana come la quantità che divide in due metà la distribuzione dei dati è certo familiare; i quarti, tranne che per il modo specifico di calcolarli, coincidono praticamente con i quartili della distribuzione. Gli altri valori lettera sono meno familiari nella pratica dell'analisi statistica. E' ragionevole chiedersi una giustificazione del modo in cui vengono calcolati; ad esempio domandarsi perchè si utilizza la formula (1) del par.2.2 e non direttamente la profondità n/4 per i quarti o n/8 per gli ottavi.

Una prima risposta è che la (1) semplifica il procedimento di calcolo, nel senso che ciascun valore lettera o corrisponde direttamente a un valore dell'insieme dei dati o alla media di due valori adiacenti. Questa, però, non è la spiegazione completa; in quanto, per esempio, per n=11 $n/4 = 2\frac{3}{4}$ e si sarebbe potuto arrotondare il

valore della profondità a 3, mentre è posta a 3 $\frac{1}{2}$. La parte rimanente della giustificazione è che si è deciso di fissare la frazione dei dati da lasciare sulla sinistra di ciascun valore lettera inferiore (e sulla destra per ciascun valore lettera superiore) alle potenze successive di 1/2. Per ricavare la relazione fra valore di profondità e corrispondente frazione di dati, diciamo, alla sinistra nella distribuzione abbiamo bisogno di un risultato preliminare e alcune considerazioni.

Supponiamo di avere una popolazione P con funzione di ripartizione F(X) derivabile. Siano $X_1 .. X_n$ le variabili casuali relative a un campione di ampiezza n estratto da P. Le osservazioni ordinati in senso crescente, espresse come $X_{(1)}, X_{(2)}, \ldots, X_{(i)}, \ldots, X_{(n)}$, sono chiamate le statistiche d'ordine del campione dato, e $X_{(i)}$ è detta statistica d'ordine di rango i.

Cerchiamo ora l'area sottesa dalla distribuzione della popolazione P che il valore mediano della variabile casuale $X_{(i)}$

lascia alla sua sinistra.

La probabilità che un valore compreso tra x e x+dx sia la statistica d'ordine di rango i in campione di dimensione n è la probabilità calcolata da una multinomiale relativa ai 3 possibili eventi $(-\infty, x)$ $(x,x+dx)$ $(x+dx,+\infty)$ che si verificano rispettivamente (i-1), 1, (n-i) volte su un totale di n prove.

Essa è pari

$$\frac{n!}{(i-1)!\,(n-i)!}\,[F(x)]^{i-1}\,[1-F(x)]^{n-i}\,f(x)dx = Pr^{i}(x)$$

in quanto F(x), 1-F(x), f(x)dx sono le probabilità rispettivamente degli eventi elementari

$$(-\infty, x)\ (x+dx, +\infty)\ (x,x+dx).$$

Il valore mediano di $X_{(i)}$ è il valore $x_{i,.5}$ tale che

$$\int_{-\infty}^{x_{i,.5}} Pr^{i}(x)dx = 1/2$$

osservando che f(x)dx = dF, si vede che basta risolvere questa equazione per la sola distribuzione uniforme per trova-

re risposta alla nostra richiesta.

Si ha: -

$$\frac{n!}{(i-1)!(n-i)!} \int_0^{F_{i,.5}} F^{i-1} (1-F)^{n-i} \, dF = 1/2$$

dove $F_{i,.5} = F(x_{i,.5})$. Interpolando le tavole della funzione beta incompleta si può arrivare ad una approssimazione ragionevole per $F_{i,.5}$.

Ora appunto il valore $F_{i,.5}$ misura, qualunque sia F(X), la area della distribuzione di P che il valore mediano della statistica d'ordine di rango i di un campione di ampiezza n lascia alla sua sinistra.

Se ora prendiamo in considerazione le famiglia di definizioni proposte da Blom [2] per descrivere le frazioni dei valori della popolazione alla sinistra dello i esimo valore più piccolo in un campione di dimensione n

$$(\text{frazione} \leq x_{(i)}) = \frac{i-\alpha}{n+1-2\alpha}$$

dove α è il parametro delle famiglie, vediamo [10] che per $\alpha = 1/3$, con cui abbiamo

$$(\text{frazione} \leq x_{(i)}) = \frac{i - \frac{1}{3}}{n + \frac{1}{3}} ,$$

tale valore è con buona approssimazione molto vicino al valore $F_{i,.5}$. Cioè che, qualunque sia la distribuzione continua della popolazione P, la frazione di dati lasciati a sinistra del valore mediano della statistica d'ordine di rango i in un campione di ampiezza n è ben approssimata dalla definizione di Blom, con

α = 1/3, della frazione dei dati che, sempre nella popolazione, restano alla sinistra della i-esimo valore più piccolo in un campione di dimensione n.

Invertendo la relazione di Blom e fissando i valori per la frazione dei dati come potenze intere di 1/2 ricaviamo le corrispondenti profondità

$$d_n = \left(n + \frac{1}{3}\right) X \left(\frac{1}{2}\right)^n + \frac{1}{3}$$

che potremo associare a degli ideali valori lettera.

Attraverso una analisi dettagliata [10], emerge che le

differenze fra le profondità di valori lettera ideali e le profondità definite dalla (1) sono non rilevanti, e precisamente:

1) la profondità definita dalla (1) è sempre più interna ai dati di quella ideale
2) la maggior parte delle volte la differenza è minore di 1/2 (sempre per i valori F, 3/4 delle volte per i valori E e D, e circa 2/3 per il seguito)
3) la differenza fra le due profondità non è mai superiore all'unità.

La semplicità di calcolo va a tutto vantaggio della (1)

2.4 "Diagramma a scatola" (Boxplot)

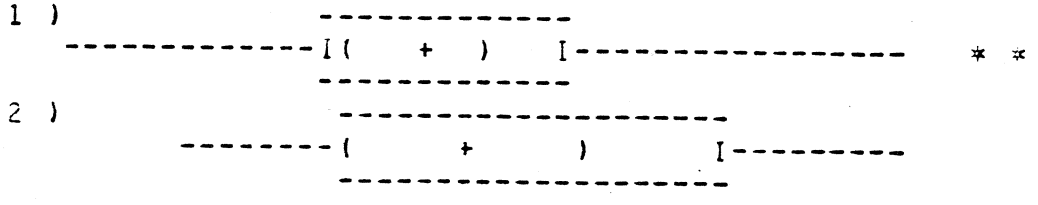
Utilizzando alcuni dei valori lettera precedentemente definiti, la mediana e i due quarti, si può ricavare una semplice rappresentazione grafica che ci consente di evidenziare, senza disperderci nella analisi dei singoli valori, le caratteristiche di localizzazione, di dispersione, di simmetria, di lunghezza delle code e la eventuale presenza di dati anomali. Questo grafico, chiamato diagramma a scatola, viene costruito nel modo seguente.

Su una linea orizzontale, scelta una scala adeguata a rappresentare l'intero gruppo dei dati, si indicherà con un segno + la posizione della mediana.

Ai lati vengono segnate con due segmenti verticali le posizioni dei due quarti superiore ed inferiore. Per evidenziare che in questa zona del grafico ricade la metà dei valori della distribuzione, uniamo gli estremi corrispondenti dei due segmenti a formare una scatola rettangolare.

Partendo dai due segmenti tracciamo verso l'esterno due linee fino ai punti che, sulla scala, corrispondono ai valori della distribuzione immediatamente superiori e inferiori rispettivamente a $F_l - 1{,}5\ d_F$, $F_u + 1{,}5 d_F$. Questi punti sono chiamati valore adiacente inferiore e superiore e ci danno una idea della lunghezza delle code della distribuzione nei quarti esterni restanti. Indicheremo con un * i valori esterni all'intervallo di troncamento $1{,}5\ d_F$; se fra questi ce ne sono alcuni lontani dal rispettivo quarto per una quantità superiore a $3d_F$ li evidenzieremo con il simbolo 0; questi ultimi li chiameremo "valori molto anomali". Riportiamo come esempio il Diagramma a scatola ai dati sulle precipitazioni dell'anno 1981 del bacino del Tevere e del bacino medio sinistro del Po, sulla medesima scala, uno sotto l'alto, per consentire il loro confronto.

Grafico 2.4 Boxplots della precipitazione del bacino del Tevere e del medio Po sinistro

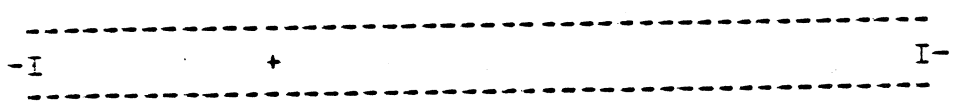


Vogliamo osservare che un uso meccanico di questo metodo può portarci a valutazioni fuorvianti e a perdere il "messaggio" contenuto nei dati. Riprendiamo dal testo del Tukey l'esempio che riporta i risultati di una serie di sperimenti effettuati da Lord Rayleigh sulla densità del 'nitrogeno' ricavato da varie fonti. Queste misurazioni evidenziarono una netta discrepanza fra la densità del gas nitrogeno prodotto dall'aria con quella ricavata per il gas prodotto dalla decomposizione di alcune sostanze chimiche, portando alla scoperta dell'argon. I risultati di Rayleigh furono i seguenti:

Pesi ricavati da Rayleigh per un volume standard di 'nitrogeno'

Data	Origine	Agente purificante	Peso
29/11/93	NO	Ferro rovente	2,30143
5/12/93	"	"	2,29816
6/12/93	"	"	2,30182
8/12/93	"	"	2,29890
12/12/93	Aria	"	2,31017
14/12/93	"	"	2,30986
19/12/93	"	"	2,31010
22/12/93	"	"	2,31001
26/12/93	N_2O	"	2,29889
28/12/93	N_2O	"	2,29940
9/1/94	NH_4NO	"	2,29849
13/1/94	NH_4NO	"	2,29889
27/1/94	Aria	Idrato di Ferro	2,31024
30/1/94	"	"	2,31030
1/2/94	"	"	2,31028

Il diagramma a scatola dei pesi del nitrogeno è il seguente



Il grafico è chiaramente fuorviante come una semplice analisi con il diagramma albero e foglia chiarisce.

```
                                  UNIT =    0.0010
  6 229.  888889
(2) 230*  11
  7 230T
  7 230F
  7 230S
  7 230.  9
  6 231*  000000
```

2.5 Alcuni esempi

Dal [10] riportiamo l'esempio relativo ad un campione di 994 famiglie cui è stato chiesto il reddito annuale in $.
Dai dati è stata ricavata la seguente tabella riportante tutti i possibili valori lettera. La tabella riporta anche, come esempio di possibile utilizzazione dei valori lettera, la media aritmetica dei loro valori superiore ed inferiore. Essi sono posti al di sotto del valore della mediana.

n=994

		L		U
M	497,5		3480	
F	249	2412	3678	4944
E	125	1788	4115	6433
D	63	1517	4400	7284
C	32	1248	4799	8350
B	16,5	963	4978	8994
A	8,5	727	5241	9754
Z	4,5	579	5394	10210
Y	2,5	345	5510	10675
	1	114	5494	10874

Le medie dei valori lettera, che chiameremo midF,midE .. mostrano in questo esempio un andamento chiaramente crescente. Il carattere sistematico di questo ci suggerisce

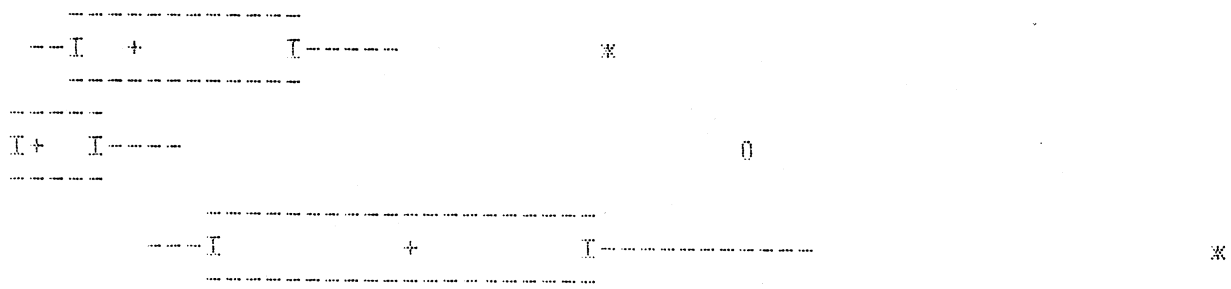
una asimmetria destra della distribuzione (valori concentrati in un piccolo intervallo sulla sinistra della distribuzione e maggiormente diluiti sulla destra); ci fa inoltre pensare che questo non sia addebitabile a qualche valore eccezionale posto alla sinistra della distribuzione come avremmo potuto pensare nel caso che solo i valori estremi avessero presentato andamento crescente.

Un'utilizzazione abbastanza naturale del Boxplot è il confronto fra differenti gruppi di dati. Esso ci consente infatti di percepire immediatamente la posizione relativa delle mediane e dei valori adiacenti e di avanzare su questo, ipotesi sulle distribuzioni che potremo successivamente valutare nella fase di conferma (analisi della varianza a una via, per esempio).

Costruiamo, sempre su suggerimento del testo citato i Boxplots relativi alle maggiori 10 città italiane, francesi ed indiane. I dati sono i seguenti (popolazione del 1967 espressa in centinaia di migliaia di individui)

Italia		Francia		India	
Roma	23,59	Parigi	28,11	Bombay	45,37
Milano	15,80	Marsiglia	7,83	Calcutta	30,03
Napoli	11,82	Lione	5,35	Delhi	22,98
Torino	11,14	Tolosa	3,30	Hyderabad	20,62
Genova	7,84	Nizza	2,98	Madras	17,25
Palermo	5,90	Bordeaux	2,54	Howrah	16,11
Firenze	4,54	Nantes	2,46	Ahmedabad	11,49
Bologna	4,44	Strasburgo	2,35	Kanpur	9,47
Catania	3,61	St.Etienne	2,03	Bangalore	0,07
Venezia	3,36	Lilla	1,99	Poona	7,21

Il grafico dei Boxplots è il seguente:



3. TRASFORMAZIONE DEI DATI

3.1 Premessa

I gruppi di dati, di cui generalmente ci si occupa in statistica, si presentano in tre ampie classi:

- ammontare di quantità e conteggi
- saldi, differenza fra quantità
- rapporti e percentuali

A questi possiamo aggiungere, per completezza, le misure che appaiono sotto forma di valori simbolici ordinati in una scala. Il primo tipo di dati, quello di cui ci occupiamo, non possono chiaramente assumere valori negativi e, a priori, non hanno un limite superiore e presentano spesso, per questo motivo, una asimmetria destra nella distribuzione. Rientrano in questo tipo di dati misure di altezze, di superficie, di distanze come anche numero dei morti, di incidenti, etc.. .

Se pensiamo il secondo tipo di dati come differenze fra quantità del primo tipo, potremo applicare indirettamente anche ad essi i procedimenti di cui si parlerà.

La prima e più immediata ragione per sottoporre a trasformazione numerica i dati è che se il campo dei loro valori è molto ampio può risultare difficile una loro rappresentazione grafica, ad esempio col diagramma "albero e foglia". Se il rapporto fra il valore massimo e minimo è molto elevato, diciamo 100 e più, è conveniente, sia per la memorizzazione che per una rappresentazione, utilizzare i logaritmi in base 10 o la radice quadrata dei dati di partenza.

Ricaviamo dal Rapporto sull'energia 1982, predisposto dall'ENI, le capacità di lavorazione di decreto, misurate in migliaia di tonnellate annue, delle raffinerie italiane; esse danno luogo al seguente diagramma "albero e foglia":

Unità 0.1

```
 0 122335
 1 013666
 2 6
 3 9999
 4 0366
 5 00122369
 6 5
 7 8
 8 0
 9
10 0
11 0
12
13 1
14 0
15
16
17
18 0
19
20 34
```

Utilizzando i logaritmi in base 10 delle capacità, otteniamo la seguente rappresentazione, con un guadagno nella visione complessiva e per la possibilità di confronti fra le varie capacità:

$\log_{10}$capacità

Unità 0.1

```
-1  0
-0. 5577
-0  03
+0  012224
+0. 666666777777777789
+1  0011233
+1.
```

Altre volte una trasformazione può avere un significato direttamente legato alla natura del fenomeno investigato, facilitandoci nella sua lettura.

In un lavoro di C.L.Hall (1934) sull'apprendimento dei ratti a percorrere un labirinto vengono riportati per 4 diversi animali, nella seconda giornata di esperimenti, i tempi di percorrenza per andata e ritorno. Rappresentandoli in un diagramma "albero e foglia", arrotondando sull'ultima cifra significativa, essi danno:

	1° Ratto	2° Ratto	3° Ratto	4°Ratto	Unità 1 s
0			43		
0.	87	9	66	56678	
1	34	23	1		
1.		9			
2		2			
2.	6				

Dal grafico sembrerebbe che il 4° Ratto abbia valori molto concentrati e che, al contrario, il 1° e il 2° abbiano tempi con notevole dispersione.

Riportando i dati invece che in s in 1000/s, misurando la velocità di percorrenza invece del tempo, il grafico

diventa:

Unità 1000/s

0	4	4		
0.	87	588	9	
1	34	1	5	42
1.			7	88
2				2
2.			6	
3			4	
3.				

Ora il 1°, 2°, 4° ratto hanno dispersioni simili; il 3° ha una dispersione maggiore dovuta, sembrerebbe, ai primi due viaggi molto rapidi.

3.2 Trasformazioni di elevamento a potenza

Per trasformare i dati ci limiteremo alla seguente famiglia di trasformazioni.

$$(1) \qquad \emptyset_p(x) = \begin{cases} \dfrac{x^p - 1}{p} & \text{per } p \neq 0 \\ \ln(x) & \text{per } p = 0 \end{cases}$$

dove x è la variabile positiva da trasformare e p, indice della trasformazione, un qualsiasi numero reale.
La $\emptyset_p(x)$, funzione reale della variabile reale positiva x è chiaramente continua per qualsiasi valore di p

Considerando l'indice p della trasformazione come a sua volta una variabile reale, la funzione $y(p,x)=\emptyset_p(x)$ risulta continua per qualsiasi valore di p. Per $p\neq 0$ è immediato (segue dalla continuità rispetto a p di x^p e di p e del loro quoziente per $p\neq 0$). Per p=0 basta osservare che

$$\lim_{p\to 0} y(p,x) = \ln(x) = y(0,x)$$

La proprietà della continuità si estende alle derivate di ordine qualsiasi rispetto ad x e p.

La famiglia di funzioni data dalla (1) si presta assai bene allo studio comparativo dei risultati derivanti dalla loro applicazione alla variabile x. Per essa si verificano le seguenti proprietà:

- qualunque sia p, $\emptyset_p(x)$ ha un andamento monotono crescente ($\emptyset'_p(x)=x^{p-1} > 0$ se $x > 0$) e conserva quindi l'ordinamento originario dei dati
- ciascuna di esse conserva la propria convessità; per $p > 1$ $\emptyset_p''(x) > 0$, per $p < 1$ $\emptyset_p''(x) < 0$.
- ciascuna di esse passa per il punto (1,0) ed ha ivi tangente con coefficiente $m=1$ ($\emptyset_p(1)=0$, $\emptyset_p'(1)=1 \forall p$).

Per x positivo e limitandoci al caso $p > 3$ verifichiamo che la differenza della variazione rispetto ad 1 della x e della variazione rispetto a 0 della trasformata può essere resa $< \varepsilon \forall \varepsilon > 0$ in quanto

$$\left| X - \frac{(1+X)^P - 1}{P} \right| = \left| \frac{(P-1)}{2} X^2 + \frac{R_2(x)}{P} \right| \leqslant$$

$$\left| \frac{(P-1)}{2} X^2 \right| + \left| \frac{R_2(X)}{P} \right| \leqslant \left| \frac{(P-1)}{2} X^2 \right| + \left| \frac{1}{P} \binom{P}{3} X^3 \right|$$

dove $R_2(x)$ è il resto dello sviluppo in serie di potenze di $(1+x)^p$ del secondo ordine.

Nell'approssimazione del 1° ordine nel punto (1,0) la scala dei valori trasformati resta inalterata. Per osservare variazioni nel grado di trasformazione dei dati occorre passare ad approssimazioni del 2° ordine. In questo caso ad una variazione simmetrica a destra e a sinistra del punto x=1 pari a Δx e $-\Delta x$ si osservano variazioni nelle trasformate pari rispettivamente a

(2) $$\Delta y = \Delta x + \frac{p-1}{2}(\Delta x)^2$$

$$\Delta y = -\Delta x + \frac{p-1}{2}(\Delta x)^2 \ ;$$

nel caso $p > 1$ la variazione per i valori trasformati maggiori di 0 supera la variazione determinata per la x; il

contrario si verifica per i valori inferiori a 0. Il comportamento opposto si verifica per $p<1$.

E' interessante studiare il caso di una trasformazione generata da una circonferenza passante per il punto (1,0) e con centro sulla retta $y = -x+1$, perpendicolare alla tangente comune alle curve della famiglia $\emptyset_p(x)$ nel punto suddetto.
In questo caso la trasformazione è esprimibile, nella approssimazione del 2° ordine, al modo seguente in un intorno di $x = 1$:

$$y(x) = (x-1) - \frac{\sqrt{2}\,(x-1)^2}{r} = \Delta x + \frac{\sqrt{2}(\Delta x)^2}{r}$$

dove $r = \sqrt{2}\,|x_0 - 1|$ è il raggio della circonferenza con centro di ascissa x_0. Quindi l'effetto di 2° ordine della trasformazione è regolata dalla curvatura $\frac{1}{r}$ della circonferenza tramite il fattore $\sqrt{2}$, (in quanto esprimiamo le variazioni del 1° ordine della lunghezza della curva tramite le variazioni nella variabile x).

Questo ci spinge a riscrivere la (2) nel caso

$p>1$ così:

$$(3)\quad \Delta y = \Delta x + \frac{\sqrt{2}(p-1)(\Delta x)^2}{2^{3/2}} = \Delta x + \sqrt{2}\,K_p(1)(\Delta x)^2$$

dove $K_p(1)$ è il valore della curvatura di $\emptyset_p(x)$ nel punto $x=1$. Si ha infatti:

$$(4)\quad K_p(x) = \frac{|P-1|\,x^{P-2}}{(1+x^{2p-2})^{3/2}}$$

e per $x = 1$

$$K_p(1) = \frac{|P-1|}{2^{3/2}}$$

Ne risulta, per inciso, che la curvatura tende a 0 sia al crescere di x sia quando x tende a 0.

Ciò che si vuole soprattutto osservare è il comportamento opposto per i casi di $p>1$ e di $p<1$, e come al crescere del valore di p l'effetto di contrazione o dilatazione della scala nel punto (1,0) tende ad aumentare in modo proporzionale alla curvatura, che è a sua volta proporzionale a $|1-p|$.

3.3 Scopi specifici della trasformazione

Come scopi specifici della trasformazione facciamo per ora riferimento ai due seguenti, in relazione a più insiemi di dati:

- dare forma simmetrica alle singole distribuzioni
- raggiungere un'uguale dispersione fra i diversi gruppi di dati.

I due obiettivi sono chiaramente legati al proseguo della analisi nella fase confermativa; le suddette caratteristiche facilitano, infatti, e rendono più efficace l'applicazione di alcuni test statistici. Già nella fase esplorativa, tuttavia, queste consentono confronti di tipo qualitativo più immediati e significativi.

Questi obiettivi non sono in linea teorica compatibili, utilizzando la famiglia di trasformazioni prima definita; può, d'altra parte, accadere che una singola tra-

sformazione ci porti, in buona approssimazione, ad entrambi i risultati.

Diamo ora due criteri empirici per valutare il coefficiente p della trasformazione $\emptyset_p(x)$ che porta i dati ad assumere, rispettivamente, simmetria nella loro distribuzione e uguale dispersione nei diversi gruppi.

Criterio per la simmetria

Costruiamo il grafico costituito dai punti aventi come ascissa il valore

$$\frac{(X_U - M)^2 + (M - X_L)^2}{4M}$$

e come ordinata il valore

$$\frac{X_U + X_L}{2} - M$$

dove M indica la mediana e X_U e X_L il valore lettera superiore ed inferiore di una data profondità.

Se questi punti, ottenuti in corrispondenza ai differenti valori lettera, si dispongono secondo una retta di penden-

za b il valore p del coefficiente della trasformazione sarà posto uguale a 1-b.

Una giustificazione per questa regola è il seguente ragionamento:

si cerca il valore di p tale che

$$X_U^p - M^p = M^p - X_L^p$$

ovvero

$$\frac{X_U^p + X_L^p}{2} = M^p$$

qualunque sia il valore lettera rappresentato dalla X.

Se sostituiamo in una di queste due espressioni, l'espansione in serie di Taylor fino al 2° ordine di X_U^p e di X_L^p intorno al valore **M**, che risultano, per $p \neq 0$ e a meno del segno per $p < 0$, rispettivamente

$$X_U^p \simeq M^p + pM^{p-1}(X_U - M) + \frac{p(p-1)}{2} M^{p-2}(X_U - M)^2$$

$$X_L^p \simeq M^p + pM^{p-1}(X_L - M) + \frac{p(p-1)}{2} M^{p-2}(X_L - M)^2,$$

otteniamo dopo semplici trasformazioni algebriche

$$\frac{X_U + X_L}{2} - M \simeq (1-P) \frac{(X_U - M)^2 + (M - X_L)^2}{4M} .$$

Perciò, possiamo dire che se il valore cercato di p esiste, esso è ricavabile con il criterio proposto.

La quantità che al secondo membro moltiplica (1-p) è una quantità crescente all'aumentare della profondità dei valori lettera. Il primo membro, in un gruppo di dati già simmetrico dovrebbe essere costantemente nullo. Nel caso che mostri un andamento crescente o, di contro, decrescente, nell'ipotesi che l'approssimazione al 2° ordine dello sviluppo in serie di Taylor sia buona, il valore approssimato di p che si ricava dalla pendenza b della retta sarà altrettanto buono per rendere simmetrica la distribuzione di partenza.

Come criterio alternativo al calcolo della pendenza, si può direttamente calcolare il rapporto di ciascun valore in ordinata col corrispondente valore in ascis-

sa; se questi valori sono fra loro approssimativamente uguali, si potrà ritenere che il valore di p che ne ricaviamo sarà efficace.

A scopo esemplificativo, ritorniamo al primo esempio del paragrafo 2.4, relativo ai redditi delle famiglie. Avevamo già notato l'andamento crescente delle medie dei valori lettera, arguendone una asimmetria destra nella distribuzione. Per valutare la possibilità di renderla simmetrica calcoliamo per ciascun valore lettera la stima relativa del valore di p. Otteniamo i seguenti valori:

Valori lettera	Stima di p
F	0.16
E	0.24
D	0.30
C	0.36
B	0.43
A	0.48
Z	0.50
Y	0.54

A crescenti valori lettera corrispondono stime di p crescenti e, quindi, trasformazioni meno potenti; evidentemente le caratteristiche di asimmetria non sono distribuite unifor-

mente e tendono ad essere maggiori per i valori lettera più interni. In pratica ciò rende impossibile trovare una trasformazione, fra quelle che abbiamo ipotizzato, che sia in grado di darci simmetria lungo tutta la distribuzione.

Criterio per l'uguaglianza in dispersione

Date n collezioni di dati, indichiamo con M_i e d_{Fi} rispettivamente la mediana e la dispersione sui quarti dell'i-esimo gruppo. Costruiamo il grafico riportante, per i diversi gruppi, in ascissa il $\log_{10} M_i$ e in ordinata il $\log_{10} d_{Fi}$. Se il grafico rappresenta un andamento approssimativamente lineare con pendenza pari a b, p=1-b è l'esponente della trasformazione cercata.

Alla base del detto criterio vi sono, naturalmente, ipotesi sulla validità di certe approssimazioni, come verrà chiarito qui di seguito.

Sia x una variabile casuale con mediana v, con

quarto inferiore pari a F_l, quarto superiore pari a F_u e dispersione sui quarti pari a $T_x(v)$. (Stiamo in realtà trattando con un insieme di variabili casuali indiciate da v).

Cerchiamo una trasformazione $y=\phi(x)$ tale che la dispersione sui quarti $T_y(\phi(v))$ sia costante. Supponiamo che la funzione ϕ ammetta derivate continue fino al 3° ordine almeno.

E' per ipotesi $F_u - F_l = T_x(v)$. Per ciascuna v esiste un $\lambda = \lambda(v)$ tale che $F_u - v = \lambda T_x(v)$ e

$$v - F_l = (1-\lambda)T_x(v)$$

Riesce $0 \leqslant \lambda \leqslant 1$.

Per ottenere un'espressione per $T_y = \phi(F_u) - \phi(F_l)$ sviluppiamo in serie di Taylor ciascuno dei due termini che lo compongono intorno al valore v, troncando lo sviluppo al 2° ordine; otteniamo

$$\phi(F_u) = \phi(v + \lambda(v)T_x(v)) = \phi(v) + \lambda(v)T_x(v)\phi'(v) + \frac{\lambda^2(v)T_x^2(v)}{2!}\phi''(v)$$

$$\phi(F_l) = \phi(v-(1-\lambda(v))T_x(v)) = \phi(v) - (1-\lambda(v))T_x(v)\,\phi'(v) + \frac{(1-\lambda(v))^2T_x^2(v)}{2!}\phi''(v).$$

Per T_y otteniamo, perciò, la seguente approssimazione

$$T_y = \emptyset(F_u)-\emptyset(F_l) = T_x(v)\emptyset'(v) + \frac{(2\lambda(v)-1)T_x^2(v)}{2!}\emptyset''(v) .$$

Consideriamo il termine di 2° ordine. Esso è composto dal fattore $(2\lambda(v) - 1)$. Osserviamo che 1) questo al massimo può assumere il valore 1, 2) se x fosse una variabile casuale simmetrica sarebbe nullo e 3) più realisticamente, se il quarto superiore fosse 2 volte più lontano dalla mediana del quarto inferiore varrebbe $\frac{1}{3}$; il valore $\emptyset''(v)$ misura la concavità di $\emptyset$ intorno al valore della mediana v; se questa non è eccessivamente vicina a 0 e $\emptyset$ è approssimativamente lineare intorno a v, allora si può valutarla come una quantità piccola. Il termine $T_x(v)$ possiamo ritenerlo dell'ordine di grandezza di uno scarto quadratico medio, misura che riteniamo finita. Queste osservazioni ci portano a ritenere trascurabile il termine di 2° ordine rispetto a quello di 1°.

In questo caso abbiamo dunque $T_y = T_x(v)\emptyset'(v)$.

Imponendo la costanza a T_y abbiamo $T_x(v)\ \emptyset'(v)=c$ che dalla conoscenza di $T_x(v)$ ci consente di ricavare la trasformazione cercata: $\emptyset(x) = c \int T_x^{-1}(x)\,dx$.

Si è voluto riportare il precedente argomento, ripreso dal testo [10], non certo per il suo rigore, quanto per consentire una valutazione sul carattere approssimativo del criterio proposto.

C'è da aggiungere, comunque, che, a detta degli autori del testo citato, tale criterio rappresenta un raffinamento nella letteratura sulla questione, anche se certo non una parola definitiva.

Specializzando il risultato di cui sopra all'ipotesi in cui $T_x(v) = kv^b$, le possibili trasformazioni che otteniamo sono, a meno di una costante moltiplicativa e di una costante additiva, quelle della famiglia delle trasformazioni di potenza. Abbiamo infatti

$$\emptyset(x) = \frac{1}{K}\int x^{-b}\,dx = \begin{cases} c_1 X^{1-b} + c2 & \text{per } b \neq 1 \\ c_3 \ln X + c4 & \text{per } b = 1 \end{cases}$$

dove c_1, c_2, c_3, c_4 sono costanti arbitrarie.

Osservando infine, che $\ln [T_x(v)] = \ln K v^b = \ln K + b \ln v$ troviamo la giustificazione alla regola empirica fornita.

A mò di esempio riportiamo i dati da Tukey [19], a loro volta ripresi da un articolo del 1950 su "Psychometrika" di Bruner, Postman e Mosteller, relativi ad un semplice esperimento, nel quale veniva presentato a diversi soggetti un disegno, la scala di Schroeder, che poteva essere facilmente visto in due differenti prospettive.
Essi avevano ricevuto differenti istruzioni sul modo di mutare la loro prospettiva, ed il numero di cambiamenti di prospettiva veniva contato per ciascuno dei 10 minuti successivi. Tralasciando i primi due minuti, per evitare eventuali effetti di 'partenza', vi sono 19 gruppi di dati, uno per ciascun soggetto sotto differenti istruzioni, ciascuno composto da 8 conteggi.

Nella tabella seguente vengono riportati per ciascuno dei 19 soggetti il valore M della mediana del gruppo

di 8 dati; il $\log_{10}$ di M, il valore dei quarti inferiore e superiore, la differenza sui quarti e il $\log_{10}$ di quest'ultima.

M	$\log_{10} M$	F_l	F_u	d_F	$\log_{10} d_F$
2	103	11	14	3	48
16.5	222	16	18	2	30
22	134	21	23	2	30
22	134	20	22	2	30
28	145	22	48	26	142
29.5	147	24	32	8	90
33.5	152	32	35	3	48
34	153	33	36	3	48
34	153	30	38	8	90
36	156	30	41	11	104
36.5	156	34	38	4	60
36.5	156	34	44	10	100
38.5	159	37	41	4	60
44	164	43	48	5	70
45	165	42	46	4	60
64	181	54	67	13	111
74.5	187	64	98	34	151
92	196	86	95	9	95
144.5	216	132	154	22	134

Tracciando su un grafico i valori relativi ai logaritmi di M e di d_F, potremmo notare un andamento lineare di tipo crescente; per stimare il valore di b possiamo scegliere coppie di punti poste agli estremi opposti rispetto ai valo-

ri di M. Se dalla nostra tabella, in cui i dati sono ordinati per valori crescenti di M, scegliamo la coppia di valori 216,134 e successivamente la prima e la seconda coppia di valori, otteniamo come stima della pendenza della retta 0.8 e 1.1 rispettivamente. Per la trasformazione dei dati potremmo, perciò, orientarci verso la trasformazione logaritmica o la radice quadrata. Un'analisi successiva sui dati trasformati ci consentirà di valutare la bontà del risultato ottenuto.

4. INTERPOLAZIONE NELLA FASE ESPLORATIVA

4.1 Premessa

Supponiamo di avere n coppie ordinate (x_i, y_i) per i=1,...,n. Nella fase esplorativa, rappresentare su di un grafico i valori y in funzione dei valori x, può certamente aiutarci ad evidenziare eventuali relazioni fra questi, specie quando non conosciamo nulla sulla connessione logica che intercorre fra le due o addirittura pensiamo che non ve ne dovrebbe essere alcuna (in quest'ultimo caso, solo se, evidentemente, non ci fidiamo completamente delle nostre idee). La ricchezza delle interrelazioni esistenti nella realtà, infatti, è spesso superiore alle nostre conoscenze (e, a volte, alla nostra fantasia).

Ciò che si vuole sottolineare è che un'ispezione di tipo visivo, con la flessibilità implicita in essa, insieme con strumenti di validità generale (stime resistenti),

riveste grande importanza nel momento in cui cominciano a raccogliere indicazioni sulla presenza di eventuali regolarità nell'insieme delle coppie. Più che la immediata applicazione di un test statistico, ad esempio, sulla casualità della distribuzione dei punti, vale all'inizio una visione complessiva che potrà suggerirci più di una via da percorrere.

Nelle situazioni più convenzionali il grafico si adegua alla semplice struttura

(fattore, risposta)

per cui nella variabile x individuiamo preventivamente un elemento esplicativo o, più spesso, la circostanza, che genera la risposta, quantificata nella variabile y.

Rilevare una struttura regolare nel caso di coppie di valori significa, in buona sostanza, individuare una funzione matematica che lega la y alla x, che spieghi a grandi linee la dispersione dei punti sul grafico.

E' indubbiamente un vantaggio, per la semplificazione che

porta alla analisi successiva, che questa relazione sia di tipo lineare.

Conviene, perciò, assumere come primo obiettivo la linearizzazione della relazione che ci siamo convinti esista fra le due variabili. Ciò si può ottenere, in una gran parte dei casi, trasformando una delle due o entrambe le variabili con funzioni appartenenti alla famiglia di trasformazioni descritta nel capitolo precedente.
Daremo poi dei metodi per raggiungere questo risultato e per valutarne la bontà.

Qui vogliamo soprattutto sottolineare come l'esistenza di una relazione lineare fra y ed x faciliti di molto il prosieguo dell'analisi. Sottraendo semplicemente ai dati originari, eventualmente trasformati, l'andamento lineare rilevato, su una scala, ampliata per il fatto che le variazioni in y si sono certamente ridotte, possiamo valutare più chiaramente, come in un ingrandimento, comportamenti locali che si discostano dalla tendenza lineare o al-

tre idiosincrasie nell'andamento generale dei dati.

Nell'ottica dell'analisi esplorativa, evidenziare una relazione di tipo lineare tra i dati, specie se incompleta, è solo un primo passo verso una comprensione effettiva dei dati. La successiva analisi dei residui ci permette di guardare, su una scala verticale espansa, più in profondità ai comportamenti che ancora caratterizzano i dati, di vedere ciò che accade ad un livello più sottile.

4.2 Una linea resistente

Il metodo più noto e diffuso nelle applicazioni per stimare la retta di regressione è certamente il metodo dei minimi quadrati. Questo richiede, come è noto, delle ipotesi piuttosto stringenti, raramente verificate dai dati.
Nella maggior parte dei casi il grafico dei punti (x_i, y_i) presenta dei valori che si discostano fortemente dalla generale tendenza lineare. La presenza di dati anomali o di errori nelle misurazioni, compromette, anche quando questi son presenti in numero limitato, la stima dei parametri della retta.

Il metodo qui descritto (Tukey, 1970), che per essere correttamente inteso va inserito nell'ottica generale descritta nella introduzione, è, al contrario, caratterizzato da una notevole resistenza a tali punti. Diciamo fin d'ora che possiamo sostituire fino ad 1/6 delle n coppie con valori anomali eccezionalmente elevati oppure molto bas-

si, senza che la nostra stima iniziale dei parametri della retta ne risulti affetta.

E' inoltre parte intrinseca al metodo la successiva analisi dei residui, che interviene in due direzioni distinte, e precisamente

- da una parte, per consentirci, attraverso la verifica della presenza di una struttura regolare, ancora di tipo lineare, in essi, di correggere la stima della pendenza della nostra retta iniziale
- dall'altra per consentirci, attraverso una analisi accurata di eventuali punti fortemente discosti dalla generale tendenza lineare, una comprensione effettiva di questi.

L'idea base è quella di scegliere tre punti che descrivano sinteticamente la struttura lineare dei dati in modo resistente, utilizzando come stime dei valori mediani.

Analiticamente procediamo al modo seguente:

- dopo aver ordinato le n coppie (x_i, y_i) rispetto ai valori di x in modo crescente, le suddividiamo in 3 gruppi di k coppie ciascuno se $n=3k$,

  k, k+1, k coppie nell'ordine se $n=3k+1$,

  k+1, k, k+1 coppie nell'ordine se $n=3k+2$;

- si calcola la mediana separatamente per le due coordinate in ciascuno dei tre gruppi. Indichiamo con $\{L\}$, $\{M\}$, $\{R\}$ l'insieme degli indici corrispondenti al 1°, 2°, e 3° gruppo rispettivamente; definiamo, intendendo per "med" il valore mediano,

$$x_L = \underset{i\in\{L\}}{\text{med}}\,(x_i) \qquad y_L = \underset{i\in\{L\}}{\text{med}}\,(y_i)$$

$$x_M = \underset{i\in\{M\}}{\text{med}}\,(x_i) \qquad y_M = \underset{i\in\{M\}}{\text{med}}\,(y_i)$$

$$x_R = \underset{i\in\{R\}}{\text{med}}\,(x_i) \qquad y_R = \underset{i\in\{R\}}{\text{med}}\,(y_i)\ ;$$

- con i tre punti così definiti calcoliamo la pendenza b_o e il valore a_o dell'ordinata corrispondente a $x=x_M$ della retta interpolante resistente $\hat{y} = a_o + b_o(x-x_M)$ ponendo

$$b_o = \frac{y_R - y_L}{x_R - x_L}$$

$$a_o = \frac{1}{3} |(y_L - b_o x_L) + (y_M - b_o x_M) + (y_R - b_o x_R)|;$$

b_o è perciò la pendenza della retta passante per i punti mediani (x_L, y_L), (x_R, y_R) dei due gruppi esterni, ed il valore a_o è calcolato come la media delle ordinate corrispondenti ad x_M delle 3 rette di pendenza b_o e passanti per i punti mediani dei gruppi $\{L\}$, $\{M\}$, $\{R\}$.

Per quanto riguarda le caratteristiche di resistenza, c'è da osservare che in ciascuno dei tre gruppi, separatamente, possiamo sostituire 1/2 dei punti con valori anomali senza che i valori mediani ne risultino affetti. Nell'ipotesi più restrittiva che tutti i punti anomali cadano all'interno di un sol gruppo, il valore massimo di punti che può essere sostituito senza che ne risultino modificati i parametri della retta è chiaramente 1/6.

La retta interpolante di parametri a_o, b_o è stata costruita seguendo un criterio essenzialmente empirico e può

non rappresentare nella sua interezza la relazione lineare che intercorre fra le y e le x. Per procedere nell'analisi calcoliamo i residui r_i^o

$$r_i^o = y_i - (a_o + b_o(x_i - x_M))$$

e verifichiamo se le coppie di valori (x_i, r_i) conservano un andamento di tipo lineare. Usando il metodo su descritto possiamo arrivare ad una relazione lineare del tipo

$$\hat{r}_i^o = \gamma_1 + \delta_1(x_i - x_M) \ .$$

Dato che riterremo soddisfacente la nostra interpolazione solo in assenza di relazione fra i residui e la variabile x, nel caso che i valori ottenuti per i parametri della retta scritta nella formula precedente siano significativamente diversi da 0, dovremo intervenire sui valori a_o e b_o a definire più efficacemente la relazione lineare tra le x e le y.

Una prima idea può essere quella di trasferire la relazione lineare presente nei residui a quella fra le x e le y ponendo

$$b_1 = b_o + \delta_1 \qquad a_1 = a_o + \gamma_1 \ .$$

La nuova retta interpolante

$$\hat{y}_i = a_1 + b_1(x_i - x_M)$$

avrà come residui le quantità r_i^1 dove

$$r_i^1 = r_i^o - \hat{r}_i^o \quad .$$

Potremmo iterare il procedimento: per un generico passo j avremmo i residui r_i^j dove

$$r_i^j = y_i - (a_j + b_j (x_i - x_M))$$

da essi ricaveremmo le quantità

$$\gamma_{j+1} = \frac{1}{3}\left\{[r_L^j - b_j(x_L - x_M)] + r_M^j + [r_R^j - b_j (x_R - x_M)]\right\}$$

$$\delta_{j+1} = \frac{r_R^j - r_L^j}{x_R - x_L} \quad ,$$

dove r_R^j , r_M^j e r_L^j sono valori mediani dei residui alla j-esima iterazione per il 3°, 2° e 1° gruppo rispettivamente, e le due quantità δ_{j+1}, γ_{j+1} rappresentano la pendenza e l'ordinata per $x=x_M$ della retta resistente che interpola le coppie (x_i, r_i^j).

Un tal modo di procedere, se ha un valore descrit-

tivo, in termini generali, dei criteri e della logica della analisi esplorativa, non presenta, come algoritmo, proprietà di convergenza, come è stato dimostrato dall'esempio riportato da A. Siegel (Biometrika 69, 1982). Considerando i seguenti dati (-4,0),(-3,0),(-2,0),(-1,0),(0,0),(1,0),(2,-5), (3,5),(12,1), il procedimento descritto acquista alla 7^a iterazione un comportamento oscillante; infatti i valori di b ed a passano alternativamente da -0.694, -1.39 a 0.8333, 1.67.

In un loro lavoro Johnstone e Vellemann (1982) hanno dato una risposta al problema della convergenza nel modo che ora ridescriviamo.

Il valore della correzione per la pendenza b_j al j-esimo passo è data da

$$\delta_{j+1} = \frac{r_R^j - r_L^j}{x_R - x_L} .$$

Tenendo conto che la differenza al numeratore dipende solo dal valore b_j, tralasciando l'indice relativo all'iterazio-

ne, possiamo scrivere che

$$\Delta r(b) = r_R(b) - r_L(b) \quad .$$

Riproponiamo ora il nostro problema nei seguenti termini: cerchiamo il valore b tale che $\Delta r(b) = 0$.

Prima di proseguire portiamo l'origine delle ascisse nel punto x_M; senza modificare la scrittura intenderemo per x_i la differenza del valore originario con x_M.

Studiamo separatamente l'andamento di $r_L(b)$ e $r_R(b)$, tenendo conto che possiamo completamente tralasciare in questa analisi il valore dell'intercetta, in quanto nella loro differenza essa scompare. Perciò, quando parleremo di residui rispetto ad un valore b della pendenza, li intenderemo a meno del valore dell'intercetta.

Consideriamo nel piano (x,y) il fascio di rette di centro l'origine, $y=bx$; esse hanno per i punti di ascissa $x=x_i$ residui pari a y_i-bx_i (che a meno dell'intercetta sono i residui di una generica retta interpolante $y=a+bx$); questa quantità equivale, chiaramente, all'intercetta del-

la retta di coefficiente angolare b passante per il punto (x_i,y_i). Consideriamo, perciò, i fasci di rette con centro in (x_i,y_i), ottenuti al variare dell'indice i; la equazione di una generica retta del generico fascio è $(y-y_i) = b(x-x_i)$; al variare di b in essa avremo sull'intercetta i valori corrispondente del residuo in (x_i,y_i); al variare del fascio considerato avremo i residui corrispondenti a ciascun punto.

Consideriamo, ora, il piano duale al piano x,y, in cui poniamo come ascissa il valore b del coefficiente angolare e in ordinate il valore della intercetta. Una retta di coefficiente angolare b e valore dell'intercetta pari a $y_i - bx_i$ sarà rappresentata dal punto (b,y_i-bx_i) nel piano duale. Il fascio di retta avente come centro il punto (x_i,y_i) sarà rappresentato da una retta passante per i punti $(0,y_i)$ e $(y_i/x_i,0)$ (in quanto per b=0, quando x=0 $y=y_i$; in quanto per valore dell'intercetta nulla $b=y_i/x_i$). La pendenza della retta è pari $-x_i$.

Per ciascun punto (x_i,y_i) avremo nel piano duale una retta che in corrispondenza di b ci dà in ordinata il valore del residuo.

Osserviamo che ai punti (x_i,y_i) con $i \in \{L\}$ corrispondono rette con pendenza positiva, e a quelli con $i \in \{R\}$ rette con pendenza negativa.

Le funzioni $r_L(b)$ e $r_R(b)$ di cui riportiamo la definizione

(1) $$r_L(b) = \underset{i\in\{L\}}{\text{med}} \; (y_i - bx_i)$$

(2) $$r_R(b) = \underset{i\in\{R\}}{\text{med}} \; (y_i - bx_i)$$

Hanno le seguenti caratteristiche:

- andamento monotono crescente e monotono decrescente rispettivamente
- sono continue per ogni b ed hanno derivata continua tranne che in un numero finito di punti ove essa non è definita;

Per dimostrare quanto detto, osserviamo inizialmente che le rette corrispondenti agli l punti $i \in \{L\}$ hanno al massimo $l(l-1)/2$ punti distinti in cui si intersecano due a due, in

corrispondenza di certi valori, che possono non essere tutti distinti, $\{b_s\}$ $s=1,\ldots,l(l-1)/2$.

In corrispondenza di b, nella (1) scegliere il valore mediano fra i residui corrisponde a scegliere una retta mediana di pendenza $-x_i$ e quota y_i, dove (x_i,y_i) è il punto che per b' dà il valore mediano dei residui.

Se $b' \notin \{b_s\}$ esiste un intorno di b' tale che la retta mediana prescelta continua ad essere quella relativa al punto (x_i,y_i) (che continua perciò a fornire, per quello intervallo di b, il valore mediano dei residui).

Se $b' \in \{b_s\}$, ma la retta mediana in questione non è interessata a questo valore, in quanto non ha punti d'intersezione in b', vale ancora la proprietà precedente.

Se $b' \in \{b_s\}$ e la retta mediana, di pendenza $-x_i$ e quota y_i, corrispondente al valore $b'-\varepsilon$ con ε piccolo a piacere, ha in tale valore d'ascissa un punto di intersezione con la retta di pendenza $-x_j$ e quota y_j, in b' il valore di $r_L(b)$ continuerà ad essere perfettamente definito dal valore co-

mune alle due rette, ma subirà un salto nella sua derivata prima, passando la sua pendenza da $-x_i$ a $-x_j$.

Un ragionamento analogo vale per $r_R(b)$.

Dato che al variare di b da $-\infty$ a $+\infty$ i valori di $r_L(b)$ variano da $-\infty$ a $+\infty$ e quelli di $r_R(b)$ da $+\infty$ a $-\infty$ ed essendo le due funzioni continue esiste un punto $\bar{b}$ tale che $r_L(\bar{b}) = r_R(\bar{b})$.

Dal punto di vista applicativo, calcolato il valore $b_0, \Delta r(b_0)$ potrà essere nullo, positivo, o negativo.

Se è nullo il valore della pendenza è quello cercato.

Se $\Delta < 0$ cercheremo un valore $b_1 > b_0$, che sicuramente esiste finito tale che $\Delta r(b_1) > 0$; il valore cercato è interno all'intervallo (b_0, b_1); procedendo iterativamente con

$$b_{j+1} = b_j + \Delta r(b_j) \frac{b_j - b_{j-1}}{\Delta r(b_j) - \Delta r(b_{j-1})}$$

possiamo avvicinarci quanto vogliamo al valore $\bar{b}$.

Sulla base di questo approccio conviene definire l'intercetta semplicemente come il valore mediano dei resi-

dui

$$a = \text{med}\ (y_i - \bar{b}x_i) \quad .$$

In tal modo la retta interpolante è fissata a una quota tale che il numero dei punti al di sopra di essa è pari a quelli che sono al di sotto.

4.3 Un criterio empirico per la linearizzazione

I tre punti mediani, che abbiamo utilizzato per costruire la nostra linea resistente, possono fornirci utili indicazioni per giudicare il grado di linearità nella relazione fra le y e le x.

Se calcoliamo, infatti, la pendenza delle due rette che passano per il punto (x_M, y_M) e rispettivamente i punti (x_L, y_L) (x_R, y_R), dal loro rapporto ci viene tale valutazione approssimativa; precisamente:

quanto più tale valore è vicino ad 1 tanto più possiamo ritenere soddisfacentemente lineare la relazione tra le due variabili; quanto più tale valore è lontano da 1 tanto più marcata sarà sul grafico la presenza di una qualche forma di concavità.

In questo secondo caso, prima di passare all'interpolazione della retta, si possono provare gli effetti sull'andamento generale, stimato come sopra, della trasfor-

mazione della y con alcune delle seguenti funzioni:

y^3, y^2, y, $\sqrt{y}$, $\log y$, $-1/\sqrt{y}$, $-1/y^2$, $-1/y^3$.

Nel caso che i tre punti mediani, una volta che siano stati uniti con segmenti di retta, presentino una concavità rivolta verso il basso, si parte dalla sinistra della scala di potenze su scritte; e dalla sua destra se la concavità è rivolta verso l'alto.

Una valutazione della concavità si può anche avere in base al segno della pendenza delle due rette parziali e dal valore del loro rapporto; infatti se la pendenza è positiva ed il rapporto fra queste è >1 la concavità sarà rivolta verso l'alto e al contrario verso il basso se quel rapporto è <1; nel caso di pendenza negativa si ha una situazione perfettamente simmetrica.

Un primo criterio [10] per valutare la trasformazione di potenza cui sottoporre la variabile y per raggiungere una relazione lineare fra le due variabili è il seguente.

Si rappresentino su un piano i punti di ascissa

$$\frac{C^2 (x - M_x)^2}{2 M_y}$$

e di ordinata

$$y - M_y - C(x-M_x) \ ,$$

dove x e y rappresentano le coppie delle osservazioni, M_x e M_y le loro rispettive mediane e C è la pendenza della linea resistente ricavata dai dati grezzi. Se questi punti si dispongono con buona approssimazione su una retta di pendenza b, il valore cercato della potenza della trasformazione è p = 1-b.

Supponiamo, infatti, che il modello reale dei nostri dati sia

(1) $\quad y^P - M_y^P = K(x-M_x)$.

Definiamo $z = y^P$; se $p \neq 0$ allora $M_y \simeq M_z^{1/P}$.

Sviluppiamo in serie di Taylor $z^{1/P}$ intorno a M_z nella approssimazione al 2° ordine:

$$y = z^{1/P} \simeq M_z^{1/P} + \frac{1}{P} M_z^{(1/P)-1} (z-M_z) + \frac{1-P}{2P^2} M_z^{(1/P)-2} (z-M_z)^2 .$$

Utilizzando la (1) otteniamo

$$y \simeq M_y + \left(\frac{K M_y}{p M_z}\right)(x - M_x) + \frac{1-P}{2M_y}\left(\frac{K M_y}{PM_z}\right)^2 (x-M_x)^2 .$$

Se il valore di $\left(\frac{K M_y}{P M_z}\right)$ è ben approssimato dal valore C della

pendenza della retta ottenuta interpolando in modo resistente i dati grezzi possiamo riscrivere l'equazione precedente come

$$y - M_y - C(x-M_x) \simeq (1-P) \frac{C^2 (x - M_x)^2}{2 M_y}$$

dove (1-p) è la sola incognita.

Per dare una valutazione approssimata e del grado della trasformazione necessaria per linearizzare, si può anche ragionare al modo seguente:

supponiamo che la y ela x siano legate da una funzione continua fino al 2° ordine $y=f(x)$; supponiamo che nel punto (x_M, y_M) la tangente ad essa abbia pendenza pari a

$$\frac{y_R - y_L}{x_R - x_L} = f'(X_M)$$

e che la derivata seconda abbia valore approssimativamente pari a

$$f''(X_M) = \frac{\frac{y_R - y_M}{x_R - x_M} - \frac{y_M - y_L}{x_M - x_L}}{X_R - X_L} .$$

Ora imponiamo che nel punto x_M, la curvatura della funzione

trasformata $\phi(x)=f^{p}(x)$ sia nulla.

In quanto è $\quad \phi'(x) = p\, f(x)^{p-1}\, f'(x)$

$$\phi''(x)= p(p-1)\, f(x)^{p-2} \left(f'(x)\right)^2 + p\, f(x)^{p-1}\, f''(x) \ ,$$

la curvatura della ϕ vale $\quad \dfrac{|p.f(x)^{p-2}|\ |(p-1)f'(x)^2+f(x)f''(x)|}{\left((1+(f'(x))^2\right)^{3/2}}$

Imponendo che tale valore sia nullo per $x=x_M$, avremo per p una stima approssimata data dalla condizione

$$(p-1)\left(f'(x_M)\right)^2+f(x_M)f''(x_M) = 0 \ .$$

Il valore di p così ricavato può essere un buon punto di partenza nella scala di potenze sopra scritta.

Se la trasformazione sulla y ha raggiunto solo un risultato parziale, oppure già in partenza abbiamo rinunciato a trasformarla, possiamo pensare di agire sulla x, usando una scala di trasformazioni identica a quella data per y. Bisogna tener conto che ora il criterio per stabilire il punto di partenza è il seguente:

partiremo dalla sinistra della scala se la concavità è rivolta verso l'asse delle y; al contrario dalla parte destra del-

la scala se la concavità è rivolta dalla parte opposta all'asse delle y.

4.4 Un esempio

Riportiamo i dati del consumo medio giornaliero di gasolio misurato in dm^3/h e le corrispondenti differenze medie di temperatura fra ambiente interno ed esterno misurato in °C (1), relative a uno stabile di Princeton per 57 giorni del periodo invernale.

Con y indicheremo il consumo di gasolio con x le differenze di temperatura

x	y	x	y
10.6	0.0	23.3	625.9
11.7	0.0	26.7	625.9
11.1	22.7	25.6	625.9
16.1	246.6	23.3	642.9
20.0	399.3	13.8	320.0
17.8	271.8	20.0	317.2
18.3	277.5	23.9	433.3
18.9	260.5	23.9	447.5
14.4	141.6	20.0	373.8
17.2	320.0	18.8	351.2
17.8	218.0	22.2	464.4
11.1	90.6	26.1	577.7
6.1	209.3	21.1	368.2
21.1	396.5	25.0	620.2
23.3	543.7	28.3	815.6
17.8	254.9	28.9	722.2
14.4	179.4	24.4	577.7
20.6	356.8	25.0	523.9
23.9	637.2	31.1	849.6
17.2	436.1	30.5	928.9
16.1	237.9	25.0	642.9
18.3	390.8	23.9	492.8
17.2	351.2	22.2	481.4
15.6	167.1	16.7	354.0
12.2	90.6	23.3	608.9
17.2	235.1	21.6	478.6
19.4	362.5	29.3	693.8
23.9	543.7	34.4	896.9
29.4	877.9		

(1) Essendo la temperatura interna praticamente costante, tali variazioni corrispondono a variazioni della temperatura esterna.

Una volta ordinate le coppie di dati per valori crescenti della x, è facile ricavare le seguenti quantità, per il cui significato ci rifacciamo al paragrafo 2 dell'attuale capitolo,

$$
\begin{array}{lll}
 & x_L = 15.85 & y_L = 226,55 \\
(1) & x_M = 20.85 & y_M = 393,65 \\
 & x_R = 25.6 & y_R = 625,9 \ .
\end{array}
$$

Da questi valori possiamo ricavare un valore iniziale delle pendenze della retta interpolante

$$b_o = \frac{y_R - y_L}{x_R - x_L} = 40.96 \ .$$

Come anche i valori delle pendenze delle rette che interpolano separatamente la parte sinistra e destra del grafico dei punti:

$$\text{L.H.S.} = \frac{y_M - y_L}{x_M - x_L} = 33.42$$

$$\text{R.H.S.} = \frac{y_R - y_M}{x_R - x_M} = 48.89 \ ,$$

dove L.H.S. e R.H.S. indicano rispettivamente la pendenza a

sinistra e a destra.

Il loro rapporto è 1.46.

Per realizzare un andamento lineare migliore proviamo ad applicare ai dati la trasformazione logaritmica; con essa il rapporto fra le due pendenze scende a 0.87. Risalendo nella scala delle trasformazioni proposte nel precedente paragrafo, applichiamo la radice quadrata; essa dà un rapporto pari a 1.13.

Si fa presente che per arrivare a questi risultati è sufficiente applicare le trasformazioni solo sulle quantità (1), dato che queste lasciano l'ordinamento dei dati inalterato.

Provando, infine, un valore intermedio fra 0 e 1/2, con la radice quarta otteniamo un rapporto fra le pendenze pari circa a 1.01.

Con questa trasformazione, utilizzando una procedura meccanografica, che applica la procedura iterativa descritta al paragrafo 2, otteniamo per la equazione della retta resistente la seguente espressione

$$\sqrt[4]{y} = 0.113\ x + 2.064 \quad .$$

Dei residui presentiamo il seguente diagramma albero e foglia.

```
                                    UNIT =    0.0100
      LO  -339, -326, -114

       5    -3. 65
       5    -3*
       6    -2. 5
      10    -2* 4441
      14    -1. 9986
      19    -1* 42210
      25    -0. 998665
      28    -0* 420
     (5)    +0* 02334
      24    +0. 55666899
      16     1* 1223
      12     1. 5
      11     2* 14
       9     2. 58
       7     3* 002
       4     3. 7

      HI    55,  59, 104
```

Esso evidenzia 1) che il valore mediano è 0, 2) che la distribuzione dei residui si concentra in modo non perfettamente simmetrico fa i valori - 2 e + 1, 3) che possibili dati anomali sono quelli corrispondenti ai residui presenti sulle righe LO e HI.

Riportiamo anche le stime successive della pendenza ottenute ai vari passi del procedimento iterativo; esso si è fermato alla 3^a iterazione in quanto la successive non portava correzioni superiori all'ordine di 10^{-4}.

Dallo stampato risulta anche il valore, probabilmente più preciso, del rapporto fra le pendenze destra e sinistra.

```
STRAIGHTNESS CHECK.
 LEFT HALF-SLOPE =     0.115027 RIGHT HALF-SLOPE =     0.115272
          RATIO =     1.002127


SLOPE   1:       0.115147
SLOPE   2:       0.112740
SLOPE   3:       0.113642
```

4.5 Estensione a più dimensioni

Si può cercare di estendere il criterio esposto nel paragrafo precedente al caso in cui le osservazioni siano costituite da n-uple di valori. Per fissare le idee poniamoci nel caso n=4. Ci poniamo, perciò, il problema di ricercare la regressione di una variabile (esogena), la risposta, in funzione di altre tre (endogene), i fattori. Da un punto di vista geometrico, ciò significa determinare un iperpiano di regressione passante per la nuvola dei punti- osservazione basandoci su di un criterio differente da quello dei minimi quadrati.

L'idea iniziale per la determinazione dei parametri, si avvicina alla procedura classica che tratta la regressione multipla come successive sottrazioni degli effetti, ma se ne distacca nel momento in cui prende in considerazione, al fine di valutare le relazioni lineari che intervengono, il metodo di interpolazione con i tre punti media-

ni precedentemente descritto.

Denotiamo le osservazioni con la quadrupla di variabili ordinate $(x_{1i}, x_{2i}, x_{3i}, y_i)$ per $i=1,\ldots,n$ e supponiamo che la y rappresenti la variabile che si vuole spiegare linearmente in termini delle altre tre.

Tramite il criterio stabilito nel paragrafo precedente, possiamo valutare le seguenti relazioni lineari

$$x_2 = a + b\,(x_1 - x_{1M})$$

$$x_3 = c + d\,(x_1 - x_{1M})$$

$$y = e + f(x_1 - x_{1M}) \quad .$$

Da queste interpolazioni possiamo facilmente ricavare i residui

$$x_2' = x_2 - a - b\,(x_1 - x_{1M})$$

$$x_3' = x_3 - c - d(x_1 - x_{1M})$$

$$y' = y - e - f\,(x_1 - x_{1M}) \quad .$$

Possiamo a questo punto impiegare tali residui in luogo delle variabili originarie, nei quali non compare, proprio per averlo eliminato nella misura in cui sia di tipo lineare, lo effetto della variabile x_1.

Costruiamo, così, le seguenti e successive regressioni in funzione della variabile x_2

$$x_3' = a' + b' x_2'$$

$$y' = c' + d' x_2'$$

e quindi i residui relativi

$$x_3'' = x_3' - a' - b' x_2'$$

$$y'' = y' - c' - d' x_2'$$

che, interpolando a loro volta linearmente, danno

$$y'' = a'' + b'' x_3'' \quad .$$

A questo punto, dal momento che il nostro scopo principale consiste nel determinare la regressione multipla, conviene mettere assieme i vari pezzi e ricomporre ordinatamente le successive regressioni effettuate. Avremo

$$y' - c' - d' x_2' = a'' + b''(x_3' - a' - b' x_2')$$

$$y - e - f(x_1 - x_{1M}) - c' - d' [x_2 - a - b(x_1 - x_{1M})] = a'' + b'' \{ x_3 - c - d(x_1 - x_{1M}) - a' - b' [x_2 - a - b(x_1 - x_{1M})] \} \quad .$$

Espressione che si semplifica nella

$$y = A + \hat{\beta}_1 (x_1 - x_{1M}) + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

ove

$A = e+c'+a(b'-d') + a'' - b'' (c+d+a')$

$\hat{\beta}_1 = f + b(b'-d') - d\,b''$, $\hat{\beta}_2 = d'-b'$, $\hat{\beta}_3 = b''$.

Conviene, a questo punto, riparametrizzare il termine A per dare alla funzione di regressione una forma simmetrica rispetto alle variabili endogene; a ciò basta porre

$$y_c = e+c'+a(b'-d')+a''+b''(c+d+a') + \hat{\beta}_2\, x_{2M} + \hat{\beta}_3\, x_{3M}$$

per avere infine

$$y = y_c + \hat{\beta}_1\,(x_1-x_{1M}) + \hat{\beta}_2\,(x_2-x_{2M}) + \hat{\beta}_3\,(x_3-x_{3M}) \ .$$

Proponiamo, alla fine di questo capitolo, una misura globale di accostamento per giudicare l'adeguatezza della interpolazione ai dati, espressa nel contesto di questo paragrafo ma valida anche nel caso unidimensiale. Una valutazione globale della bontà dell'accostamento si può ottenere ragguagliando la differenza sui quarti (o differenza interquartile) dei residui: $y - y_c - \hat{\beta}_1\,(x_1-x_{1M}) - \hat{\beta}_2(x_2-x_{2M}) - \hat{\beta}_3\,(x_3-x_{3M})$, indicato come $F_u\,(res) - F_L\,(res)$, all'analoga misura di variabilità delle osservazioni y : $F_u(y)-F_L(y)$. Il complemento a 1

di tale rapporto

$$RR = 1 - \frac{F_u(res) - F_L(res)}{F_u(y) - F_L(y)}$$

fornisce un indice di adeguatezza della interpolazione resistente (RR), che risulta certamente pari ad 1 quando almeno la metà dei residui rispetto al piano interpolante sono nulli e, di contro, si annulla nel caso in cui la variabilità intorno al piano di regressione resta identica a quella della sola variabile y. Possiamo interpretare, in analogia trasparente con l'accezione riservata al quadrato del coefficiente di correlazione multipla, tale quantità come la riduzione proporzionale della variabilità della variabile y dopo aver effettuato l'operazione di interpolazione, ovvero come la parte di variabilità "spiegata" dal piano di regressione.

5. ANALISI DI UNA TABELLA A DOPPIA ENTRATA

5.1 Premessa

Una tabella a doppia entrata è un insieme di dati in cui le osservazioni sono date nella forma matriciale

$$(y_{ij}) \qquad i=1,\ldots,I,\ j=1,\ldots,J$$

Questo tipo di struttura coinvolge tre variabili: il fattore riga che ha I livelli possibili, il fattore colonna che ha J livelli e la risposta di cui possediamo IxJ osservazioni. (La intersezione di una riga e di una colonna è detta cella).

E' una struttura che capita assai frequentemente di incontrare: se abbiamo, per esempio, tassi di natalità per varie combinazioni dell'età della madre con qualche caratteristica sociale (paese, livelli di reddito, numero di figli viventi, etc.), abbiamo una tabella a doppia entrata; se abbiamo la produzione di grano, per esempio, per diverse combinazioni di qualità del grano con quantità e tipo di fertiliz-

zante usato, abbiamo una tabella a doppia entrata; e così di seguito.

E' una struttura interessante anche per la diversità e la ricchezza delle analisi a cui può introdurre, a motivo delle diverse ipotesi che si possono proporre per spiegare le possibili combinazioni di dati.

Tuttavia, in questa sede, ci limiteremo a considerare la più semplice delle relazioni pensabili tra le tre variabili.

Ci poniamo nell'ipotesi che i due fattori, riga e colonna, contribuiscano separatamente ed additivamente alla variabile risposta y. Se i dati dovessero allontanarsi sistematicamente da questa struttura, invece che tentare altri possibili modelli, cercheremo di rimuovere la non additività riesprimendo i dati con trasformazioni di potenza.

Il modello additivo può essere formalmente scritto come:

$$(1) \qquad y_{ij} = m + a_i + b_j + e_{ij} \quad .$$

Qui m è un valore tipico della tabella, una sorta di livello comune rispetto a cui si dispongono le osservazioni. Il contributo incrementale del livello i del fattore riga rispetto al livello globale m è espresso da a_i, l'"effetto riga". Allo stesso modo b_j rappresenta il contributo incrementale del livello j del fattore colonna, l'"effetto colonna". Infine e_{ij} rappresenta lo scostamento di y_{ij} dal modello puramente additivo $m+a_i+b_j$; spesso si cerca di trattarli come fluttuazioni casuali.

In questo contesto, analisi di una tabella a doppia entrata significa decomporre l'insieme dei valori nei quattro elementi specificati nella (1).

Prima di passare a descrivere la tecnica di decomposizione proposta nell'ambito dell'EDA, vogliamo brevemente accennare al metodo classico basato sull'uso delle medie. In questo caso, per decomporre la tabella (y_{ij}) secondo il modello (1), basta porre

$$(2a) \qquad \hat{m} = \frac{1}{IJ} \sum_i \sum_j y_{ij}$$

(2 b) $\hat{a}_i = \frac{1}{J} \sum_j (y_{ij} - \hat{m})$ $\qquad i=1,\ldots,I$

(2 c) $\hat{b}_j = \frac{1}{I} \sum_i (y_{ij} - \hat{m})$ $\qquad j=1,\ldots J$.

Il metodo non ha bisogno di iterazioni e consente di ottenere in un solo passo l'effetto globale, gli effetti riga e colonna.

I vantaggi teorici del metodo sono legati al modo classico di stabilire la qualità dell'adattamento di un modello ai dati osservati.

Definiamo i residui come

$$r_{ij} = y_{ij} - \hat{m} - \hat{a}_i - \hat{b}_j$$

e la somma dei residui al quadrato come

$$SQR = \sum_i \sum_j r_{ij}^2$$

Il modello additivo che meglio si adatta ai dati è definito, nell'ambito classico, come quello per cui i valori di $\hat{m}$, $\hat{a}_i$, e $\hat{b}_j$ rendono minima la SQR. E' ben noto che le (2a), (2b)e(2c) definiscono i valori con questa proprietà.

La stima dei minimi quadrati è ottimale in svariati sensi, su un piano teorico, quando i dati hanno certe proprietà speciali; si assume spesso, per esempio, che le e_{ij} della (1) siano errori o fluttuazioni distribuiti in modo indipendente secondo una distribuzione gaussiana con media zero e uguale varianza. Quando, però, queste proprietà non valgano, altri criteri che non quello dei minimi quadrati potrebbero rilevarsi più efficaci.

In modo più immediato, se nei dati sono presenti errori grossolani, capitati nella tabella e inconsistenti con la struttura sottostante ad essa, questi esercitano sicuramente un effetto sensibile sulle quantità calcolate nelle formule (2).

Nel prossimo paragrafo descriveremo una tecnica di analisi indubbiamente più resistente, di modo che disturbi violenti ed isolati in poche celle non influenzino fortemente la stima del valore dell'effetto globale e degli effetti riga e colonna, e siano, così, riflessi nei residui.

5.2 Il metodo delle mediane

La tecnica proposta [10] nell'ambito dell'EDA per l'analisi di una tabella a doppia entrata consiste in un processo iterativo di sottrazioni successive alternantesi dei valori mediani di ciascuna riga dagli elementi della riga stessa e dei valori mediani di ciascuna colonna dagli elementi della stessa.

Più discorsivamente, partendo dalle righe, una iterazione completa procede così: 1) si calcola il valore mediano di ciascuna riga, lo si sottrae ai singoli elementi della riga corrispondente; 2) sulle quantità residue si calcolano le mediane di ciascuna colonna e le si sottraggono dagli elementi delle colonne corrispondenti; 3) dalla colonna di mediane,ottenuta nel primo passo, calcoliamo il valore mediano che sottraiamo ai valori della colonna stessa; dalla riga di mediane del secondo passo calcoliamo il valore mediano che sottraiamo ai valori della riga stessa.

La somma di questi due valori mediani dà una prima stima dell'effetto globale, la colonna e la riga delle mediane depurate dai lori valori mediani ci dà una prima stima degli effetti riga e colonna rispettivamente.

Questo procedimento può essere iterato finchè qualche riga o colonna dei residui non abbia mediana nulla.

Per una presentazione formale del metodo, poniamo che il modello stimato della (1) dal metodo delle mediane dopo n iterazioni è il seguente

$$y_{ij} = m^{(n)} + a_i^{(n)} + b_j^{(n)} + e_{ij}^{(n)} \quad , \tag{3}$$

con valore iniziale, per n=0 ,

$$\begin{aligned} m^{(0)} &= 0 \\ a_i^{(0)} &= 0 \qquad\qquad i = 1,\ldots,I \\ b_j^{(0)} &= 0 \qquad\qquad j = 1,\ldots,J. \end{aligned} \tag{4}$$

Il procedimento del metodo di sottrazione delle mediane, nel caso che si inizi la iterazione dalle righe, è formalmente descritto dalle seguenti 6 equazioni

Righe:

$$(5a) \quad \Delta a_i^{(n)} = \text{med}\left\{ e_{ij}^{(n-1)} \mid j = 1,\ldots,J \right\} ; \; i=1,\ldots,I$$

$$(5b) \quad \Delta m_b^{(n)} = \text{med}\left\{ b_j^{(n-1)} \mid j=1,\ldots,J \right\} ;$$

$$(5c) \quad d_{ij}^{(n)} = e_{ij}^{(n-1)} - \Delta a_i^{(n)} ; \; j=1,..,J; \; i=1,\ldots,I$$

Colonne:

$$(5d) \quad \Delta b_j^{(n)} = \text{med}\left\{ d_{ij}^{(n)} \mid I=1,\ldots I \right\} ; \quad j=1,\ldots J$$

$$(5e) \quad \Delta m_a^{(n)} = \text{med}\left\{ a_i^{(n-1)} + \Delta a_i^{(n)} \mid i=1,\ldots,I \right\} ;$$

$$(5f) \quad e_{ij}^{(n)} = d_{ij}^{(n)} - \Delta b_j^{(n)} ; \; i=1,\ldots,I; \; j=1,\ldots,J$$

Valore globale ed effetti:

$$(5g) \quad m^{(n)} = m^{(n-1)} + \Delta m_a^{(n)} + \Delta m_b^{(n)} ;$$

$$(5h) \quad a_i^{(n)} = a_i^{(n-1)} + \Delta a_i^{(n)} - \Delta m_a^{(n)} ; \; i=1,\ldots,I$$

$$(5i) \quad b_j^{(n)} = b_j^{(n-1)} - \Delta m_b^{(n)} + \Delta b_j^{(n)} ; \; j=1,\ldots,J$$

Le equazioni (5a), (5b), e (5c) descrivono la sottrazione delle mediane per le righe compresa la riga degli effetti colonna e l'aggiornamento dei residui. Le (5d), (5e) e (5f) rappresentano le operazioni corrispondenti per le colonne; le (5g), (5h) e (5i) si riferiscono all'aggiornamento del valore globale e degli effetti riga e colonna. In aggiunta si vuole solo osservare che l'esecuzione della equazione (5b) deve seguire quella del-

le (5i) e che deve essere seguita dall'aggiornamento di $b_j^{(n)}$ e $m^{(n)}$ se vogliamo preservare la centratura a mediana 0 dei $b_j^{(n)}$. Comunque ogni operazione di centratura può essere lasciata all'ultimo passaggio dell'iterazione, come avviene nell'algoritmo presentato da Velleman e Hoaglin [21]. Per dimostrare le proprietà di convergenza dell'algoritmo descritto dalle (5), partiamo dalla considerazione che, come la media è la quantità che minimizza la somma dei quadrati dei residui, così la mediana è un valore che minimizza la somma dei valori assoluti dei residui. Formalmente definita tale somma come

$$S A R = \sum_i \left| x_i - c \right| \quad ,$$

riesce che il valore c che rende minimo SAR è

$$c = \text{med} \left\{ x_1, \ldots x_n \right\} .$$

Se n è un numero pari, qualsiasi valore c compreso nell'intervallo centrale $(x_{(k)}, x_{(k+1)})$ è soluzione del problema di minimo.

Fatta questa premessa, vediamo che il metodo delle mediane,

se si comincia dalle righe, equivale a minimizzare, prima, la somma su j dei residui rispetto a ciascun a_i, tenendo fissate le b_j; e successivamente a minimizzare rispetto a b_j la somma dei residui su i, tenendo fissati gli a_i ai valori precedentemente calcolati.

Formalmente, il primo passo di una iterazione corrisponde, considerando m inglobato in a_i, al criterio

$$(6) \qquad \min_{a_i} \sum_j | y_{ij} - a_i - b_j | \qquad i=1,\ldots,I \quad ; \quad b_j \text{ fissati}$$

e il secondo passo a

$$(7) \qquad \min_{b_j} \sum_i |y_{ij}-a_i-b_j| \qquad j=1,\ldots,J \quad ; \quad a_i \text{ fissati} \quad .$$

Essendosi modificati i valori b_j, tornando alla (6) occorre ricalcolarsi i valori di a_i che la rendono minima. Soltanto nel caso in cui le mediane dei residui sulle righe e sulle colonne siano contemporaneamente nulle il procedimento si ferma. Sulla base di questa prospettiva risulta chiara la proprietà di convergenza del metodo: nel passare da una iterazione al-

l'altra la somma dei valori assoluti dei residui non si incrementa come risulta dell'equivalenza di questa alla successione delle operazioni descritte dalla (7) e dalla (8). La sequenza della somma dei valori assoluti dei residui prodotti dall'iterazione del metodo per sottrazione delle mediane è una sequenza non crescente di valori non negativi. Essa, e con esso il metodo, converge.

L'equivalenza del metodo al criterio descritto dalla (7) e dalla (8) prese in successione ci consente anche un'altra osservazione.

Nel caso di una tabella con I e J numeri dispari, se una riga o una colonna hanno mediana diversa da zero, allora l'ap-

plicazione di una iterazione del metodo delle mediane fa decrescere la somma dei valori assoluti dei residui.

Potremmo, a questo punto chiederci se il metodo delle mediane corrisponda al criterio di minimizzare tale somma; cioè se, definita la somma dei valori assoluti dei residui come

$$S\ A\ R = \sum_{ij} |y_{ij} - m - a_i - b_j| \quad ,$$

l'applicazione successiva delle (7) e (8) equivalga al criterio

$$\min_{a_i, b_j} \sum_{ij} |y_{ij} - m - a_i - b_j| =$$

$$= \min_{a_i b_j} S\ A\ R.$$

Un semplice esempio (A. Siegel) nega questa possibilità.

Data la tabella 3x3

1	6	3
5	9	2
6	4	7

,

il metodo delle mediane, cominciando dalle righe ,dà la seguente decomposizione:

-2	0	0	-2
0	1	-3	0
0	-5	1	1
0	3	0	5

dove la tabella 3x3 dei residui ha ai suoi bordi gli effetti colonna, globale e riga; partendo dalle colonne il metodo dà

-4	0	0	0
0	3	-1	0
0	-3	3	1
0	1	-2	5

La soluzione di minimo SAR è invece

-1	0	0	-3
0	0	-4	0
0	-6	0	1
-1	3	0	6

La soluzione del metodo delle mediane, su cui non è più possibile agire iterativamente, danno somma dei valori assoluti di residui 12 e 14, rispettivamente partendo dalle righe e dalle colonne; il metodo di minimo SAR dà 11.

Un altro aspetto che questo esempio fa balzare agli occhi è la differente soluzione ottenuta dal metodo delle mediane con condizioni di partenza differenti. Dal punto di vista delle applicazioni queste differenze non sembra [10] rivestono grossa importanza, anche se certamente da un punto di vista matematico molte questioni restano aperte. Di alcuni risultati ci occuperemo nel prossimo paragrafo.

Vogliamo terminare il seguente confrontando, con un esempio, il metodo proposto con il metodo di minimi quadrati.

Consideriamo una ipotetica tabella così formata

9	0	0
0	0	0
0	0	0

in cui tutti i valori sono nulli tranne che nella cella(1,1). Un'analisi col metodo dei minimi quadrati porta alla seguente decomposizione

4	-2	-2	2
-2	1	1	-1
-2	1	1	-1
2	-1	-1	1

dove bordiamo i residui con i diversi effetti. Il metodo delle mediane porta a

9	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

L'esempio espone in maniera molto chiara il trasferimento di quantità numerica da una singola cella ai valori stimati e ai residui di tutte le altre celle, quando si utilizza uno stimatore come la media.

Se una sola cella contiene un dato anomalo, ciò influenza negativamente tutta l'analisi. Nel nostro caso possiamo ancora individuare nella cella (1,1) un dato anomalo. Se siamo in presenza di due o tre dati anomali in tabelle più grandi, alcuni sulla stessa riga o colonne, può diventare molto difficoltoso riuscire ad individuarli (Daniel C. 1978).

5.3 Il metodo delle mediane: alcune proprietà

In questo paragrafo riportiamo alcuni risultati relativi ad una tabella 3x3 e i valori del punto di rottura del metodo delle mediane.

Si è visto nel paragrafo precedente che il metodo delle mediane non coincide con il metodo che minimizza la SAR.

Tuttavia è stato dimostrato (A. Siegel) che, in una tavola 3x3, ogni volta che il primo passo di una iterazione produce una intera riga o colonna di zeri, il metodo delle mediane converge in una iterazione e un passo, al massimo, alla soluzione del minimo valore assoluto per i residui.

Prima di ricavare delle conseguenze da questo risultato, premettiamo le seguenti osservazioni.

Supponiamo che $Y_{ij} = A_i + B_j$

è esattamente la somma di un effetto riga e di un effetto colonna. Allora le possibili tavole di residui, e^*_{ij}, per

$$Y_{ij} + y_{ij} = m^* + a^*_i + b^*_j + e^*_{ij}$$

sono esattamente le tavole dei residui possibili per

$$y_{ij} = m + a_i + b_j + e_{ij}.$$

E quindi y_{ij} e $y_{ij}+Y_{ij}$ hanno lo stesso valore minimo per SAR.

Dalla proposizione dimostrata di A. Siegel, e dalla precedente osservazione ne segue che :

1) data una tavola 3x3 y_{ij}, se poniamo

$$Y_{ij} = \begin{pmatrix} x & x & x \\ 0 & 0 & 0 \\ -x & -x & -x \end{pmatrix}$$

allora, per un valore sufficientemente elevato di x, il primo passo di una iterazione, che parta dalle colonne, applicato a $(y_{ij}+Y_{ij})$ soddisfa le ipotesi della proposizione di Siegel. Così gli (e^*_{ij}) raggiungono il valore minimo di SAR e la corrispondente analisi per (y_{ij}) sarà una soluzione con minimo valore di SAR.

2) se da una tavola 3x3 sottraiamo una matrice y tale da rendere nulla una sua riga e da lasciare inalterate le restanti, allora il metodo delle mediane che parte dalle righe porta al-

la soluzione di minimo SAR.

Questi risultati danno la possibilità di intervenire sulla tabelle 3x3 di partenza in modo che la soluzione cui si perviene tramite il metodo delle mediane è di minimo SAR.

Essi non possono essere estesi a tabelle 3x4 e si sta tentando di estenderle a tabelle 3x5.

Prima di valutare il valore del punto di rottura del metodo delle mediane, ricordiamo tale valore per la mediana; esso è

$$(8) \qquad \frac{1}{2} - \frac{2 - d(n)}{2n}$$

dove n è il numero dei dati e d(n) è la funzione parità.

Questa quantità esprime la frazione massima di dati che in un insieme di n valori può essere sostituito senza restrizione in modo che la variazione del valore della mediana resti limitata. Tale valore fa riferimento al caso in cui i dati sostituiti si distribuiscano nel modo più sfavorevole possibile. Potremmo perciò pensare, all'opposto, di definire un punto di rottura nel caso di distribuzione favorevole dei valori sostituiti.

Nel caso della mediana esso è esattamente il doppio del valore della (8)

$$1 - \frac{2-d(n)}{n} .$$

Nel caso del metodo delle mediane, il valore del punto di rottura è determinato dal numero massimo di osservazioni che, sostituiti nella tabella, non determinano grosse variazioni nei valori dei parametri stimati, m, a_i e b_j e che perciò si ripercuotano essenzialmente sul valore dei residui e_{ij}.

Nel caso di distribuzione sfavorevole di dati anomali all'interno della tabella, essi si andranno a collocare sulla stessa riga o sulla stessa colonna. Nel caso che I sia maggiore di J, la situazione più sfavorevole si precisa come quella in cui tutte le osservazioni cattive cadano tra le J osservazioni di una singola riga.

Un numero eccessivo di dati cattivi sulla riga, farà variare in modo sensibile la mediana di quella riga inficiando il proseguo della analisi. Perciò il punto di rottura nell'ipotesi di distribuzione degli errori la più sfavorevole, PRS, sarà

il punto di rottura per la mediana di un gruppo di dati la cui dimensione è la minore fra i numeri I e J.

$$(9) \qquad PRS = \frac{\frac{1}{2}\min(I,J) + \frac{1}{2}\,d[\min(I,J)] - 1}{IJ}$$

$$= \frac{1}{2\max(I,J)} - \frac{2 - d[\min(I,J)]}{2IJ}$$

Nel caso di distribuzione favorevole delle osservazioni cattive all'interno della tavola vale la seguente proposizione.

Il punto di rottura, nel caso di distribuzione favorevole degli errori nella tavola, del metodo delle mediane per una tavola I x J è B/IxJ dove

$$(10) \quad B = \frac{1}{2}\min\left\{J\,[\,I-2+d(I)\,],\, I\,[\,J-2+d(J)\,]\right\} =$$

$$= \frac{1}{2}IJ - \frac{1}{2}\max\left\{2J-Jd(I),\ 2I-Id(J)\right\}.$$

Se $I \leq J$ allora B è dato dalle formule

1) per I dispari, J pari e $J < 2I$

$$(11) \qquad B = \frac{1}{2}IJ - I$$

2) negli altri casi

$$(12) \quad B = \frac{1}{2}IJ - \frac{1}{2}J\,[\,2-d(I)\,] = \begin{cases} \frac{1}{2}IJ - J & \text{per I dispari} \\ \frac{1}{2}(IJ-J) & \text{per I pari} \end{cases}$$

Cosicchè $B \geq \frac{1}{2} IJ - \max(I,J)$ e

$$B/IJ \geq \frac{1}{2} - \frac{1}{\min(I,J)} .$$

Dimostriamo innanzitutto che B non può superare il valore alla destra nell'espressione (10).

Dalla (8) ricaviamo che $\frac{1}{2} I - 1 + \frac{1}{2} d(I)$ è il valore più alto di osservazioni cattive tollerabili in ciascuna colonna. Allo stesso modo, $\frac{1}{2} J - 1 + \frac{1}{2} d(J)$ è il numero massimo di osservazioni cattive tollerabile in ciascuna riga. Il numero ammissibile di cattive osservazioni nella intera tabella non può eccedere nè $J\left[\frac{1}{2} I - 1 + \frac{1}{2} d(I)\right]$ nè $I\left[\frac{1}{2} J - 1 + \frac{1}{2} d(J)\right]$, cosicchè B non può eccedere il valore minimo fra i due. Per dimostrare completamente la proposizione, ci rimane da far vedere che esiste una configurazione dei valori cattivi per cui B assume proprio il valore a destra della (10).

Ragioniamo per assurdo; supponiamo di avere di fronte a noi la configurazione con B massimo e che B sia strettamente inferiore al valore nella destra della (10). Senza perdere di

generalità assumiamo che la prima riga abbia un numero di valori cattivi minore di $\frac{1}{2}\,J-1+\frac{1}{2}\,d(J)$ e la prima colonna in numero inferiore a $\frac{1}{2}\,I-1+\frac{1}{2}d(I)$. Se la cella (1,1) non contiene già un valore cattivo, possiamo sostituirlo senza che l'analisi ne sia inficiata, contraddicendo l'ipotesi di massimo per B. La cella (1,1) conterrà allora un valore cattivo. Permutando opportunamente righe e colonne possiamo assumere che i rimanenti elementi cattivi della colonna 1 e rispettivamente della riga 1 siano nelle prime posizioni.

Avremo perciò R elementi cattivi nelle prime R posizioni della colonna 1 e C elementi cattivi nelle prime C posizioni della riga 1. Dalle nostra assunzione deriva chiaramente che

$R<\frac{1}{2}\,I-1+\frac{1}{2}\,d(I)$ e $C<\frac{1}{2}\,J-1+\frac{1}{2}\,d(J)$. Definiamo ora Q la parte della tavola di (I-R) righe e (J-C) colonne posta nell'angolo in basso a destra. Se Q contenesse solo valori buoni, l'ipotesi di massimo per B sarebbe violata, in quanto qualsiasi elemento di Q potrebbe essere sostituito con un valore cattivo senza eccedere il numero di $\frac{1}{2}\,J-1+\frac{1}{2}\,d(J)$ elementi cattivi nel-

le sue righe e di $\frac{1}{2}$ I-1+$\frac{1}{2}$ d(I) elementi cattivi nelle sue colonne. Perciò Q contiene almeno un elemento cattivo, che supponiamo che sia nella cella (i,j). Quindi scambiamo il valore cattivo in (i,j) con il valore buono in (i,1) e il valore cattivo in (1,1) con il valore buono (1,j). Se sostituiamo ora il valore buono nella cella (1,1), incrementiamo il valore di B a B+1 senza che ciò porti alla rottura del metodo. Questo contraddice l'ipotesi che B sia massimo e porta alla dimostrazione della proposizione. Nel caso di una tabella 5x8 il punto di rottura è in base alla(9)2/40. Nella configurazione più favorevole possibile il valore del punto di rottura è dalla (10) pari 15/40. Queste due quantità ci danno una idea più precisa dei limiti di tollerabilità del metodo a valori anomali, che non il solo valore del punto di rottura definito per la configurazione più sfavorevole.

5.4 Trasformazione dei dati

Per verificare la validità della struttura additiva ipotizzata nel modello (1) e, allo stesso tempo, per avere una valutazione della trasformazione di potenza cui sottoporre i dati per rimuovere la non additività, eventualmente presente nei dati, ci serviamo del grafico costituito dai punti aventi come ascissa

x i "valori di confronto" definiti come

$$cv_{ij} = \frac{a_i x b_j}{m},$$

dove m, a_i e b_j sono il valore globale, gli effetti riga e colonna stimati sui dati grezzi,

e come ordinata

y i valori dei residui e_{ij} ottenuti sottraendo dei dati grezzi y_{ij} il modello $m+a_i+b_j$.

Se tale grafico non presenta un significativo andamento li-

neare, possiamo concludere che i dati non si discostano in modo consistente da un modello additivo. Altrimenti, le pendenze della retta, su cui si dispongono le coppie di valori (cv_{ij}, e_{ij}), ci suggerisce la trasformazione cui sottoporre i dati per rimuovere la non additività. Se tale pendenza è pari a b, una potenza vicina a 1-b può servire allo scopo. (In effetti, in questo ultimo caso, potrebbe accadere che la potenza 1-b applicata ai dati non appaia nè plausibile nè appropriata, nel qual caso potrebbe essere utile considerare di aggiungere al modello il termine $bx cv_{ij}$.)

Per giustificare l'uso del grafico su descritto ai fini che ci siamo proposti, ragioniamo al modo seguente. Supponiamo che esista un valore p tale il modello additivo valga esattamente cioè che

$$y_{ij}^{p} = m+a_i+b_j$$

in modo che

$$y_{ij} = (m+a_i+b_j)^{1/P} \qquad (13)$$

Notiamo che i valori di a_i e b_j non hanno qui i valori otte-

nuti interpolando direttamente i dati grezzi; si fa inoltre osservare che supponiamo di lavorare con quantità positive.

Per confrontare il modello (13) con il modello additivo semplice, usiamo una approssimazione al secondo ordine della (13). Riscrivendola come

$$y_{ij} = m^{1/P} \left(1 + \frac{a_i}{m} + \frac{b_j}{m}\right)^{1/P} ,$$

espandendo il secondo fattore $(1+t)^{1/P}$ in serie di Taylor fino al 2° termine rispetto a $t = \frac{a}{m} i + \frac{b}{m} j$, si ottiene

$$(14) \qquad y_{ij} \simeq m^{1/P} \left[1 + \frac{1}{P} \left(\frac{a_i}{m} + \frac{b_j}{m} \right) + \frac{1-P}{2P^2} \left(\frac{a_i}{m} + \frac{b_j}{m} \right)^2 \right] .$$

Raggruppando opportunamente i termini abbiamo

$$(15) \qquad y_{ij} \simeq m^{1/P} \left[1 + \left(\frac{1}{P} \frac{a_i}{m} + \frac{1-P}{2P^2} \frac{a_i^2}{m^2} \right) + \right.$$
$$\left. + \left(\frac{1}{P} \frac{b_j}{m} + \frac{1-P}{2P^2} \frac{b_j}{m^2} \right) + \left(\frac{1-P}{2P^2} \frac{2 a_i b_j}{m^2} \right) \right] .$$

Per semplificare la notazione poniamo

$$(16a) \qquad D = m^{1/P}$$

$$(16b) \qquad \frac{A_i}{D} = \frac{1}{P} \frac{a_i}{m} + \frac{1-P}{2P^2} \frac{a_i^2}{m^2}$$

$$(16c) \quad \frac{B_j}{D} = \frac{1}{P}\frac{b_j}{m} + \frac{1-P}{2P^2}\frac{b_j^2}{m^2}$$

$$(16d) \quad \frac{C_{ij}}{D} = \frac{1-P}{P^2}\frac{a_i b_j}{m^2} \; .$$

Cosicchè la 15 diventa

$$(17) \quad y_{ij} \simeq D + A_i + B_j + C_{ij} \; .$$

Se osserviamo che, sempre al secondo ordine,

$$\frac{A_i B_j}{D \times D} \simeq \frac{1}{p2}\frac{a_i}{m}\frac{b_j}{m}$$

la (16d) può essere approssimata da

$$\frac{C_{ij}}{D} \simeq (1-P)\frac{A_j}{D}\frac{B_j}{D}$$

e la (17), utilizzando questa approssimazione, diventa

$$(18) \quad y_{ij} \simeq D + A_i + B_j + (1-p)\frac{A_i B_j}{D} \; .$$

Concludendo questa prima parte del ragionamento, diciamo che se y_{ij}^P è approssimato da un modello additivo, allora y_{ij}, in una approssimazione al 2° ordine è data dalla equazione (18).

Di converso, se y_{ij} è dato approssimativamente dalla (18), allora per y^{P}_{ij} vale approssimativamente il modello

$$y^{P}_{ij} \simeq m+a_i+b_j.$$

Nell'ipotesi che valga la (18), i residui che otteniamo interpolando i dati grezzi col modello additivo dovrebbero all'incirca valere

$$R_{ij} = (1-p) \frac{A_i B_j}{D},$$

dove le quantità D, A_i e B_j sono l'effetto globale, il fattore riga e colonna ottenuti con la stessa interpolazione.

Rappresentando su in grafico i punti $(R_{ij}, \frac{A_i B_j}{D})$, essi si disporranno all'incirca su una retta di pendenza (1-p).

In pratica, si scambia il ragionamento e dalla presenza di un andamento lineare nel grafico con pendenza 1-p, ricaviamo che una trasformazione di potenza p applicata ai dati porterà a una struttura additiva nella tabella.

5.5 Un esempio

Dalla pubblicazione ISTAT, Collana di informazioni, Recenti livelli e caratteristiche della mortalità infantile in Italia, Anno VII, n° 4, 1983, ricaviamo la seguente tabella a doppia entrata relativa ai quozienti di mortalità infantile per gli anni '74-'81 nelle 5 grandi ripartizioni territoriali :

	'74	'75	'76	'77	'78	'79	'80	'81
Italia								
Nord occidentale	20.4	18.6	17.4	16.0	16.2	13.4	12.2	12.9
Nord orientale	18.0	17.2	15.3	14.2	13.4	12.0	11.2	11.0
Centrale	19.3	16.8	16.3	15.3	14.4	13.4	12.3	12.2
Meridionale	28.6	26.8	22.8	21.8	20.8	19.0	16.8	16.7
Insulare	26.0	25.3	22.6	20.7	17.7	18.7	17.7	16.6

Utilizzando un programma in Basic fornito da [21] otteniamo per 2 iterazioni complete del metodo di sottrazione delle mediane, partendo dalle righe, la seguente decomposizione:

0	-.52	.12	-.12	.92	-.52	-.82	0	0
-.47	0	-.05	0	.05	0	.1	.02	-1.92
-.22	-1.4	-.1	.05	0	.34	.15	.17	-.87
2.67	2.15	0	.15	0	-.45	-.1.75	-1.72	5.52
.92	1.5	.55	-.1	-2.25	.1	0	-.97	4.67
4.70	3.42	1.57	.42	-.42	-.1.77	-2.67	-2.80	15.70

Osserviamo che le colonne dei residui hanno già raggiunto mediana nulla e generalmente molto vicino allo zero è anche quella delle righe. Mediane nulle presentano anche gli effetti riga e colonna.

Calcolati i valori di confronto e posti su un grafico i punti con ascisse i valori di confronto e ordinate i residui della decomposizione ottenuta, non risulta un omogeneo andamento lineare, tale da far pensare a una trasformazione dei dati.

C'è comunque da osservare per la ripartizione meridionale un andamento decrescente abbastanza netto e tale da far pensare a una non completa risoluzione della struttura.

6. CONCLUSIONI

Ripercorrendo brevemente le varie parti di questo lavoro si può constatare la presenza di quattro temi fondamentali che ne sono alla base

1) il problema della resistenza delle stime
2) l'analisi dei residui
3) la riespressione dei dati tramite trasformazioni monotone.
4) l'utilizzazione di metodi grafici per rilevare direttamente le caratteristiche dei dati.

Questi temi li ritroviamo riversati nelle varie tecniche di investigazione dei dati proposte nell'ambito dell'EDA:

1) il diagramma "albero e foglia"
2) il "diagramma a scatola"
3) i "valori lettera,"

per quanto riguarda l'analisi e il confronto di gruppi di dati non strutturati,

4) la tecnica della linea resistente

5) il metodo per sottrazione delle mediane

per quanto riguarda dati che si ripresentano in coppie di valori e in una tabella a doppia entrata.

Tali tecniche, per essere correttamente intese, vanno valutate rispetto alla logica di tipo diagnostico ed indiziario che l'EDA si propone.

Esse non presumono di corrispondere a criteri di ottimalità e di univocità, come, ad esempio, il metodo delle mediane nell'analisi di una tabella a doppia entrata; tentano costruzione di modelli che rispecchino più fedelmente possibile l'andamento dei dati. Nell'esempio specifico, la non univocità della soluzione, ai fini di un'analisi preliminare di tipo investigativo, non costituisce un problema grave, tenendo conto che le differenze fra le soluzioni non sembrano tali da inficiare l'analisi.

Che poi il terreno sia aperto e allo approfondimento delle tecniche già formulate che alle formulazioni di nuo-

vi strumenti d'indagine è un opinione riscontrabile nelle posizioni dei vari autori che ad esse hanno maggiormente contribuito.

L'aver aperto la statistica al problema di come aiutare le scienze empiriche ad acquisire nuove cognizioni e quindi al problema delle definizioni di metodi atti a scoprire nuovi fenomeni, evitando il limite di preoccuparsi solo del controllo di ipotesi pre-formulate è, il merito ascrivibile all'Analisi Esplorativa di Dati.

Da un'altra prospettiva c'è da rilevare l'influenza che l'EDA ha esercitato sul filone più prettamente metodologico della determinazione di nuove classi di stimatori con proprietà nuove rispetto a quelle classiche. Ci riferiamo,in particolare, alla proprietà di robustezza [11], che in termini generali può essere definita come l'invarianza di proprietà ottimali dello stimatore rispetto a variazioni limitate della distribuzione ipotizzata. In questi studi, accanto ad analisi di natura prettamente matematica, si associano analisi di simu-

lazione per la verifica delle proprietà nel caso di piccoli campioni.

NOTA SUI PROGRAMMI UTILIZZATI

Rimandando al testo di Vellemann e Hoaglin [21], per una visione completa dei programmi utilizzati per gli esempi, qui vogliamo brevemente sottolinearne alcuni aspetti.

Nel programma per l' "albero e foglia" viene scelto un numero di righe prossimo a $10 \log_{10} n$, dove n è il numero dei dati da rappresentare.

Per una visione critica della scelta e di altre possibilità si può vedere anche [10]. Un interessante punto di confronto è la scelta del numero delle classi per l'istogramma basato sul criterio di minimo errore quadratico per la densità stimata rispetto alla teorica [13].

Nel programma per il "diagramma a scatola" c'è la possibilità di scegliere una rappresentazione che, accanto ai quarti, segnala i punti posti a una distanza $1.58 d_F / \sqrt{n}$ simmetricamente rispetto alla mediana. Questi consentono di effettuare, per due gruppi di dati, una analisi della varian-

za a una via. Se, ad esempio, gli intervalli determinati da questi punti e relativi a due gruppi di dati non si sovrappongono, possiamo dire che a un livello di confidenza del 95% le mediane delle popolazioni sono differenti. Sulla determinazione dell'ampiezza di tale intervallo si può vedere [21].

Il programma che calcola i parametri della linea resistente utilizza l'algoritmo descritto dalla (1) nel paragrafo 4.2. C'è da sottolineare che il programma inizia con la determinazione dei tre gruppi, assicurando che i punti con uguale valore di ascissa cadano nella medesima regione e che i gruppi esterni contengano almeno 3 punti. Nel caso che ciò non sia possibile la stima viene effettuata raggruppando i dati in due sole regioni. Se neanche ciò è possibile il programma si ferma per errore.

Per il metodo delle mediane nella analisi di una tabella a doppia entrata, il relativo programma prevede una versione standard con 2 iterazioni complete (4 passi) con partenza dalle righe. E' possibile stabilire, però, in modo arbitra-

rio il numero dei singoli passi e il tipo di partenza (righe o colonne). Il programma calcola anche i valori di confronto.

Tutti i programmi vengono presentati sia in linguaggio Fortran che Basic. Tuttavia c'è da sottolineare che alcuni di essi non sono direttamente utilizzabili, a causa della presenza di errori in alcune istruzioni così come riportate.

"QUADERNI DI DISCUSSIONE PUBBLICATI"

84.01 REY G. M.
Le statistiche ufficiali e l'attività della Pubblica Amministrazione
Giugno 1984

85.01 CRESCENZI F.
Nota su alcune metodologie per la classificazione di unità territoriali
Febbraio 1985

85.02 CORTESE A.
Alcune considerazioni sulle prospettive del censimento della popolazione
Marzo 1985

85.03 MATURANI G.
Stima delle ore di lavoro effettivamente prestate dai lavoratori occupati negli anni 1960-1983
Aprile 1985

85.04 NAPOLITANO P.
Esposizione di alcune tecniche per la investigazione dei dati

B I B L I O G R A F I A

(1) ANDREWS D.F. (1974) "A robust method for multiple linear Regression", Technometrics, vol. 16, 523-531.

(2) BLOM G. (1958) "Statistical estimates and transformed beta-variables". New York: Wiley.

(3) BODE H., MOSTELLER F., TUKEY J.W., WINSOR (1949) "The education of a scientific generalist" Science, 109, pag. 553.

(4) BOX, G.E.P. (1980) "Sampling Bayes' inference in scientific modelling and robustness" with discussion, Journal of the Royal Statistical Society, Series A, 143, 383-430.

(5) CLEVELAND W.S., GUARINO R. (1976) "Some robust statistical procedures and their application to air pollution data", Technometrics, vol. 18, 401-409

(6) DAJOZ, R.C.(1972)Manuale di Ecologia; ISEDI

(7) GOWER J.C., e ROSS G.J.S., (1969), "Minimum spanning

trees and single linkage cluster analysis, Appl. Stat. 18, 54-64

(8) HAMPEL, F.R. (1968) "Contributions to the theory of Robust Estimation". Ph . D. Thesis, Department of Statistics, University of California, Berkeley.

(9) HAMPEL, F.R. (1971) "A general qualitative definition of robustness", Annals of Mathematical Statistics, 42, 1887-1896.

(10) HOAGLIN, D.C., MOSTELLER, F., TUKEY T.W., (editors) (1983) Understanding robust and exploratory data analysis; Wiley

(11) HUBER P.J., (1981) Robust statistics, Wiley

(12) MORINEAU A. (1978) "Règression robustes. Mèthods d'ajustement et de validation". Reveu de Statistique Appliquè, 1978, vol. XXVI, n° 3

(13) SCOTT, D.W. (1979) "On optimal and data-based histograms", Biometrika, 66, 605-610.

(14) SIEGEL, H.F. (1982) "Robust regression using repeated

medianes," Biometrika, 69, 242-244

(15) TUKEY J.W. (1949) "One degree of freedom for non-additivity", Biometrics, 5, 232-242.

(16) _____ (1957) "On the comparative anatomy of trasformations", Annals of Mathematical Statistics, 28, 602-632

(17) _____ (1960) "A survey of sampling from contaminated distributions" in Contributions to Probability and Statistics; I. Olkin, Stanford University Press.

(18) _____ (1972) "Data analysis, computation and mathematics", Quarterly of Applied Mathematics, April, 1972, 51-65

(19) _____ (1977) Exploratory data analysis; Reading MA: Addison.-Wesley

(20) _____ (1980) "We need both exploratory and confirmatory ", the American Statistician, 34, 23-25

(21) VELLEMAN P.F. e HOAGLIN, D.C. (1981) Applications, Basics and Computing of Exploratory Data Analysis, Boston MA: Duxburg Press

(22) WAINER, H. e THISSEN, D. (1981) "Graphical data analysis", Annual Review of Psychology, 32, 191-241.