L'analisi dei dati in biblioteca

Come un test statistico può contribuire a individuare una corretta strategia gestionale

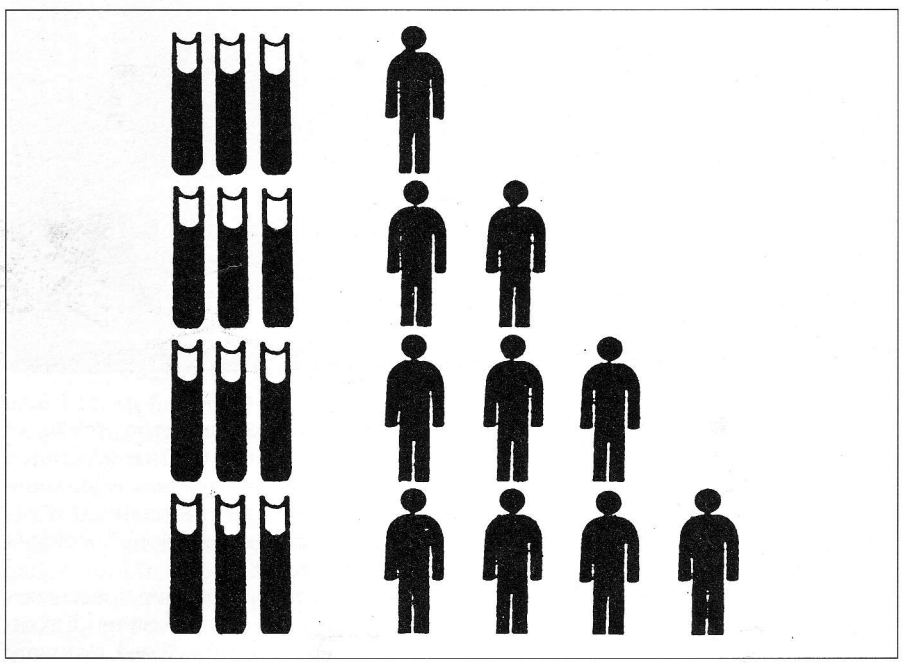
di Mario Sebastiani

La biblioteca, considerata come un sistema, è caratterizzata da molteplici variabili di tipo quantitativo che consentono di descrivere il suo funzionamento in termini statistici e matematici rigorosi. Collezionando i dati relativi ai prestiti, agli utenti, al patrimonio, si può descrivere in maniera precisa il funzionamento effettivo di una biblioteca.
Questi dati sono importanti anche al fine della corretta gestione della biblioteca e questa è indubbiamente una considerazione banale. Meno banale invece è l'interrogativo su *come* questi dati possono essere utili alla gestione della biblioteca. Per rispondere dobbiamo chiederci cosa significa *gestire*. In un certo senso gestire significa risolvere problemi e per risolvere i problemi occorrono *risposte*, cioè *informazioni utili*. Ma un qualsiasi dato numerico, statistico o meno, non costituisce di per sé un'informazione: occorre prima metterlo in relazione con il problema che dobbiamo risolvere, stabilire la sua rilevanza, confrontarlo con altri dati: occorre, in altre parole, elaborare il dato.

Il primo articolo della serie dedicata alla "Ricerca operativa" è comparso sul numero di novembre dello scorso anno (p. 40-45).

1. Dai dati statistici alle informazioni

Un problema di gestione è sostanzialmente un interrogativo relativo a quale, tra diverse opzioni possibili, è la scelta migliore da perseguire. I dati grezzi non possono fornire indicazioni in merito. Sono i dati sottoposti ad opportune elaborazioni quelli che forniscono le informazioni utili a selezionare l'opzione migliore. Solo disponendo di opportune procedure possiamo estrapolare dai dati statistici le informazioni implicite in essi.
Per gestire una biblioteca quindi, così come ogni altro sistema, occorre applicare ai dati statistici rilevanti opportune procedure di *analisi dei dati*.
Ma quali sono queste procedure? In che modo dobbiamo analizzare i dati statistici al fine di gestire il sistema-biblioteca? Per questo occorre, ne abbiamo già parlato in un precedente articolo, una scienza adeguata.
Questa scienza esiste ed è conosciuta con vari nomi: analisi dei sistemi, scienza dell'organizzazione, ecc. In realtà si tratta di un corpus di tecniche matematiche che consentono di ampliare e migliorare il controllo del sistema complessivo e la sua gestione. Il



termine che si è andato pian piano affermando, nell'indicare questo insieme di tecniche, è *ricerca operativa* (in inglese *operation research*).
Le origini di questa disciplina datano a circa 50 anni fa. Durante la seconda guerra mondiale i pianificatori militari angloamericani dovettero affrontare, tra i molti, anche questo drammatico dilemma: al fine di minimizzare la quantità di navi affondate dai sottomarini tedeschi nell'Oceano atlantico, conveniva che le navi che trasportavano viveri e rifornimenti viaggiassero quanto più possibile sparpagliate sulle superficie del mare, oppure che venissero concentrate in grossi assembramenti denominati convogli?
Quella che si affermò e che risultò vincente, come è noto, fu la seconda soluzione. Questo fu anche conseguenza dell'utilizzazione, da parte dei pianificatori militari, di alcune di quelle tecniche matematiche ricomprese nella ricerca operativa. Comunemente si identifica in questo episodio la nascita e la prima applicazione della ricerca operativa. In seguito questa disciplina ha conosciuto una grande diffusione non solo in campo militare ma anche in campo industriale e, seppur in modo più limitato, anche in campo sociale.
La ricerca operativa nasce come scienza fortemente connessa alle discipline statistiche e al calcolo delle probabilità. Anzi si può dire che la ricerca operativa coincide in buona misura con la statistica, o meglio con quella parte della statistica nota come statistica non parametrica. Nella statistica parametrica ci si occupa principalmente delle variabili quantitative che caratterizzano gli individui appartenenti ad una data popolazione: peso, altezza, età, ecc. A partire da queste variabili quantitative vengono calcolati alcuni parametri che descrivono, nel suo insieme, le caratteristiche quantitative essenziali della popolazione. I parametri statistici fondamentali sono: il valor medio, lo scarto quadratico medio, l'andamento della distribuzione (cioè la ripartizione delle variabili quantitative all'interno della popolazione). Per affrontare tanto la statistica che la ricerca operativa è auspicabile una certa dimestichezza — anche solo a livello concettuale — con queste grandezze.[1]
La statistica non parametrica invece non si occupa delle variabili quantitative degli individui di una popolazione bensì dei loro attributi: sesso, colore dei capelli, caratteristiche sociali, lavorative, ecc. Se nella statistica parametrica il fine è quello di descrivere, nei suoi aspetti quantitativi essenziali, una data popolazione di individui, nella statistica non parametrica, invece, l'obiettivo è quello di individuare relazioni tra gli attributi di una popolazione e misurare l'intensità di questa relazione.
Ad esempio: c'è un rapporto tra il colore dei capelli e la scelta del lavoro? Oppure: c'è relazione tra l'età ed il consumo di hamburger? Sono tutte domande che tendono ad ipotizzare una relazione tra gli attributi di una popolazione. La statistica non parametrica include strumenti per poter affermare se la relazione sussiste, non sussiste, quanto è forte.
Tra i test più comuni di cui ci si avvale in questo genere di elaborazioni vi è il cosiddetto test del chi-quadrato (rappresentato dalla lettera greca χ^2) inventato nel 1900 dallo statistico Karl Pearson.

2. Test del chi-quadrato e coefficienti di contingenza

Il test del chi-quadrato fornisce, essenzialmente, una misura della *distanza tra gli attributi* degli individui: se questa distanza è inferiore ad una certa soglia critica possiamo accettare l'ipotesi nulla, vale a dire l'ipotesi che gli attributi siano indipendenti tra loro. Accettare l'ipotesi nulla significa affermare che non c'è correlazione tra il modo in cui un dato attributo si ripartisce tra i membri di una data popolazione e il modo in cui si ripartiscono gli altri attributi.
Nel caso invece che la distanza risulti superiore ad una certa soglia critica, allora potremo rigettare l'ipotesi nulla e, contestualmente, accettare l'ipotesi alternativa, vale a dire l'ipotesi che gli attributi siano tra loro dipendenti. Accettare l'ipotesi alternativa significa affermare che c'è correlazione tra gli attributi ovvero che la ripartizione di un dato attributo tra gli individui dipende, in una certa misura, dalla ripartizione degli altri attributi. Quanto più la distanza supera la soglia critica, tanto più forte è questa dipendenza.
Come calcolare il chi-quadrato? L'essenza di questo test consiste nel mettere a confronto le frequenze osservate degli attributi degli individui di una data popolazione con le frequenze teo- ➤

Tab. 1 - Tavola di contingenza di 2 righe per 3 colonne

	Agricoltura	Industria	Servizi	Totali
Maschi	2.000	6.000	6.000	14.000
Femmine	1.100	1.800	3.200	6.100
Totali	3.100	7.800	9.200	20.100

riche che si dovrebbero avere nel caso che tra gli attributi non vi sia alcuna relazione ovvero che gli attributi siano indipendenti tra loro. Dalla teoria statistica però si possono ricavare modalità che implicano solo il calcolo dei dati osservati.

Queste modalità, nel caso di *tavole di contingenza* di 2 righe per n colonne (con n maggiore o uguale a 3), sono abbastanza semplici. Per tavola di contingenza si intende la tabella nella quale sono riportati, ordinati per righe e colonne, i dati numerici relativi agli attributi considerati.

Ad esempio la Tab. 1 (p. 33) — che indica la ripartizione di una data popolazione tra le aree lavorative fondamentali — è una tavola di contingenza di 2 righe per 3 colonne. Gli attributi riportati in questa tavola sono 5: maschio, femmina, agricoltore, addetto all'industria, addetto ai servizi. La tavola è completata con i totali di riga e di colonna.

Nella Tab. 2 è riportata la rappresentazione generale di una tavola di contingenza di 2 righe per 3 colonne (completa dei totali). Nella Tab. 3 è riportata una formula che consente di calcolare il valore del chi-quadrato in tavole di contingenza di 2 righe per 3 colonne (naturalmente per motivi di spazio tralasciamo ogni questione attinente la dimostrazione di questa formula).[2]

Per calcolare il chi-quadrato di tavole di contingenza di 2 righe per n colonne (con n maggiore di 3) è sufficiente generalizzare la formula della Tab. 3 (estendendo le somme tra parentesi quadre anziché fino a 3, fino al valore n). Naturalmente una tavola di n righe per 2 colonne può essere facilmente convertita e rappresentata sotto forma di tavole di 2 righe per n colonne.

Tab. 2 - Rappresentazione generale di una tavola di contingenza di 2 righe per 3 colonne

	Attr. 1	Attr. 2	Attr. 3	Totali
Attr. a	a_1	a_2	a_3	N_a
Attr. b	b_1	b_2	b_3	N_b
Totali	N_1	N_2	N_3	N

Per cui la formula di Tab. 3 (generalizzata al valore n, cioè proseguendo le somme tra parentesi quadra sino al valore a_n/N_n e b_n/N_n) può applicarsi, previa conversione della tavola, anche a tavole di n righe per 2 colonne.

Una volta calcolato il chi-quadrato occorre valutare se questo costituisce un'informazione a favore oppure contro l'ipotesi nulla. Per valutare il chi-quadrato occorre confrontarlo con determinati valori critici che devono essere individuati in funzione dei *gradi di libertà* del test (valore che dipende dal numero di attributi) e dal *grado di precisione* del test (valore che indica la probabilità di errore nel caso che si respinga l'ipotesi nulla). I buoni manuali statistici riportano in appendice una tabella in base alla quale è possibile individuare — in funzione dei gradi di libertà e del grado di precisione — i valori critici per il test del chi-quadrato.

Tab. 3 - Calcolo del chi-quadrato per una tavola di contingenza di 2 righe per 3 colonne

$$\chi^2 = \frac{N}{N_a}\left[\frac{a_1}{N_1} + \frac{a_2}{N_2} + \frac{a_3}{N_3}\right] + \frac{N}{N_b}\left[\frac{b_1}{N_1} + \frac{b_2}{N_2} + \frac{b_3}{N_3}\right] - N$$

Tab. 4 - Risultato del chi-quadrato[1]

$\chi^2 = 319{,}362$

[1] Per il calcolo è stata utilizzata la formula di Tab. 3 applicata alla tavola di contingenza in Tab. 1.

I gradi di libertà di una tavola di k righe e n colonne si calcolano con una formula abbastanza semplice a condizione però di conoscere i parametri necessari per stimare le frequenze teoriche della popolazione (cioè le frequenze che si avrebbero nel caso che tutti gli attributi fossero effettivamente indipendenti). Ma questo non sempre è possibile.

Esistono tuttavia modalità per la valutazione del chi-quadrato che consentono di prescindere dall'individuazione dei valori critici. È possibile calcolare, il cosiddetto *coefficiente di contingenza*. Questo parametro varia tra 0 e 1 e si può calcolare, a partire dal chi-quadrato, nel modo indicato nella Tab. 5. Nella Tab. 6 è riportato il valore del coefficiente di contingenza per la tavola (cfr. Tab. 1).

Tab. 5 - Coefficiente di contingenza

$$C = \frac{\chi^2}{\chi^2 + N}$$

Tab. 6 - Coefficiente di contingenza del risultato[1]

C = 0,1250606

[1] Il calcolo si riferisce ai dati di Tab. 4.

Al fine di valutare il grado di associazione degli attributi basta tener presente che quanto più è grande il coefficiente di contingenza, tanto maggiore è l'associazione tra loro degli attributi.[3]

3. Un'applicazione del coefficiente di contingenza

Per elaborare un esempio di applicazione del coefficiente di contingenza nel campo delle biblioteche occorrono naturalmente dati statistici rilevanti. Uno dei pochi punti di riferimento, a questo riguardo, è rappresentato dalle statistiche sulle biblioteche pubbliche statali pubblicate ogni anno sul "Notiziario dell'Ufficio studi del Ministero per i beni culturali e ambientali".[4]

Abbiamo utilizzato queste statistiche e il coefficiente di contingenza per dare una risposta (approssimata) a questo interrogativo: l'*utenza* di una biblioteca è maggiormente correlata al *patrimonio* della biblioteca oppure alle dimensioni del *bacino di utenza* (almeno di quello potenziale)? Una biblioteca con un ricco patrimonio, è chiaro, avrà più utenti di una biblioteca di dimensioni ridotte mentre una biblioteca collocata in un vasto bacino di utenza (considerando la provincia come unità territoriale di misura) avrà sicuramente più utenti di una biblioteca collocata in una provincia meno abitata. Ma con quale di questi due fattori è maggiormente associata la dimensione dell'utenza: con l'entità del patrimonio o con l'entità del bacino di utenza? Un piccolo test può fornirci indicazioni interessanti in proposito.

I dati utilizzati per il test sono riportati nella Tab. 7; si tratta dei dati relativi all'anno 1990 sull'utenza e il patrimonio delle biblioteche pubbliche statali — aggregati per provincia — nonché i dati, validi fino al 1989, relativi al totale degli abitanti delle province.[5]

Costruiamo ora, a partire da ➤

Tab. 7 - Dati statistici relativi alle biblioteche statali pubbliche

Provincia	Utenza[1]	Patrimonio[2]	Abitanti[3]
Avellino	1.340	163.288	447.822
Bari	46.414	342.420	1.515.742
Bologna	41.492	1.114.237	917.016
Cagliari	18.347	530.038	759.076
Cremona	89.059	590.218	228.613
Firenze	261.341	8.694.293	1.193.310
Frosinone	2.580	66.797	478.931
Genova	35.770	781.941	1.006.711
Gorizia	13.056	231.377	527.362
Lucca	37.243	460.907	382.882
Milano	22.674	1.330.340	3.978.658
Modena	55.503	783.816	595.610
Napoli	152.283	2.720.864	3.087.246
Padova	272.722	780.009	816.226
Parma	3.175	43.149	396.491
Pavia	101.000	529.302	501.470
Pisa	58.399	613.578	388.620
Potenza	10.387	49.460	412.361
Rieti	265	38.094	145.475
Roma	730.047	7.245.875	2.815.457
Salerno	1.600	85.184	1.053.766
Sassari	47.359	289.736	448.055
Torino	118.694	1.165.063	2.292.068
Trieste	50.806	161.974	269.878
Venezia	42.981	923.704	837.170

[1] Presenze nelle biblioteche pubbliche statali della provincia (dati 1990).
[2] Patrimonio complessivo delle biblioteche pubbliche statali della provincia (dati 1990).
[3] Abitanti della provincia (dati 1989).

Tab. 8 - Valori del coefficiente di contingenza per le tavole di contingenza di 2 righe per 25 colonne formate a partire dai dati della Tab. 7

Attributi riga	Attributi colonna	Coefficiente
Utenza, Patrimonio	Province	0,1969612
Utenza, Abitanti	Province	0,2844803
Patrimonio, Abitanti	Province	0,4261118

Tab. 9 - Programma di immissione dati (QBasic)

```
   REM immissione dati
   CLS
   REM file che contiene i nomi delle provincie:
   OPEN "a:\cities.txt" FOR OUTPUT AS #1
   REM file presenze nelle bibl.stat. della prov. (1990):
   OPEN "a:\p1.txt" FOR OUTPUT AS #2
   REM file patrimonio delle bibl.stat. della prov.(1990):
   OPEN "a:\p2.txt" FOR OUTPUT AS #3
   REM file che contiene gli abitanti della provincia (1989)
   OPEN "a:\p3.txt" FOR OUTPUT AS #4
 1 INPUT "Provincia    "; a1$
   INPUT "Presenze     "; a2
   INPUT "Patrimonio   "; a3
   INPUT "Abitanti     "; a4
   PRINT #1, a1$
   PRINT #2, a2
   PRINT #3, a3
   PRINT #4, a4
 3 INPUT "Ancora (s/n)"; z$
   IF z$ = "s" THEN CLS : GOTO 1
   IF z$ = "n" THEN CLS : GOTO 2
   CLS : GOTO 3
 2 CLOSE 1: CLOSE 2: CLOSE 3: CLOSE 4
   PRINT "concluso!": END
```

questi dati, le tre tavole di contingenza di 2 righe per 25 colonne che si ottengono considerando come attributi di colonna le 25 province e come attributi di riga gli attributi utenza, patrimonio e abitanti presi due alla volta. Calcoliamo su queste tre tavole di 2 righe per 25 colonne il coefficiente di contingenza e mettiamo a confronto i risultati. Il confronto è riportato nella Tab. 8.

Appare chiaro, dal confronto tra i coefficienti di contingenza, che vi è un discreto legame associativo tra il patrimonio bibliotecario di una provincia e il numero di abitanti (ovviamente si tratta di considerazioni riferite solo alle biblioteche pubbliche statali). In altri termini, i patrimoni bibliotecari più ricchi sono concentrati, tendenzialmente, nelle province più numerose. Vi è anche un certo legame — comunque più debole — tra il livello dell'utenza e gli abitanti della provincia. In altri termini si registrano, tendenzialmente, livelli di utenza più forti nelle province con più abitanti.

Molto basso invece è il legame associativo tra l'utenza e l'entità del patrimonio: in questo caso si può affermare, con una ragionevole probabilità di non sbagliare, che il livello di utenza e l'entità del patrimonio sono attributi indipendenti. In altre parole la ripartizione dell'utenza in relazione al patrimonio, nelle 25 province, sarebbe casuale. Questo può significare che l'entità del patrimonio, cioè la disponibilità di libri, non determina una maggior o minor propensione a frequentare la biblioteca. Altri fattori, diversi dalla disponibilità di libri, hanno una maggior influenza sull'afflusso in biblioteca.

Generalizzando potremmo affermare che — come linea di tendenza — i fattori decisivi circa l'entità dell'utenza vanno cercati *fuori* e non *dentro* la biblioteca. In sintesi, sembrerebbe questa l'informazione fondamentale desumibile, tramite la Tab. 8, dai dati statistici riportati nella Tab. 7. Occorre sempre ricordare però che queste considerazioni hanno dei fondamenti esclusivamente statistici. Questo implica che "il fatto che esista una correlazione tra due variabili non costituisce prova dell'esistenza di una relazione causale, e se una relazione causale appare possibile non è necessariamente ovvio in quale direzione essa agisca".[6]

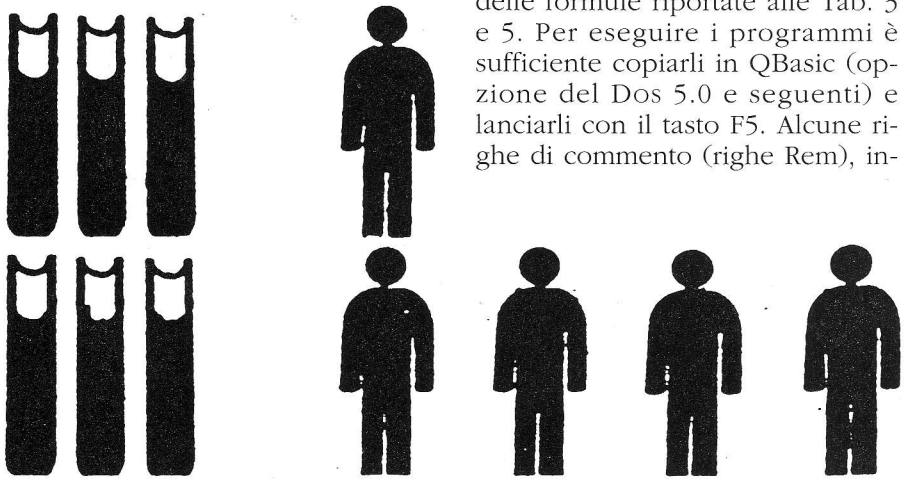
Con questo naturalmente non si è certo esaurito il tema di questo articolo: l'analisi dei dati statistici. Una trattazione completa di questo argomento richiederebbe, come è evidente, spazi molto più estesi. Crediamo comunque che,

Tab. 10 - Programma per il calcolo del chi-quadrato e del coefficiente di contingenza (QBasic).

```
REM  Chi-quadrato e coefficiente di contingenza
REM  per tavole di 2 righe per n colonne
REM        -------- input dei dati --------
CLS : INPUT "Quante colonne"; n
DIM matrix(2, n), nn(n), cities$(n)
OPEN "a:\cities.txt" FOR INPUT AS #1
FOR j = 1 TO n: INPUT #1, cities$(j): NEXT: CLOSE 1
CLS : INPUT "Primo file"; first$
CLS : INPUT "Secondo file"; second$
OPEN first$ FOR INPUT AS #1
OPEN second$ FOR INPUT AS #2
FOR j = 1 TO n
INPUT #1, a1
INPUT #2, a2
matrix(1, j) = a1:  matrix(2, j) = a2
NEXT: CLOSE 1: CLOSE 2
REM         -------- calcoli --------
na = 0: FOR j = 1 TO n: na = na + matrix(1, j): NEXT
nb = 0: FOR j = 1 TO n: nb = nb + matrix(2, j): NEXT
FOR j = 1 TO n: nn(j) = matrix(1, j) + matrix(2, j): NEXT
nnn = na + nb: suma = 0: sumb = 0: kiqua = 0
FOR j = 1 TO n
suma = suma + (matrix(1, j) ^ 2 / nn(j))
sumb = sumb + (matrix(2, j) ^ 2 / nn(j))
NEXT
kiqua = kiqua + (nnn / na) * suma + (nnn / nb) * sumb -nnn
ccc = SQR(kiqua / (kiqua + nnn))
REM     -------- visualizzazione dei risultati --------
CLS : PRINT "Confronto "; first$; " "; second$
PRINT : PRINT "Indice Chi-quadro:", kiqua
PRINT : PRINT "Indice C          :", ccc: PRINT
FOR i = 1 TO 2: FOR j = 1 TO n
PRINT matrix(i, j);
NEXT: PRINT : NEXT: PRINT : INPUT "Stampo (s/n)"; z$
IF z$ = "n" THEN END
REM             -------- stampa dei risultati --------
IF first$ = "a:\p1.txt" THEN first$ = "Presenze:"
IF first$ = "a:\p2.txt" THEN first$ = "Patrimonio:"
IF first$ = "a:\p3.txt" THEN first$ = "Abitanti:"
IF second$ = "a:\p1.txt" THEN second$ = "Presenze:"
IF second$ = "a:\p2.txt" THEN second$ = "Patrimonio:"
IF second$ = "a:\p3.txt" THEN second$ = "Abitanti:"
LPRINT : LPRINT
LPRINT "Provincia:     "; first$; "       "; second$: LPRINT
FOR j = 1 TO n
LPRINT USING "\                        \"; cities$(j);
LPRINT matrix(1, j), matrix(2, j)
NEXT: LPRINT "Coefficiente di contingenza:", ccc
END
```

per chi volesse proseguire l'approfondimento di questo argomento, le considerazioni riportate fin qui costituiscano un punto di partenza sufficientemente semplice e chiaro. Per chi infine volesse ripetere il test illustrato o compierne altri (con le dovute cautele circa le conclusioni affrettate) riportiamo, nelle Tab. 9 e 10, i due programmi Basic utilizzati per il nostro esempio.

Il programma (Tab. 10) si avvale, per il calcolo del chi-quadrato e del coefficiente di contingenza, delle formule riportate alle Tab. 3 e 5. Per eseguire i programmi è sufficiente copiarli in QBasic (opzione del Dos 5.0 e seguenti) e lanciarli con il tasto F5. Alcune righe di commento (righe Rem), intercalate alle istruzioni, facilitano la lettura del programma a chi ha dimestichezza con il linguaggio Basic. ■



Note

[1] Per queste considerazioni e, più in generale, per un approccio veramente *morbido* alla statistica, consigliamo, per chi ne sia completamente digiuno: G. KENNEDY, *Introduzione alla statistica*, Roma, Editori Riuniti, 1985, p. 152.

[2] M. SPIEGEL, *Statistica*, Milano, Etas libri, 1976, p. 204.

[3] *Ivi.*

[4] MINISTERO PER I BENI CULTURALI E AMBIENTALI, *Notiziario dell'Ufficio studi*, Direzione generale per gli affari generali amministrativi e del personale, VII. 38 / luglio-settembre 1992.

[5] *Ivi*, p. 28-33; *Il nuovo libro Garzanti della geografia, 1-l'Italia degli anni '90*, Milano, Garzanti, 1989.

[6] G. KENNEDY, *op. cit.*, p. 169.