QUADERNI DEL DIPARTIMENTO DI
SCIENZE DELL'UOMO

DSU 01/2005

Giovanni Delli Zotti (*dellizottig@sp.units.it*)

COME "FARE A FETTE"
UNA DISTRIBUZIONE DI FREQUENZA

Nuova edizione – Luglio 2007

Università degli Studi di Trieste
www.dsu.units.it

Quaderni del Dipartimento di Scienze dell'Uomo
Università degli Studi di Trieste
DSU: 01/2005
(www.dsu.units.it)

Come "fare a fette" una distribuzione di frequenza [*]

di Giovanni Delli Zotti

Sommario

Premessa

Nell'analisi monovariata ci si trova di fronte a due situazioni fondamentali: o si analizzano *classificazioni* (a *categorie non ordinate* o *ordinate*), oppure *variabili cardinali*[1]. Nel primo caso lo strumento adatto è la tabella della distribuzione di frequenza: le categorie in genere non sono molto numerose e sono pochi gli strumenti statistici a disposizione per sintetizzare altrimenti la distribuzione. Si tratta della *moda*, che non sintetizza, oppure della *mediana*, che non è molto utile quando le categorie della variabile ordinale sono poco numerose.

Nel caso di una variabile cardinale ci si trova invece spesso in una situazione opposta, in quanto si può sintetizzare la variabile anche mediante la *media* e proprio la tabella della distribuzione di frequenza si rivela uno strumento d'analisi poco utile. Nei casi più estremi la distribuzione di frequenza non sintetizza per nulla l'informazione contenuta nella colonna della matrice dei dati: si pensi alla popolazione dei comuni di una regione, in cui la misurazione accurata porta quasi sempre alla rilevazione di un valore diverso per ognuno dei casi. Accade lo stesso se registriamo il fatturato di un campione di aziende o il reddito così come è indicato precisamente nella dichiarazione dei redditi di qualche centinaio di contribuenti.

Realizzando la distribuzione di frequenza ci si limita infatti ad ordinare i valori in modo crescente (o decrescente, se si preferisce). Pur non riducendosi il numero delle informazioni di partenza, anche con questo tipo di variabili l'analisi della distribuzione di frequenza è utile, in quanto l'ordinamento dei valori permette, se non altro, di individuare agevolmente i valori

[*] Nuova edizione (luglio 2007) riveduta anche sulla base delle indicazioni di Alberto Marradi che ringrazio per la disponibilità. In questa versione è aggiunto un paragrafo finale nel quale è illustrato il modo per realizzare le procedure di segmentazione con l'ausilio del modulo di Spss® denominato "segmentazione grafica".

[1] Una variabile *categoriale* è relativa a proprietà in cui la registrazione dei valori non consente di utilizzare le proprietà ordinali e cardinali dei numeri (gli stati sulla proprietà sono *discreti non ordinabili* – ad es. la residenza).

Una variabile *ordinale* è relativa a proprietà in cui la registrazione dei valori consente di utilizzare le proprietà ordinali dei numeri, non quelle cardinali (gli stati sulla proprietà sono *discreti ordinabili* – ad es. l'istruzione).

Una variabile *cardinale* è relativa a proprietà in cui la registrazione dei valori consente di utilizzare le proprietà cardinali dei numeri. Deriva da una proprietà *continua misurabile* (ad es. l'altezza) o da una proprietà con *stati enumerabili* (ad es. il numero di figli) (adattato dal *Glossario* in G. Delli Zotti, *Introduzione alla ricerca sociale. Problemi e qualche soluzione*, Angeli, Milano, 2004, nuova edizione riveduta ed aggiornata).

minimo e massimo della distribuzione, e inoltre di procedere alla suddivisione della distribuzione in un insieme di classi di valori contigui.

A volte è necessaria, o preferibile, questa soluzione poiché, ad esempio, può essere ritenuto troppo semplificante mostrare della distribuzione originaria l'estrema sintesi costituita dalla media aritmetica dei valori, anche fosse corredata da qualche altro valore caratteristico sintetico o posizionale che mostri la dispersione dei dati. Inoltre, potremmo voler realizzare una tabella perché la stima del numero di intervistati che ricade all'interno di una determinata fascia di reddito, o di aziende che hanno una specifica dimensione in termini di forza lavoro, è proprio il tipo di informazione che vogliamo ottenere dai nostri dati.

Detto questo, quali sono le strategie per "fare a fette" la distribuzione? Cioè, *quante* devono essere le fette e *dove* si deve tagliare? L'esperienza e la riflessione su quanto accade in concreto quando si affronta questo problema hanno permesso a chi scrive di individuare quattro strategie fondamentali: una lista di opzioni che non pretende di essere esaustiva, ma che è abbastanza differenziata da tenere conto delle esigenze e circostanze che molto frequentemente si possono presentare.

Le quattro strategie possono avvalersi di tecniche diverse di segmentazione, le quali possono utilizzare:

1 – intervalli di valori di *uguale ampiezza,*
2 – gruppi di casi di *uguali dimensioni,*
3 – valori significativi o *valori soglia,*
4 – "fratture" o *discontinuità* nella distribuzione.

1. Costruire intervalli di valori di uguale ampiezza

Suddividere il campo di variazione della variabile in intervalli di uguale ampiezza è piuttosto usuale; probabilmente l'esempio più noto è quello della distribuzione di frequenza dell'età. Nelle tabelle statistiche proposte dall'Istat è consueta la suddivisione in classi di età costruite utilizzando intervalli decennali o quinquennali. Su questa base possono essere costruite le c.d. *piramidi d'età* (vedi fig. 1).

Siccome gli intervalli sono stati scelti come uguali, mediante la tabella della distribuzione per classi d'età e la sua eventualmente rappresentazione mediante la piramide, ciò che si vuole osservare e valutare è evidentemente la diversa consistenza delle classi rappresentate. Si potrà così notare, ad esempio, con quale rapidità la consistenza delle classi si riduce, specialmente per le fasce d'età più elevate ed in particolare, se utilizziamo la piramide d'età, sul versante dei maschi.

Nell'esempio proposto (tab. 1) la distribuzione delle età si riferisce ad un campione di utenti di un ente di patronato, cui ci si rivolge specialmente per pratiche riguardanti le pensioni. Non vi sono perciò individui con un'età inferiore ai 20 anni e la classe superiore è aperta verso i valori più elevati, per quanto si registrino solo due individui di età pari a 91 anni.

Nella distribuzione sono stati inseriti ed evidenziati in giallo (grigio chiaro) i valori per il quali non si sono riscontrati intervistati di età corrispondente. Le righe in rosso evidenziano i punti di taglio della distribuzione e l'integrazione dei valori non registrati nella rilevazione permette di mostrare anche graficamente come gli intervalli, disposti su tre colonne, abbiano pari ampiezza. Tra l'altro – ritorneremo in seguito su questo punto – riteniamo non sia condivisibile che il software solitamente utilizzato elimini arbitrariamente dalle tabelle di distribuzione di frequenza le categorie vuote. Ciò può essere sensato nel caso di una variabile categoriale (nemmeno le categorie previste dal ricercatore è detto che siano per forza esaustive e comunque potrebbero essere diverse da quelle utilizzate), ma nel caso di una variabile cardinale con valori interi questa semplificazione impedisce di visualizzare facilmente gli eventuali "buchi" nella distribuzione.

Il risultato della segmentazione in classi della distribuzione di tabella 1 è riportato nella tabella 2, dalla quale si nota chiaramente come ad un'uguale ampiezza della gamma di valori compresi in ognuna delle categorie corrisponda una consistenza della classe assai differente, da 150 intervistati tra i giovani da 20 a 29 anni, a quasi un migliaio tra i cinquantenni, a soli due intervistati con un'età superiore a 90 anni.

Tab. 1 – Distribuzione di frequenza dell'età

Valori	N	%	% cum.	Valori	N	%	% cum.	Valori	N	%	% cum.
20	1	,0	,0	**50**	71	2,5	32,9	**80**	13	,5	98,9
21	4	,1	,2	**51**	50	1,7	34,6	**81**	10	,3	99,3
22	5	,2	,3	**52**	83	2,9	37,5	**82**	8	,3	99,5
23	12	,4	,8	**53**	118	4,1	41,6	**83**	5	,2	99,7
24	13	,5	1,2	**54**	98	3,4	45,1	84	0	,0	99,7
25	11	,4	1,6	**55**	111	3,9	49,0	85	0	,0	99,7
26	18	,6	2,2	**56**	95	3,3	52,3	**86**	3	,1	99,8
27	24	,8	3,1	**57**	102	3,6	55,8	**87**	1	,0	99,9
28	33	1,2	4,2	**58**	101	3,5	59,4	**88**	1	,0	99,9
29	30	1,0	5,3	**59**	132	4,6	64,0	**89**	1	,0	99,9
30	31	1,1	6,4	**60**	109	3,8	67,8	90	-	-	-
31	39	1,4	7,7	**61**	85	3,0	70,8	**91**	2	,1	100,0
32	37	1,3	9,0	**62**	81	2,8	73,6	92	-	-	-
33	24	,8	9,9	**63**	88	3,1	76,7	93	-	-	-
34	31	1,1	10,9	**64**	89	3,1	79,8	94	-	-	-
35	35	1,2	12,2	**65**	94	3,3	83,1	95	-	-	-
36	35	1,2	13,4	**66**	60	2,1	85,2	96	-	-	-
37	39	1,4	14,7	**67**	59	2,1	87,2	97	-	-	-
38	41	1,4	16,2	**68**	37	1,3	88,5	98	-	-	-
39	34	1,2	17,4	**69**	42	1,5	90,0	99	-	-	-
40	23	,8	18,2	**70**	42	1,5	91,4	**Totale**	**2862**	**100,0**	**100,0**
41	26	,9	19,1	**71**	28	1,0	92,4				
42	42	1,5	20,5	**72**	38	1,3	93,7				
43	32	1,1	21,7	**73**	38	1,3	95,1				
44	38	1,3	23,0	**74**	17	,6	95,7				
45	35	1,2	24,2	**75**	27	,9	96,6				
46	39	1,4	25,6	**76**	12	,4	97,0				
47	45	1,6	27,1	**77**	15	,5	97,6				
48	42	1,5	28,6	**78**	13	,5	98,0				
49	51	1,8	30,4	**79**	13	,5	98,5				

N.B.
rosso: limite delle 8 classi decennali
azzurro: valore che suddivide la distribuzione in quartili
verde: valori che suddividono in due parti i quartili
giallo: valori senza casi omessi nella distribuzione di frequenza prodotta dal computer

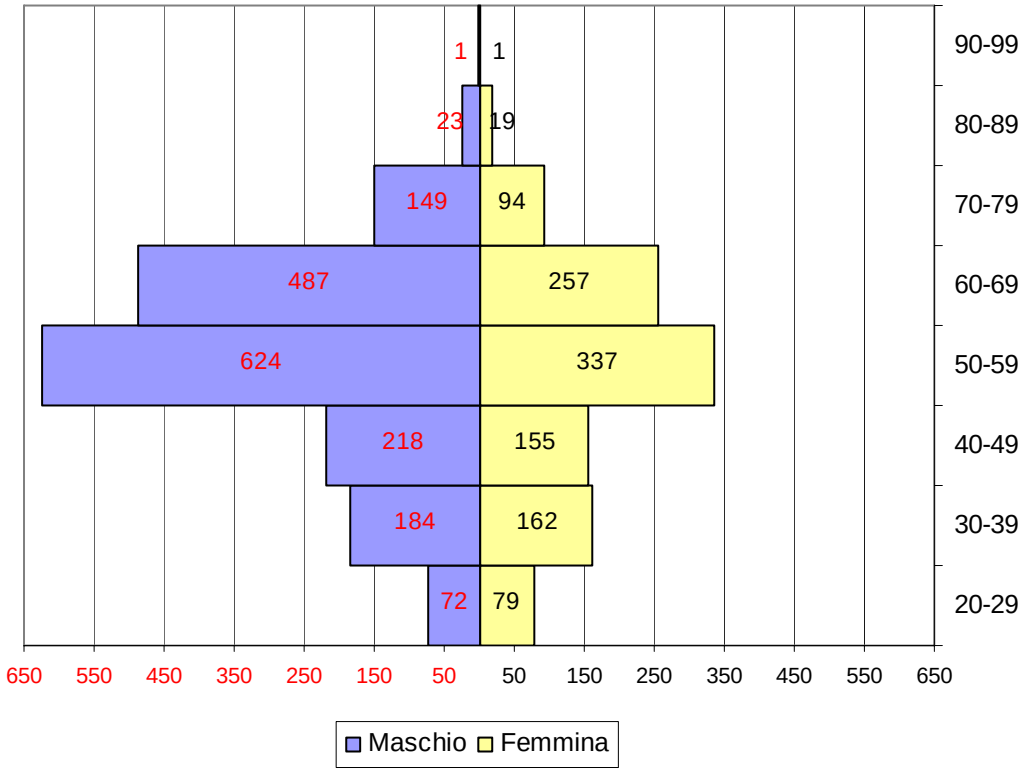
Tab. 2 – Distribuzione di frequenza delle classi d'età (intervalli di valori uguali)

Valori	N	%	% cum.
20-29	151	5,3	5,3
30-39	346	12,1	17,4
40-49	373	13,0	30,4
50-59	961	33,6	64,0
60-69	744	26,0	90,0
70-79	243	8,5	98,5
80-89	42	1,5	99,9
90-99	2	,1	100,0
Totale	**2862**	**100,0**	

La situazione è rappresentata graficamente nella figura 1, dove si è deciso di mostrare la già ricordata piramide d'età, nella quale vengono rappresentate contrapponendole le distribu-

zioni di frequenza dell'età dei maschi e delle femmine[2]. Trattandosi della distribuzione di frequenza di un campione di persone un po' particolare, la figura rappresentata assume la forma di una botte, con una base molto esigua che poi cresce fino a raggiungere l'ampiezza massima in corrispondenza della classe d'età per la quale il problema della pensione diventa imminente. Si può anzi dire che si tratta della distribuzione dell'età di due sottopopolazioni: coloro che si rivolgono al patronato perché direttamente interessati alle pratiche della pensione e coloro che, molto più giovani, si rivolgono al patronato per conto del lavoratore anziano.

Fig. 1 – Piramide di età



Un altro caso piuttosto usuale è quello del reddito annuale. Se, ad esempio, costruiamo intervalli con 5.000 euro di ampiezza, possiamo controllare se nella popolazione studiata vi è una consistente classe media, oppure una base piuttosto ampia di persone a basso reddito o, ancora, di quanto si assottigli la consistenza delle classi mano a mano che il reddito sale.

Per finire, è plausibile e particolarmente appropriato ridurre a 10 classi di uguale ampiezza la distribuzione di una variabile cardinale costruita percentualizzando un valore osservato, come avviene nei rapporti di composizione (ad esempio, la percentuale di voti ottenuti da un particolare partito o la percentuale della popolazione alfabetizzata). Nell'esempio (tab. 3), tratto dal file World95, presente tra quelli forniti per effettuare le esercitazioni all'interno del programma di elaborazione dati Spss, è rappresentata la percentuale di popolazione che, nei

[2] Tra i programmi maggiormente utilizzati, Spss solo dalla versione 13 ha introdotto questo tipo di rappresentazione nella sua galleria di tipi di grafici. La piramide di età non è esplicitamente presente tra i grafici proposti da Excel®, ma può essere ugualmente realizzata ricorrendo ad un espediente: i dati di una delle due categorie vengono scritti come valori negativi e si possono poi utilizzare per la rappresentazione le barre suddivise (barre in pila nella terminologia di Excel). Si può anche evitare che i numeri vengano effettivamente scritti con il segno meno, che può apparire fuorviante, ricorrendo alla forma di rappresentazione come numeri debitori scritti in rosso (vedi fig.1). I diagrammi a barre in forma di piramide possono essere utilizzati anche in altre situazioni in cui una delle due variabili sia una dicotomia: per qualche applicazione di questo tipo si veda, ad esempio, G. Delli Zotti (2004), *L'Università di Trieste: studenti e docenti*, Quaderni del Dipartimento di Scienze dell'Uomo, Quad-DSU 2/2004, Trieste, scaricabile all'indirizzo: www.dsu.units.it/quaderni/dsu0204.pdf.

108 paesi considerati[3], vive in zone urbane: una variabile costruita rapportando la popolazione urbana al totale della popolazione e moltiplicando per 100 la proporzione risultante.

Siccome questo scritto ha eminentemente finalità didattiche, non sarà inutile chiarire che la tabella può ingenerare confusione nel lettore meno avvertito. I valori scritti nella prima colonna sono semplici etichette che indicano l'intervallo (di percentuali) al quale ci si riferisce in ognuna delle righe della tabella. Le ultime due colonne invece si riferiscono alla percentuale di osservazioni che ricade all'interno dell'intervallo (di percentuali) di volta in volta considerato e, nel caso delle percentuali cumulate, alla percentuale di osservazioni che ricadono in quell'intervallo e in quelli precedenti. Ad esempio, sono pari al 14,8% i paesi nei quali nelle zone urbane vive una percentuale di popolazione variabile tra il 41 e il 50% del totale della popolazione e accanto leggiamo l'informazione che nel 39,8% dei casi la popolazione degli stati rappresentati nella distribuzione non supera il 50%.

Tab. 3 - Percentuale di popolazione che vive in città in 108 paesi del mondo

Valori	N	%	% cum.
1-10	2	1,9	1,9
11-20	8	7,4	9,3
21-30	10	9,3	18,5
31-40	7	6,5	25,0
41-50	16	14,8	39,8
51-60	14	13,0	52,8
61-70	14	13,0	65,7
71-80	16	14,8	80,6
81-90	14	13,0	93,5
91-100	7	6,5	100,0
Totale	**108**	**100,0**	

Da ciò non si può concludere necessariamente che nel mondo prevale la popolazione urbana. Il complemento a 100 di 39,8 ci dice infatti che nel 60,2 % dei paesi prevale la popolazione urbana, ma non si tratta di una stima corretta del totale della popolazione mondiale che vive in città. Per poter effettuare una stima corretta si devono pesare i casi, e cioè tenere conto della dimensione della popolazione di ognuno di paesi per i quali si è rilevata la percentuale di popolazione urbana. Infatti, dai dati si evince che la popolazione urbana in Cina è pari al 26%, il che corrisponde a oltre 300 milioni di abitanti; la percentuale è simile a quella del Botswana (25%), ma che corrisponde a meno di 350.000 abitanti e, al contrario, il 91% di islandesi che vive in città corrisponde a soli 240.000 abitanti circa.

Tab. 4 - Percentuale di popolazione che vive in città (casi pesati in base alla popolazione)

Valori	N	%	% cum.
1-10	14400	,3	,3
11-20	315500	6,1	6,4
21-30	2449385	47,2	53,5
31-40	273400	5,3	58,8
41-50	296056	5,7	64,5
51-60	179700	3,5	68,0
61-70	328511	6,3	74,3
71-80	1033744	19,9	94,2
81-90	253944	4,9	99,1
91-100	46863	,9	100,0
Totale	**5191503**	**100,0**	

[3] Dalla matrice dei dati sono evidentemente esclusi i c.d. mini-Stati, presenti in particolare nell'area caraibica, ma dei quali vi sono diversi esempi anche in Europa (Lussemburgo, Cipro, Monaco, Andorra, San Marino, ecc.).

Il ricalcolo, effettuato pesando i casi in base alla variabile "popolazione", porta ai risultati della tab. 4[4] nella quale i casi non sono più i 108 originali, ma oltre 5 milioni, corrispondenti ad oltre 5 miliardi di persone (nella matrice dei dati registrati in migliaia). In questa nuova distribuzione ogni singolo paese non conta più per uno, ma in relazione al numero dei suoi abitanti. Si vede allora, ad esempio, che poco meno della metà della popolazione vive in paesi nei quali la quota di popolazione urbana varia dal 21 al 30% e che più della metà (53,5%) vive in paesi nei quali la popolazione urbana non supera il 30%. Inoltre, dalla tabella precedente si leggeva che i paesi nei quali la popolazione urbana varia tra il 91 e il 100% erano pari al 6,5% del totale dei paesi, ma dalla tabella 4 si vede che la popolazione che vive in questo tipo di paesi è pari a meno dell'1% del totale degli abitanti.

Per concludere questo breve inciso, va osservato che anche i valori medi non sono stimati correttamente senza la ponderazione dei casi. Infatti, la media della variabile originale, prima della sua riduzione in classi di ampiezza pari a 10 punti percentuali, risultava essere pari a 56,5 (significa la percentuale di popolazione urbana nei 108 paesi è mediamente pari al 56,5%) . Da ciò non si può dedurre che in media oltre il 50% della popolazione vive in città. La media calcolata sulla base dei casi pesati risulta infatti essere pari a 43,7% ed è questa la stima più corretta della popolazione urbana nel mondo. I paesi più popolosi (Cina e India, ad esempio) hanno una percentuale di popolazione urbana più esigua della media e, quando viene loro attribuito il peso che loro spetta in base alla popolazione (anziché contarli solo per un caso), fanno sentire questo loro peso abbassando la media generale.

2. Costruire gruppi di uguali dimensioni

La seconda strategia di segmentazione si base sulle frequenze percentuali cumulate della distribuzione in quanto ci si propone di costruire le classi in modo che contino al loro interno un uguale numero di casi. Se non ci sono ragioni che consiglino altre soluzioni, ci è in questa procedura un vantaggio tecnico non trascurabile: se il numero di classi non è troppo elevato in rapporto all'entità complessiva dei casi, il ricercatore può essere certo di non ritrovarsi con una o più classi comprendenti solo pochi casi, come può accadere quando sono invece uguali gli intervalli di valori. Ciò è utile perché, ad esempio, nell'analisi bivariata una categoria con solo pochi casi diventa pressoché inutilizzabile, in quanto statisticamente poco significativa. Inoltre, alcuni indicatori sono basati proprio sulla suddivisione dei casi in gruppi di uguale dimensione. La c.d. "soglia di povertà" può essere stabilita utilizzando questa strategia di segmentazione delle distribuzioni, individuando, ad esempio, il *quartile inferiore* della distribuzione (che corrisponde al 25% di famiglie con il reddito più basso)[5].

Nell'esempio proposto (tab. 5) si è analizzata la distribuzione della ricchezza a livello globale, utilizzando nuovamente i dati dell'archivio Word95 di Spss. Dei 109 paesi, il quarto più povero non supera i 1.000 dollari di Prodotto Nazionale Lordo (PNL) pro capite, il secondo quarto di paesi non raggiunge i 3.000 dollari. Infatti, al valore di 2.995 dollari si colloca la mediana (confine superiore del secondo quartile), che individua il paese che si trova nella po-

[4] È semplice ottenere questo ricalcolo e la tabella 4 con Spss: basta attivare la funzione "pesa casi" dal menù "dati" e quindi indicare la variabile "popolazione in migliaia" che dovrà essere utilizzata per pesare i casi.

[5] In realtà, vi sono diversi modi per individuare la soglia di povertà e la definizione di "povertà relativa" considera al di sotto di questa soglia le famiglie di due persone che spendono mensilmente per consumi un importo inferiore alla spesa mensile pro capite. Per famiglie di diversa ampiezza, la linea di povertà va individuata applicando una "scala di equivalenza" per tenere conto delle economie di scala realizzabili all'aumentare del numero di componenti. Per una rassegna delle diverse tecniche utilizzate e proposte per "misurare" la povertà, si veda, ad esempio, Zaickzyk F. (1993), *Problematiche teoriche e metodologiche per la misurazione della povertà*, in M. Palumbo (a cura di), *Classi, disuguaglianze e povertà. Problemi di analisi*, Angeli, Milano e, inoltre, le diverse pubblicazioni sul tema della povertà scaricabili dal sito dell'Istat (www.istat.it).

sizione centrale della distribuzione ordinata dei casi. Sui 7.467 dollari troviamo infine l'ultima soglia, al di sopra della quale è collocato il 25% di paesi più ricchi.

Tab. 5 – Prodotto Nazionale Lordo (PNL) pro capite in 109 paesi del mondo nel 1995

Valori	N	%	% cum.
122	1	,9	,9
202	1	,9	1,8
205	1	,9	2,8
208	1	,9	3,7
230	1	,9	4,6
260	1	,9	5,5
263	1	,9	6,4
275	1	,9	7,3
282	1	,9	8,3
292	1	,9	9,2
323	1	,9	10,1
325	1	,9	11,0
351	1	,9	11,9
357	1	,9	12,8
377	1	,9	13,8
383	1	,9	14,7
406	1	,9	15,6
409	1	,9	16,5
447	1	,9	17,4
457	1	,9	18,3
573	1	,9	19,3
681	1	,9	20,2
730	1	,9	21,1
744	1	,9	22,0
748	1	,9	22,9
867	1	,9	23,9
993	1	,9	24,8
1000	1	,9	25,7
1030	1	,9	26,6
1034	1	,9	27,5
1062	1	,9	28,4
1078	1	,9	29,4
1085	1	,9	30,3
1107	1	,9	31,2
1157	1	,9	32,1
1342	1	,9	33,0
1350	1	,9	33,9
1382	1	,9	34,9
1429	1	,9	35,8
1500	2	1,8	37,6
1538	1	,9	38,5
1800	1	,9	39,4
1955	1	,9	40,4
2031	1	,9	41,3
2126	1	,9	42,2
2340	1	,9	43,1
2354	1	,9	44,0
2397	1	,9	45,0
2436	1	,9	45,9
2591	1	,9	46,8
2677	1	,9	47,7
2702	1	,9	48,6
2829	1	,9	49,5
2995	1	,9	50,5
3000	1	,9	51,4
3098	1	,9	52,3
3128	1	,9	53,2
3131	1	,9	54,1
3408	1	,9	55,0
3604	1	,9	56,0
3721	1	,9	56,9
3831	1	,9	57,8
4283	1	,9	58,7
4429	1	,9	59,6
4500	1	,9	60,6
5000	1	,9	61,5
5249	1	,9	62,4
5487	1	,9	63,3
5910	1	,9	64,2
6000	1	,9	65,1
6500	1	,9	66,1
6627	1	,9	67,0
6651	1	,9	67,9
6680	1	,9	68,8
6710	1	,9	69,7
6818	1	,9	70,6
6950	1	,9	71,6
7055	1	,9	72,5
7311	1	,9	73,4
7400	1	,9	74,3
7467	1	,9	75,2
7875	1	,9	76,1
8060	1	,9	77,1
9000	1	,9	78,0
12170	1	,9	78,9
13047	1	,9	79,8
13066	1	,9	80,7
14193	1	,9	81,7
14381	1	,9	82,6
14641	1	,9	83,5
14990	1	,9	84,4
15877	1	,9	85,3
15974	1	,9	86,2
16848	1	,9	87,2
16900	1	,9	88,1
17241	1	,9	89,0
17245	1	,9	89,9
17500	1	,9	90,8
17539	1	,9	91,7
17755	1	,9	92,7
17912	1	,9	93,6
18277	1	,9	94,5
18396	1	,9	95,4
18944	1	,9	96,3
19860	1	,9	97,2
19904	1	,9	98,2
22384	1	,9	99,1
23474	1	,9	100,0
Totale	**109**	**100,0**	

Nella tabella sono evidenziati (con una linea tratteggiata) anche altri due valori-soglia solitamente utilizzati in questo tipo di analisi, e cioè quelli che individuano il primo e l'ultimo *decile*: il primo decile comprende il 10% di paesi più poveri e l'ultimo decile il 10% di paesi più ricchi. Dalla tabella è possibile ricavare il pnl medio pro capite nei paesi più poveri sommando direttamente il pnl di ognuno dei paesi (la frequenza per ogni valore essendo sempre pari a 1) e dividendolo la somma per 11 (il valore ottenuto è pari a 242 dollari). Si può operare analogamente nell'altro versante della distribuzione, ottenendo un pnl medio nei paesi considerati di 19.268 dollari circa: dal che si ricava che i paesi più ricchi sono poco meno di 100 volte più ricchi dei paesi più poveri.

Con una riga continua abbiamo inoltre indicato un altro insieme di valori che verrà più ampiamente commentato nel paragrafo seguente: si tratta di cifre tonde, che assumono un particolare rilievo perché facili da memorizzare e/o costituiscono una soglia facilmente apprezzabile nel suo significato reale. Si sono scelti i valori che indicano il non superamento del livello dei 500, 1.000, 2.000, 5.000 e 10.000 dollari di pnl, ma evidentemente si possono effettuare scelte diverse, anche tenendo conto di quanti siano i gruppi che si vogliono formare.

Anche in questo caso i valori calcolati non sono corretti se vogliamo inferire il pnl medio pro capite del 10% più povero della popolazione mondiale. Il 10% di paesi con il pnl più basso infatti non corrisponde al 10% della popolazione in quanto, ad esempio, tra questi (al-

l'ottavo posto partendo dal basso) si trova l'India, che da sola conta per circa il 20% della popolazione mondiale. Come abbiamo visto in precedenza, se ponderiamo i paesi in base alla loro popolazione ricaviamo un'altra distribuzione. Siccome il caso (ponderato) che corrisponde al primo decile è proprio l'India, se si comprende l'intero paese la percentuale cumulata della popolazione mondiale supera largamente il 10% (23,7%) ed il pnl pro capite medio si innalza leggermente rispetto alla distribuzione precedente. Nonostante il fatto che il valore soglia scenda da 323 a 275 dollari, il pnl pro capite medio si innalza in quanto, comprendendo l'intera popolazione indiana, abbiamo largamente superato, rispetto al calcolo precedente, il 10% della popolazione, in quanto vengono compresi molti redditi che dovrebbero essere attribuiti alla fascia successiva. Si potrebbe effettuare un calcolo più preciso includendo solo i redditi della quota di popolazione indiana utile al raggiungimento del 10% della popolazione mondiale, ma affidiamo questo compito alla diligenza e buona volontà del lettore. Noi ci limitiamo a segnalare che all'epoca della rilevazione quasi un quarto della popolazione mondiale viveva con un pnl pro capite di circa 250 dollari.

Ponderando i paesi secondo la popolazione, anche nell'altro versante della distribuzione la situazione cambia: i paesi che contengono il 10% di popolazione più ricca si riducono da 11 a 8 (dato non visualizzato in tabella) ed il reddito medio della popolazione di questi paesi (pari al 9,8% della popolazione mondiale) gode in media di un pnl pro capite di 21.552 dollari.

Tab. 6 - Paesi del mondo per fasce di Prodotto Nazionale Lordo pro capite

Fasce di pnl in $	**N**	**%**	**% cum.**	**Pnl pro capite medio**	**Scarto tipo**
da 122 a 323	11	10,1	10,1	242	55,8
da 325 a 17.245	87	79,8	89,9	4876	4924,4
da 17.500 a 23.464	11	10,1	100,0	19268	2006,1
Totale	**109**	**100,0**		**5860**	**6479,8**
(paesi ponderati in base alla popolazione)					
da 122 a 275	1231200	23,7	23,7	256	38,3
da 282 a 17.500	3462703	66,6	90,2	2979	4597,4
da 17.539 a 23.464	508000	9,8	100,0	21552	2048,8
Totale	**5201903**	**100,0**		**4148**	**6968,2**

È interessante a questo punto illustrare la tecnica grafica di rappresentazione delle distribuzioni che si chiama *box plots* (grafici a scatola), concepita proprio al fine di mostrare in forma sintetica la posizione sull'asse dei valori di alcuni valori posizionali[6]. I valori usati per costruire l'altezza della scatola (*box*) (o la larghezza se viene costruita in orizzontale) sono il quartile superiore ed inferiore della distribuzione, mentre la linea centrale rappresenta la posizione della mediana. Le linee che si estendono dal *box* rappresentano le code della distribuzione e si estendono fino ad una distanza massima di una volta e mezza la differenza interquartile, mentre gli *outliers* (casi devianti) sono marcati individualmente (i *box-plots* prodotti da Spss distinguono i casi devianti dai casi estremi).

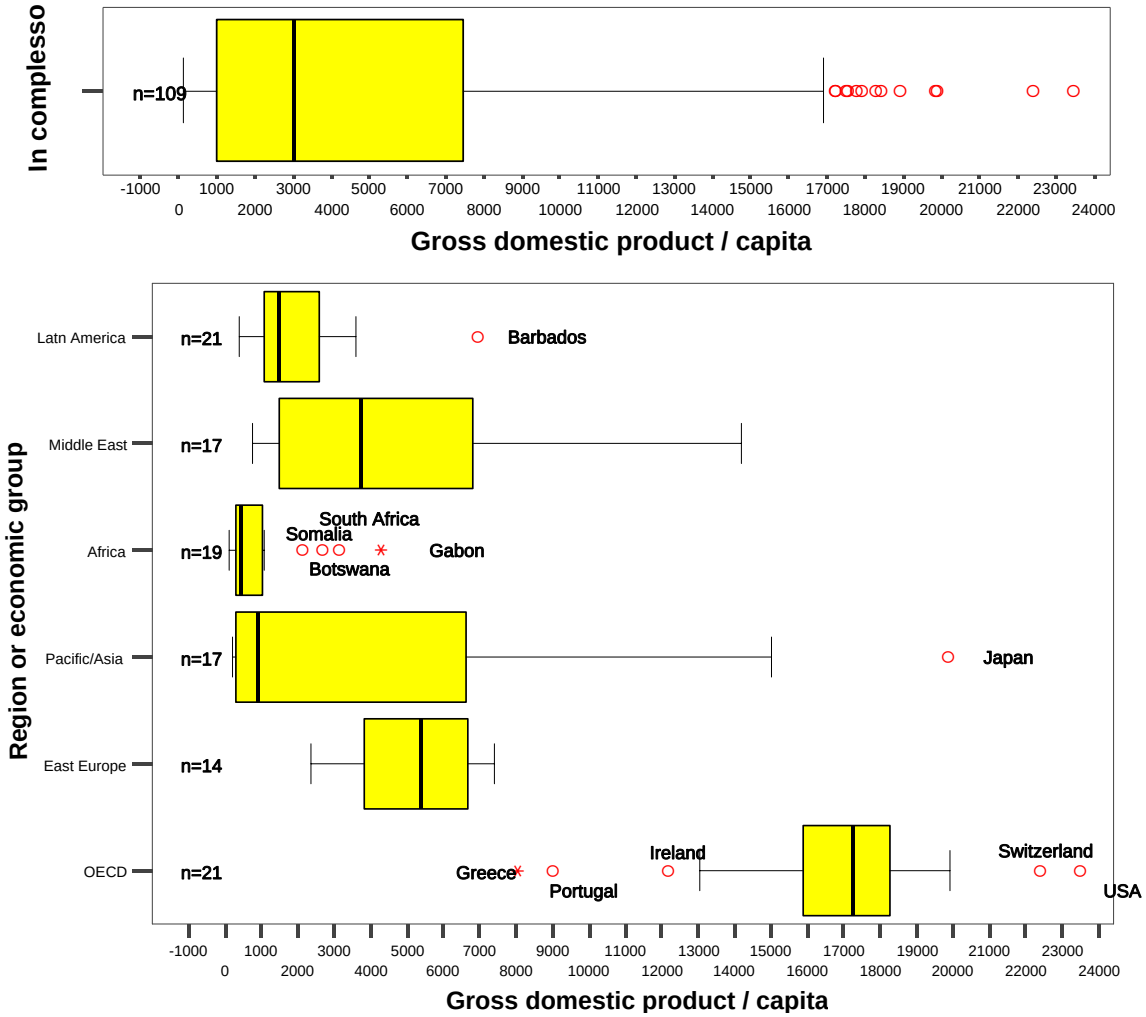
Nel *box plot* proposto (fig. 2) è innanzitutto rappresentata la distribuzione di frequenza del pnl pro capite del complesso dei 109 paesi, al fine di visualizzare l'estensione del campo di variazione del pnl dei 4 gruppi contrassegnati nella tabella 5 dall'evidenziazione in azzurro (grigio chiaro). Si è poi aggiunta le serie di *box-plots* che rappresentano la suddivisione in quartili dei paesi a seconda della regione o gruppo di appartenenza.

Il lettore dovrà prestare particolare attenzione nel valutare queste rappresentazioni grafiche in quanto i *box-plots*, pure utilissimi, si prestano ad equivoci interpretativi per il lettore disattento o non sufficientemente addestrato. All'interno di una specifico box, il fatto che la parte a

[6] Assieme ai grafici *stem-and-leaf* (ramo-foglia), dei quali si vedrà un esempio nel paragrafo 4, i *box-plots* sono stati proposti da J.W. Tukey, *Exploratory Data Analysis*, Reading, Addison-Wesley, 1977.

destra e a sinistra della mediana abbiano una diversa estensione non significa che nelle due metà del box vi sia un diverso numero di paesi. Per definizione, infatti, nelle due metà c'è lo stesso numero di paesi; quello che cambia è la loro dispersione all'interno di un campo di variazione dei valori più o meno ampio. Nel caso della regione del Pacifico/Asia, se si toglie il caso deviante costituito dal Giappone, i quattro paesi rappresentati dalla quasi invisibile linea che si estende sulla sinistra sono "schiacciati" all'interno di un campo di variazione che si discosta poco dai 200 dollari, altri quattro paesi sono contenuti nel semi-box a sinistra, altri quattro nel semi-box a destra che, partendo da meno di 1.000 dollari, supera largamente i 6.000. Infine, i quattro paesi più ricchi della regione sono contenuti nel campo di variazione che si estende da quasi 7.000 dollari fino a circa 15.000 (il programma Spss può ovviamente fornire i valori precisi che corrispondono ai diversi quartili).

Fig. 2 – *Box-plot* del reddito di 109 paesi nel complesso e per regione o gruppo economico



La strategia che utilizza i valori posizionali per suddividere le distribuzioni di frequenza è particolarmente utile quando i valori non hanno un particolare significato, come nel caso di punteggi su indicatori "astratti" (qualità della vita), oppure di punteggi di valutazione della prestazione di istituzioni, come è il caso dell'università nell'esempio che proponiamo[7].

[7] Si tratta dei punteggi attribuiti alle facoltà di ingegneria nell'indagine del Censis, che ogni anno pubblica le graduatorie di eccellenza delle facoltà italiane. I dati dell'esempio sono stati pubblicati su *La Repubblica* del 23 luglio 2000 e sono stati tratti da: G.P. Zaccomer, *Relazione sulle graduatorie Censis Servizi pubblicate su La grande guida all'Università – La Repubblica*, CESV, Udine, 2000.

Come abbiamo fatto in precedenza, nella tabella 7 sono evidenziati, oltre ai valori che contengono i casi che si collocano ai confini dei quartili, anche quelli che sarebbero stati scelti se si fosse deciso di sezionare la distribuzione in intervalli uguali, corrispondenti ad incrementi di 100 punti. Siccome ad ogni valore evidenziato nella distribuzione corrisponde in genere un solo caso, a prima vista gli intervalli evidenziati dalla segmentazione della distribuzione in quartili potrebbero sembrare uguali. Ciò appare perché lo sguardo si sofferma sul numero di righe della tabella e, ovviamente, siccome in ogni quartile ci è un uguale numero di casi, nei quattro quartili ci sono le stesse righe (o meglio, ci sarebbero le stesse righe se non fosse per un paio di valori che registrano la frequenza di due casi). In realtà, ad ogni quartile corrispondono intervalli nei valori di entità molto diseguale: il primo contiene infatti valori (punteggi) che variano da 108 a 260 (con una differenza di 152 punti), con il secondo si passa da 260 a 379 (la differenza è di 119 punti), con il terzo da 379 a 443 (la differenza è di soli 64 punti); infine, il quarto quartile abbraccia valori che passano da 443 a 717 punti (la differenza è in questo caso di ben 274 punti).

Paradossalmente, sembrano invece diseguali gli intervalli di valori, in quanto le righe continue che li delimitano sono più o meno ravvicinate proprio perché gli intervalli sono uguali, ma comprendono un numero diseguale di casi. Come nella tabella 1, anche questi intervalli apparirebbero uguali se integrassimo la tabella con i valori per i quali non si è registrata alcuna frequenza – cosa ovviamente assai poco pratica da realizzare quando i valori sono molto numerosi. Però, come vedremo a proposito della quarta strategia, con gli istogrammi si può ovviare a questa distorsione percettiva.

Comunque, per meglio evidenziare l'aspetto della distanza tra i singoli valori, sul quale torneremo nel paragrafo successivo, nella tabella siamo ricorsi all'espediente di inserire, a fianco della colonna dei valori stessi, una colonna aggiuntiva nella quale per ogni valore è annotato l'intervallo rispetto al valore precedente (che corrisponde al numero delle righe vuote che sono state omesse dalla tabella). Si notano in effetti alcuni salti piuttosto ampi. Come vedremo, proprio queste fratture nella distribuzione possono essere utilizzate per un'ultima strategia di sezionamento delle distribuzioni.

Tab. 7 – Punteggi attribuiti alle facoltà di ingegneria dall'indagine Censis del 2000

Valori	Intervallo	N	%	% cum.
108	-	1	2,8	2,8
197	89	1	2,8	5,6
212	15	1	2,8	8,3
229	17	1	2,8	11,1
235	6	1	2,8	13,9
246	11	1	2,8	16,7
252	6	1	2,8	19,4
257	5	1	2,8	22,2
260	3	2	5,6	27,8
268	8	1	2,8	30,6
277	9	1	2,8	33,3
284	7	1	2,8	36,1
290	6	1	2,8	38,9
323	33	2	5,6	44,4
357	34	1	2,8	47,2
379	22	1	2,8	50,0
382	3	1	2,8	52,8
384	2	1	2,8	55,6
389	5	1	2,8	58,3
392	3	1	2,8	61,1
399	7	1	2,8	63,9
424	25	1	2,8	66,7
438	14	1	2,8	69,4
439	1	1	2,8	72,2
443	4	1	2,8	75,0
458	15	1	2,8	77,8
461	3	1	2,8	80,6
466	5	1	2,8	83,3
510	44	1	2,8	86,1
515	5	1	2,8	88,9
580	65	1	2,8	91,7
622	42	1	2,8	94,4
689	67	1	2,8	97,2
717	28	1	2,8	100,0
Totale		**36**	**100,0**	

Non ne abbiamo parlato nel paragrafo precedente in quanto la distribuzione delle età della tabella 1 partiva dalla cifra tonda di 20 anni ma, osservando la tabella 7, risulta più evidente

che, quando si decide di costruire le classi sulla base di intervalli uguali di valori, ci si trova di fronte ad un'ulteriore scelta: decidere se gli intervalli uguali devono partire dal minimo teorico o dal minimo empirico. Che si tratti di età, di reddito, o di qualche tipo di punteggio, la prima classe può partire dal valore zero (o da qualsiasi altro valore che costituisca il valore minimo teorico) oppure in alternativa gli intervalli possono iniziare dal valore minimo riscontrato nella distribuzione empirica dei casi. Inoltre, in particolare quando si utilizzano punteggi "astratti", può essere opportuno suddividere in intervalli il reale campo di variazione della variabile, tenendo conto anche del massimo empiricamente osservato, invece che del massimo teorico.

Nella tabella 7 sono stati evidenziati con righe tratteggiate quattro intervalli uguali così definiti: il campo di variazione della variabile è pari a 606 punti, che divisi per 4 danno un intervallo di 152,25 punti per raggiungere 717 punti a partire da 108 con quattro intervalli. Perciò, gli intervalli uguali vengono individuati tagliando la distribuzione a 152,25, 412,5 e 564,75 punti.

Confrontando la seconda strategia qui illustrata con quella precedente, si può sinteticamente affermare che, in generale, quando i valori hanno un significato, è opportuno segmentare la distribuzione secondo i valori: in questo caso è interessante vedere se nelle classi c'è un numero di casi diverso (vedi la piramide d'età). Quando invece i valori non hanno significati condivisi o noti (o si tratta di valori convenzionali e mutevoli nel tempo come è il caso del reddito), è meglio creare classi di dimensioni uguali e vedere a quale valore corrispondono il quartile più alto, il più basso, etc.

Visto che manca un preciso significato attribuibile ai punteggi e la suddivisione viene effettuata mediante i valori posizionali (i quali non hanno altro significato che quello di individuare fasce più o meno ampie all'interno della distribuzione di frequenza originaria), appare convincente impiegare termini come alto, medio, basso, oppure inferiore, intermedio, superiore per definire le fasce stesse. Si possono anche immaginare una serie di opzioni riguardo ai termini impiegabili, a seconda del numero di categorie che si sceglie di istituire.

Numero	**Denominazione categorie**				
2	Basso				Alto
3	Basso		Medio		Alto
4	Basso	Medio-basso		Medio-alto	Alto
5	Basso	Medio-basso	Medio	Medio-alto	Alto

3. Individuare valori significativi o valori soglia

Anche la terza strategia (come la prima) parte dai valori della variabile ma, anziché costruire meccanicamente intervalli di uguale ampiezza, si ripropone di individuare alcuni valori con un particolare significato. In genere si tratta di *valori soglia*, e cioè valori che vengono per consuetudine considerati *valori standard* che indicano una "transizione". In alcuni casi si tratta di valori fissati per legge: nel caso dell'età si tratta, ad esempio, dell'età lavorativa minima (14 anni), della maggiore età (18 anni) o dell'età della pensione (65 anni per alcune categorie di lavoratori) o, ancora, dell'età alla quale si vota per il Senato (25 anni), oppure quella che permette di essere candidati al Senato stesso (40 anni).

Nel caso del reddito un valore significativo può essere quello della c.d. "soglia di povertà", oppure quello dell'esenzione dal pagamento delle tasse sulle persone fisiche e i diversi livelli di reddito che individuano gli scaglioni in cui si passa da una aliquota dell'Irpef all'altra. Come abbiamo visto nel paragrafo precedente, i valori soglia possono anche essere rappresentati da cifre tonde che il lettore è in grado di figurarsi facilmente; ad es., nel caso della popolazio-

ne di stati, il superamento dei 100.000 abitanti, del milione, dei 10 milioni, dei 100, oppure del miliardo[8].

Un esempio interessante può essere costituito dal problema della conversione del sistema di votazione in trentesimi adottato in Italia al sistema unificato europeo ECTS. In questo caso si tratta proprio di individuare i punti di transizione da una valutazione all'altra, seguendo la tabella di conversione. In questo modo è possibile sezionare la distribuzione di frequenza in modo da attribuire alla votazione conseguita una delle lettere che il sistema europeo, per lo meno in una delle formulazioni a noi note, attribuisce ai diversi esiti delle prove, partendo dalla valutazione A = *excellent*, per finire con F – *failed*.

Tab. 8 – Voti agli esami e corrispondenti giudizi nel sistema ECTS

	Voti	N	% sul totale	% su tot. superato	% cum. superato	
Non	7	3	2,1			
superato	8	2	1,4			
	9	2	1,4			
	10	1	,7			
	12	3	2,1			**F – failed**
	13	3	2,1			
	14	4	2,8			
	15	1	,7			
	16	3	2,1			
	17	1	,7			
	Sub-totale	**23**	**16,2**			
Superato	18	4	2,8	3,4	3,4	
	19	5	3,5	4,2	7,6	
	20	5	3,5	4,2	11,8	**E – sufficient**
	21	8	5,6	6,7	18,5	
	22	8	5,6	6,7	25,2	
	23	14	9,9	11,8	37,0	
	24	8	5,6	6,7	43,7	**D – satisfactory**
	25	8	5,6	6,7	50,4	
	26	10	7,0	8,4	58,8	
	27	8	5,6	6,7	65,5	**C – good**
	28	13	9,2	10,9	76,5	
	29	8	5,6	6,7	83,2	
	30	9	6,3	7,6	90,8	**B – very good**
	31	2	1,4	1,7	92,4	
	32	4	2,8	3,4	95,8	**A – excellent**
	33	5	3,5	4,2	100,0	
	Sub-totale	**119**	**83,8**	**100,0**		
	Totale	**142**	**100,0**			

Nella tabella 8 le percentuali sono calcolate dapprima sul totale delle prove effettuate e nella colonna accanto ricalcolate considerando solo le prove che hanno raggiunto la sufficienza. Sulla distribuzione delle percentuali cumulate relative alle prove superate sono stati indicati i valori che individuano i quartili della distribuzione di cui si è visto nel paragrafo precedente[9].

[8] Il fatto che si passi con una certa uniformità da un valore all'altro non deve trarre in inganno e fare apparire gli intervalli come uguali. Si tratta infatti, come è ovvio, di una progressione logaritmica (in base 10).

[9] Nella tabella appaiono alcuni valori superiori a 30 trentesimi: tali punteggi, risultanti dalla procedura adottata per il calcolo dell'esito del test, possono poi essere trasformati nella valutazione finale di 30 e lode.

4. Individuare discontinuità nella distribuzione

Forse meno usuale, ma a nostro avviso ugualmente importante, è un'ultima strategia che vorremmo suggerire: si osserva la distribuzione di frequenza assoluta o percentuale e si cerca di sezionarla tagliando dove c'è una rarefazione di casi, in modo da evitare, per quanto possibile, che troppi casi che condividono una situazione simile finiscano per appartenere a classi diverse, per quanto contigue.

Come si fa però ad individuare queste rarefazioni? In teoria è molto semplice ma, come abbiamo evidenziato in precedenza, i programmi di elaborazione dati, come Spss, al momento di compilare la tabella della distribuzione di frequenza, elidono le categorie che non presentano casi (cioè, prive di referenti empirici). Perciò bisogna ispezionare attentamente la distribuzione per accorgersi che ci sono valori mancanti, e può essere difficile apprezzare a prima vista dove vi sia maggiore rarefazione. Certamente è più facile effettuare questo controllo se il campo di possibile variazione dei dati è ristretto, o almeno noto. Ciò accade, ad esempio, quando è conosciuto l'intervallo di età della popolazione alla quale è stato somministrato un determinato questionario.

Più difficile è scorgere l'eventuale frattura nella distribuzione quando la misurazione è effettuata con precisione e, di fatto, non esiste un numero di valori ragionevolmente finito che possiamo attenderci di trovare nella distribuzione. Se stiamo studiando i redditi ricavati dalla dichiarazione dell'Irpef di una serie di soggetti, alcune decine di migliaia di numeri diversi potrebbero apparire nella distribuzione di frequenza, ma la maggior parte di essi, di fatto, non figurerà nella distribuzione e diventa difficile apprezzare dove siano collocate eventuali fratture.

In pratica, ci si trova di fronte a due diverse situazioni:

1 - la variabile cardinale è costituita da un numero relativamente piccolo di valori; spesso si tratta di valori interi, in quanto la proprietà è costituita da stati enumerabili. In tali casi è facile identificare i valori mancanti: la frattura spesso consisterà non tanto in una mancanza di valori nella distribuzione, quanto nel fatto che per determinati valori la frequenza osservata è più bassa, o molto più bassa rispetto alle frequenze limitrofe.

2 - la variabile cardinale è costituita da un numero molto elevato di valori, in genere con cifre decimali perché la definizione operativa prevede la misurazione precisa di una variabile i cui stati sulla proprietà sono continui: tendenzialmente ad ogni valore corrisponde un solo caso e dunque serve a poco osservare la distribuzione delle frequenze dei singoli valori. Per frattura in questo caso si intende un intervallo molto più elevato del solito tra due valori successivi nella serie.

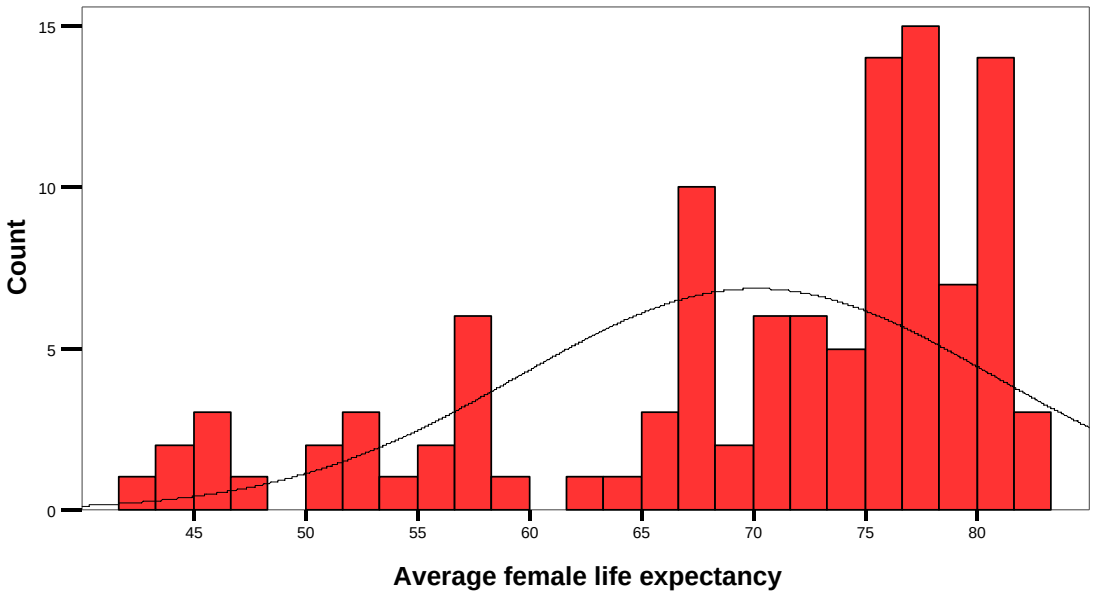
La frattura in quest'ultima circostanza può essere identificata calcolando la distanza tra due valori successivi, ad esempio, mediante un foglio elettronico (come abbiamo già suggerito e mostrato nella seconda colonna della tabella 7). Altrimenti - non c'è bisogno di inventare nulla - c'è una tecnica grafica di rappresentazione (l'*istogramma*) nata per questo scopo: studiare la forma della distribuzione di frequenza. Della forma possono fare parte anche eventuali fratture nella distribuzione, in quanto non sempre accade ciò che gli statistici (ed anche i sociologi) auspicano, e cioè che la distribuzione per lo meno approssimi la normalità, cioè una forma a campana con una concentrazione di casi verso il centro e una loro rarefazione verso gli estremi della distribuzione stessa. Se, ad esempio, possiamo riscontrare che la distribuzione dei dati ha una forma bimodale, ecco individuata una frattura.

Perché questa preoccupazione per la necessità di tagliare seguendo le fratture naturali? Si tratta della stessa preoccupazione, se vogliamo, che hanno i tagliatori di diamanti i quali cercano, per quanto possibile, di tagliare la pietra grezza seguendo le fratture naturali già presenti nella pietra. Detto altrimenti, bisogna cercare di non dividere ciò che la natura ha unito; se mai, è opportuno dividere ciò che nei fatti è già separato. Ai fini dell'analisi non è conveniente che vengano assegnati a due diversi gruppi casi che presentano valori molto simili sulla

proprietà che stiamo rilevando e finiscano nello stesso gruppo casi con valori assai distanti l'uno dall'altro. Si deve invece realizzare una segmentazione della distribuzione tesa a rendere massima la diversità (media) tra i gruppi e minima la diversità (dei casi) all'interno dei gruppi. Quanto bene siamo riusciti ad ottenere questo scopo può essere misurato calcolando la media e lo scarto tipo (deviazione standard) per le diverse categorie, una volta che siano state costruite e verificando se una segmentazione diversa della variabile potrebbe portare ad una maggiore variabilità tra (*between*) le categorie, rispetto alla varianza interna (*within*) alle categorie.

Vista la difficoltà di individuare le fratture guardando alla distribuzione di frequenza sotto forma di tabella, in questo caso inizieremo mostrando (fig. 3) per primo l'istogramma della distribuzione, che rappresenta l'aspettativa di vita delle femmine alla nascita nei paesi compresi nell'archivio World95.

Fig. 3 – Aspettativa media di vita delle femmine



Esaminando l'istogramma è possibile individuare diverse fratture che separano gruppi di paesi con aspettative di vita piuttosto difformi e raggruppano invece paesi con valori simili[10]. Un piccolo gruppo di paesi ha aspettative di vita intorno ai 45 anni, poi c'è in gruppo di paesi con un'aspettativa di vita intorno ai 55 anni in cui, se i casi fossero più numerosi, si potrebbe effettuare un'ulteriore suddivisone. Su valori più elevati si può notare un'altra concentrazione poco al di sotto dei 70 anni, un'altra poco oltre i 75 e, infine, un ultimo "picco" intorno agli 80 anni che serve ad individuare il gruppo di paesi più fortunati.

Il grafico deve essere utilizzato come orientamento generale, e solo l'esame dei valori riportati nella tabella permetterà di identificare con maggiore precisione i valori da adottare come confini di classe. L'istogramma infatti viene costruito automaticamente e, specialmente se il computer decide di costruirlo utilizzando pochi intervalli, alcune fratture possono essere

[10] All'istogramma è stata sovrapposta la curva normale che mostra quale sarebbe stata la forma della distribuzione se l'aspettativa di vita nei 109 paesi si distribuisse normalmente: numerosi paesi con un'aspettativa di vita intorno ai 70 anni ed un calo progressivo verso gli estremi della distribuzione. La distribuzione empirica è asimmetrica per la presenza di alcuni paesi non in grado di assicurare condizioni igienico sanitarie sufficienti a garantire la sopravvivenza, ma anche per l'impossibilità, con l'attuale livello di conoscenze scientifiche e di prevenzione dei comportamenti a rischio, di superare agevolmente la soglia degli 80 anni.

occultate o create artificialmente. Si dovrà perciò "pilotare" la realizzazione dell'istogramma, fissando manualmente il campo di variazione ed il numero di micro-intervalli da rappresentare mediante le colonne dell'istogramma e magari procedere con diversi tentativi, per prova ed errore, controllando a vista la stabilità delle fratture individuate.

Tab. 9 – Aspettativa media di vita delle femmine in 109 paesi

Valori	N	%	% cum.
43	1	,9	,9
44	2	1,8	2,8
45	2	1,8	4,6
46	1	,9	5,5
47	1	,9	6,4
50	2	1,8	8,3
52	2	1,8	10,1
53	1	,9	11,0
54	1	,9	11,9
55	2	1,8	13,8
57	2	1,8	15,6
58	4	3,7	19,3
59	1	,9	20,2
63	1	,9	21,1
64	1	,9	22,0
65	2	1,8	23,9
66	1	,9	24,8
67	5	4,6	29,4
68	5	4,6	33,9
69	2	1,8	35,8
70	5	4,6	40,4
71	1	,9	41,3
72	3	2,8	44,0
73	3	2,8	46,8
74	5	4,6	51,4
75	9	8,3	59,6
76	5	4,6	64,2
77	6	5,5	69,7
78	9	8,3	78,0
79	7	6,4	84,4
80	7	6,4	90,8
81	7	6,4	97,2
82	3	2,8	100,0
Totale	**109**	**100,0**	

Esaminando la tabella 9 si osserva che, per individuare le fratture nella distribuzione, si deve tenere conto del relativo affollamento di casi intorno ad alcuni valori e anche della mancanza di alcuni valori nella distribuzione. Ad esempio, il primo gruppo di valori viene individuato da un confine tracciato al di sotto del valore 50 non tanto a causa delle frequenze, che per tutti i valori circostanti sono sempre pari a 1 o 2 casi, quanto per la mancanza dei valori 48 e 49. Situazione analoga si realizza per il confine di classe successivo, che sfrutta la mancanza dei valori 60, 61 e 62. Nel caso del confine di classe tracciato al di sotto del valore 77 si utilizza invece il minimo relativo costituito dai valori contigui 76 e 77 che hanno frequenze pari a 5 e 6 casi, mentre i valori 75 e 78 presentano frequenze più elevate (9 casi).

Si noti che, procedendo per intervalli di valori uguali e scegliendo cifre tonde (decine di anni), in questo esempio la suddivisione avrebbe separato molti paesi con aspettativa di vita simile: infatti, con intervalli di valori decennali, che si includa il valore 80 nella classe "81-80" oppure nella classe "80 e più", si sarebbero separati valori contigui che hanno la stessa frequenza.

Ovviamente, procedendo in quest'ultimo modo, le classi possono presentare una numerosità assai dissimile; ma se lo scopo prioritario è quello di creare classi omogenee, il problema deve essere considerato secondario. Se invece si preferisce creare classi con un numero di casi uguale, o almeno simile, si possono usare i quartili che, come in precedenza, sono stati evidenziati in azzurro (grigio chiaro) nella tabella 9. In questo modo la prima classe si estende dal valore 43 al 67 (con una differenza di 24 anni di aspettativa di vita) e invece la terza classe abbraccia solo i valori da 75 a 78 anni (con una differenza di 4 anni). Ciò si traduce in una variabilità interna alle classi molto elevata, da evitare se non si vuole perdere in capacità discriminatoria della variabile: è evidente che fa molta differenza, sotto diversi aspetti, vivere in paesi dove l'aspettativa di vita è di soli 43 o 44 anni, rispetto a paesi dove le donne in media possono attendersi di vivere fino a 65-66 anni.

Gli stessi dati, rappresentati nella figura 3 mediante un istogramma e poi riproposti nella distribuzione di frequenza della tabella 9, sono raffigurati nella figura 4 mediante il grafico ramo-foglia che, come detto, è stato introdotto da Tukey. Si tratta di un tipo di grafico non

molto conosciuto ma, a conferma della sua utilità, si consideri che è stato inserito tra le opzioni grafiche di Spss all'interno della procedura "analizza/statistiche descrittive/esplora". Si tratta di un uso assai ingegnoso degli stessi valori che costituiscono la distribuzione di frequenza al fine di realizzare una raffigurazione grafica che assomiglia molto all'istogramma, con le barre disposte in orizzontale anziché in verticale. Come nel caso dell'istogramma, anche con il grafico ramo-foglia vengono costruite automaticamente classi di valori di uguale ampiezza, ma con il vantaggio che i valori registrati su ognuno dei casi possono essere direttamente letti sul grafico.

Nel grafico che proponiamo notiamo innanzitutto che vengono raggruppati in un'unica categoria, e denominati come estremi, i valori pari o inferiori a 50 anni; è poi possibile leggere direttamente dal grafico che vi sono due paesi che registrano un'aspettativa di vita media delle femmine di 52 anni e un altro paese dove essa è pari a 53 anni. All'altro estremo della distribuzione si legge che vi sono tre paesi con un'aspettativa di vita pari a 83 anni e, poco sotto un gruppo di 14 paesi, per 7 dei quali il valore è pari a 81 anni e per altri 7 pari a 80. Si individuano poi abbastanza agevolmente le fratture nella distribuzione, per quanto in questo caso una di esse sia nascosta all'interno del gruppo dei casi estremi.

Fig. 4 – Grafico ramo-foglia (*stem-and-leaf*) dell'aspettativa media di vita delle femmine in 109 paesi

```
Frequency    Stem &  Leaf

     9 Extremes    (=<50)
            5 .
     3      5 .  223
     3      5 .  455
     2      5 .  77
     5      5 .  88889
            6 .
     1      6 .  3
     3      6 .  455
     6      6 .  677777
     7      6 .  8888899
     6      7 .  000001
     6      7 .  222333
    14      7 .  44444555555555
    11      7 .  66666777777
    16      7 .  8888888889999999
    14      8 .  00000001111111
     3      8 .  222

Stem width:    10
Each leaf:       1 case(s)
```

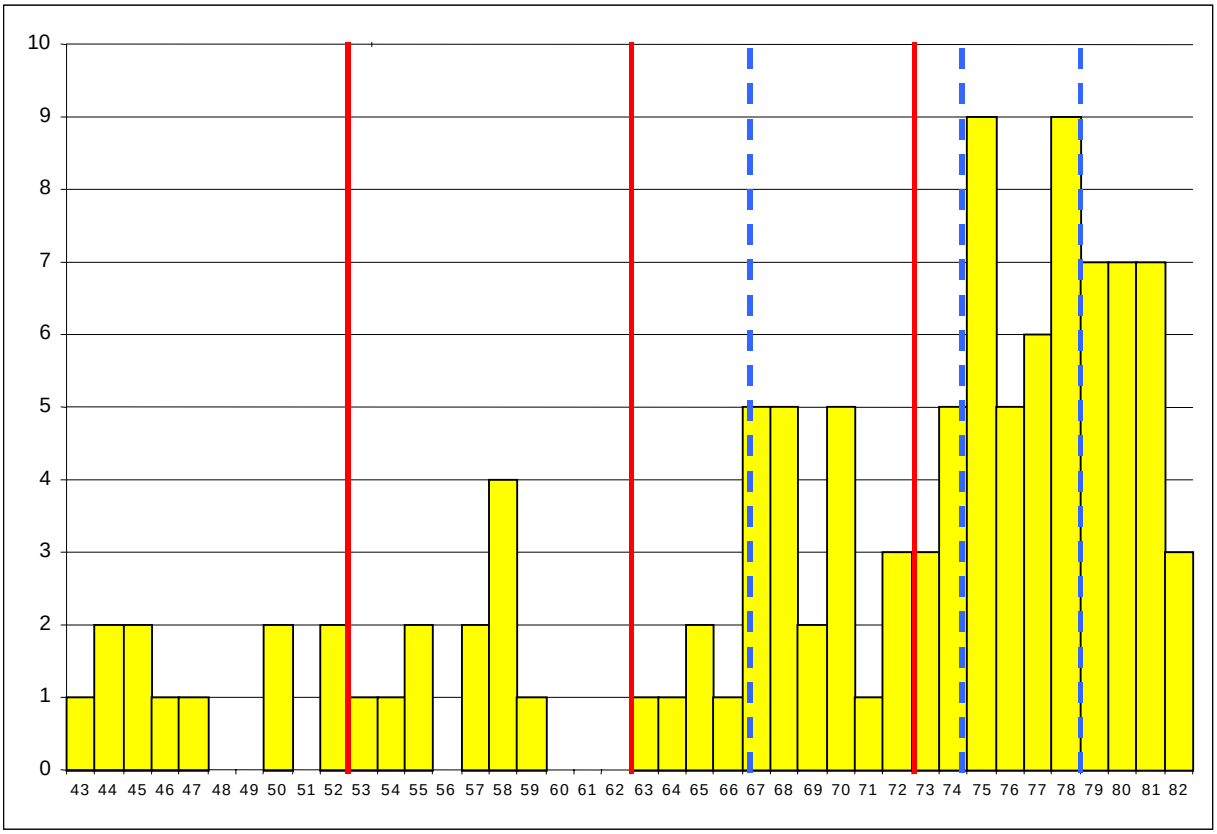
5. Intervalli uguali e uguale numerosità a confronto

Per concludere, vorremmo mostrare con maggiore precisione gli effetti della scelta tra due tecniche per certi versi opposte: uguali intervalli di valori oppure uguale numerosità dei casi. Per rendere meglio confrontabili gli effetti abbiamo deciso di realizzare con entrambe le tecniche un pari numero di classi. Partendo dal valore minimo riscontrato nella distribuzione (43 anni) e sfruttando il fatto che il valore massimo (82) consente una ripartizione senza dare resto, sono state infatti realizzate quattro classi decennali che vengono confrontate con le quattro classi create utilizzando i quartili della distribuzione.

Per visualizzare la posizione dei tagli sulla distribuzione utilizzando le due tecniche, abbiamo riproposto il grafico della figura 3, realizzato questa volta mediante un grafico a barre

verticali, una tecnica solo apparentemente uguale all'istogramma. Ricordiamo che l'istogramma crea automaticamente classi di ampiezza uguale e le barre sono accostate, per sottolineare che è stata ridotta in classi una variabile con la quale si sono rilevati valori di una proprietà continua. Vengono realizzate le classi perché spesso le variabili cardinali presentano una distribuzione di frequenza con i singoli valori rilevati in un unico caso. Ciò si è visto, ad esempio, nella tabella 5: se la distribuzione originaria venisse rappresentata senza un'aggregazione in classi dei valori, il diagramma avrebbe mostrato barre di altezza unitaria, ad eccezione di due soli paesi per i quali il pnl pro capite era pari a 1.500 dollari, e perciò sarebbe risultato poco utile a mostrare la forma della distribuzione.

Fig. 5 – Aspettativa media di vita delle femmine



Anche la variabile qui analizzata in origine presentava probabilmente una distribuzione di frequenza simile a quella evidenziata nella tabella 5: un solo caso per ogni valore (il calcolo dell'aspettativa di vita media aveva certamente prodotto decimali che poi sono stati arrotondati all'anno intero). Ecco perché vi sono più paesi con la stessa aspettativa di vita, il che consente di realizzare il diagramma con barre che non abbiano tutte un'altezza unitaria.

Un'altra differenza consiste nel fatto che, mentre con l'istogramma viene rappresentato l'intero campo di variazione della variabile e visualizzati gli eventuali intervalli privi di osservazioni, i diagrammi a barre vengono costruiti dai programmi di elaborazione statistica standard mostrando solo i valori che presentano una frequenza non nulla. Abbiamo perciò integrato la distribuzione con gli anni mancanti, come è stato fatto nella tabella 1, e le barre sono state accostate per sottolineare la contiguità tra i valori[11].

Graficamente si vede immediatamente la differenza tra le due tecniche: le linee verticali rosse sono ugualmente spaziate ed invece la prima linea verticale azzurra tratteggiata che se-

[11] Per rendere nulla la distanza, in Excel, dopo avere selezionato le barre che rappresentano la serie dei dati, deve essere attivata la sequenza "formato/serie dei dati/opzioni" e impostata sul valore "0" la "distanza tra le barre".

gna il confine del primo quartile è posizionata sui 67 anni, ben oltre la metà del campo di variazione della variabile.

Tutto ciò, come detto, ha delle conseguenze sulla composizione dei gruppi di paesi che vengono costruiti; nella tabella 10 mostriamo alcune statistiche descrittive per illustrare le differenze tra queste due e un'ulteriore modalità di segmentazione che verrà illustrata nell'ultimo paragrafo. Nell'aggregazione in classi decennali l'intervallo di valori (detto anche *range* o campo di variazione) dovrebbe essere per definizione sempre lo stesso e pari a 9, ma nella seconda classe esso è pari a 6 in quanto, come avevamo notato, mancano nella distribuzione gli anni 60, 61 e 62. Il numero dei casi è invece molto diversificato; nell'ultima classe è pari a oltre la metà del totale. L'aspettativa di vita media nelle classi si innalza di circa 10 anni da una classe all'altra con poca diversificazione all'interno, a parte la seconda classe, anche a causa di un intervallo di valori meno ampio.

Tab. 10: Aspettativa di vita della femmine aggregata in classi decennali e in quartili

	Intervallo	N	Media	Scarto tipo
Classi decennali				
43-52	9	11	47,1	3,33
53-62	6	11	56,5	1,97
63-72	9	26	68,2	2,42
73-82	9	61	77,6	2,58
Totale	**39**	**109**	**70,2**	**10,57**
Classi in quartili				
43-66	23	27	54,2	7,10
67-74	7	29	70,3	2,57
75-78	3	29	76,5	1,24
79-82	3	24	80,3	1,03
Totale	**39**	**109**	**70,2**	**10,57**
Classi basate su media e scarto tipo				
43-60	17	22	51,8	5,53
61-70	9	22	67,6	2,02
71-80	9	55	76,5	2,48
81-82	1	10	81,3	0,48
Totale	**39**	**109**	**70,2**	**10,57**

Invece, nella suddivisione della variabile in quartili la prima classe è molto ampia, abbracciando un intervallo di valori di 23 anni, mentre le due ultime classi sono comprese all'interno di intervalli di soli tre anni (differenza tra il valore minimo e massimo della classe). Di conseguenza, rispetto alla distribuzione precedente, l'aspettativa media di vita tra le classi è molto più diversificata, e così pure lo scarto tipo.

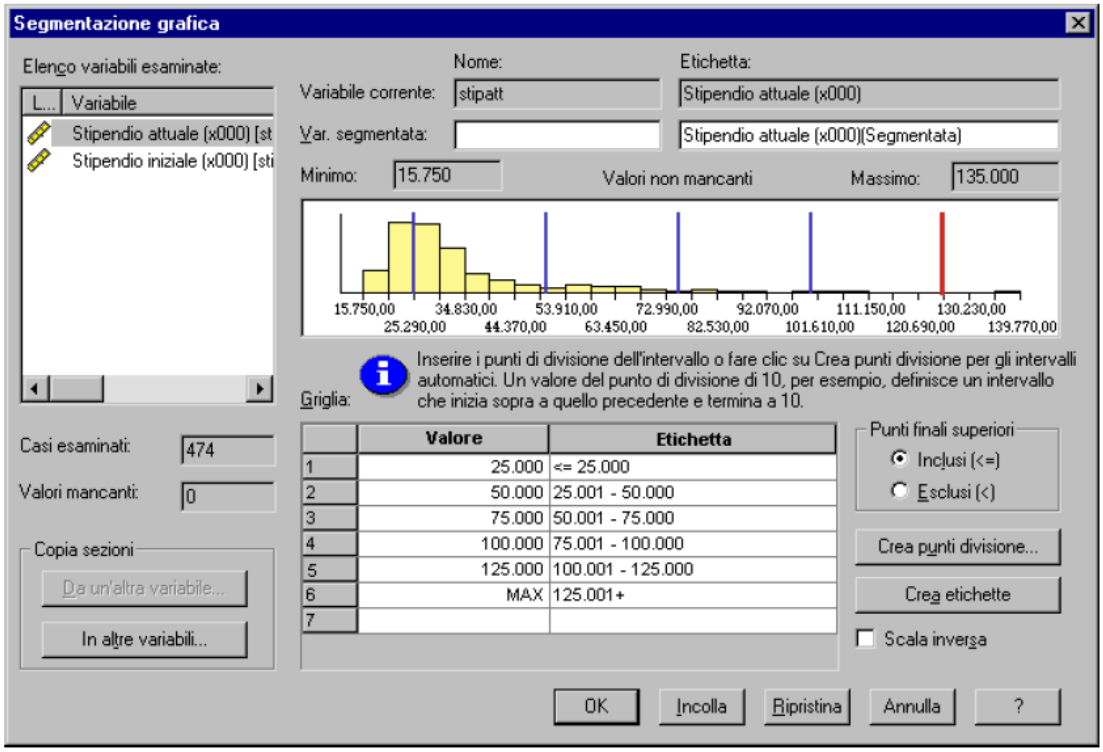
In compenso, le classi hanno dimensioni simili, ma non uguali, perché più casi condividono i valori che costituiscono i punti di taglio della divisione in quartili. Anziché suddividere questi paesi arbitrariamente affinché i conti tornino, è opportuno assegnarli tutti ad una delle due classi contigue in modo da minimizzare la differenza tra l'esito finale e l'obiettivo di costruire classi con lo stesso numero di casi.

Conclusioni: come "fare a fette" con Spss

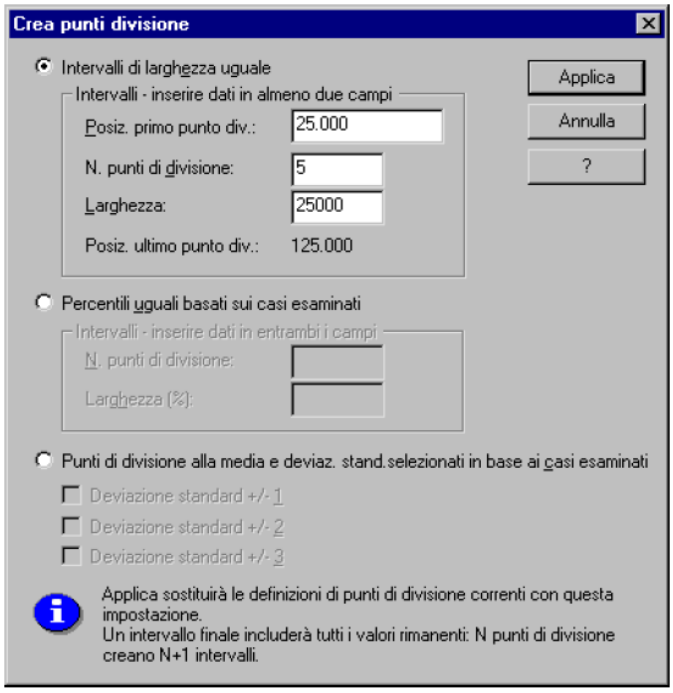
L'esperienza ed il ragionamento hanno guidato chi scrive nell'individuazione delle quattro opzioni fondamentali tra cui scegliere quella più adatta per realizzare la segmentazione delle variabili.

Un importante riscontro che le opzioni si possono ricondurre a quelle qui esposte e, inoltre, che le tecniche grafiche sono un supporto fondamentale per guidare la scelta tra le opzioni per

realizzare la segmentazione, proviene da Spss che ha introdotto nelle più recenti versioni un modulo denominato "segmentazione grafica". Nella videata iniziale infatti viene presentato l'istogramma della distribuzione della variabile selezionata, sul quale sono tracciate barre verticali che mostrano la posizione dei punti di taglio a seconda della modalità di segmentazione usata.



Selezionando il bottone "crea punti di divisione", si accede a una finestra di dialogo che consente di scegliere se segmentare creando intervalli di larghezza uguale oppure percentili uguali: le prime due opzioni descritte nel presente lavoro.



Come si può notare, è presente una terza opzione che in questa sede non avevamo considerato e che vale la pena di commentare in quanto per certi aspetti costituisce una quinta strategia di segmentazione. Questa opzione è simile alla scelta di segmentare sulla base dei valori posizionali (si basa pur sempre sulla distribuzione di frequenza empirica dei casi), ma utilizza la media e la deviazione standard, invece di valori posizionali come la mediana e i quartili.

Come risultato avremo la creazione di gruppi contenenti un numero diseguale di casi che dovrebbe essere simmetrico intorno alla media. Per definizione, infatti, nell'intervallo individuato dalla distanza di uno scarto tipo (deviazione standard) in più o in meno della media è contenuto il 68% circa dei casi. Ciò avviene però solo se i valori sono distribuiti normalmente, e invece la distribuzione empirica della variabile "aspettativa di vita" è fortemente asimmetrica; ecco perché le classi create con questo metodo nel nostro caso (vedi tab. 10) sono molto lontane dalla ripartizione teorica che dovrebbe vedere raggruppati circa il 34% dei casi nelle due classi centrali e il 16% in entrambe le altre due.

Visto che le variabili rilevate nelle ricerche sociali sono spesso molto lontane dalla teorica distribuzione normale, riteniamo che perciò sia di gran lunga preferibile la scelta di segmentare utilizzando valori posizionali della distribuzione quali la mediana e i quartili. L'istogramma visualizzato nella finestra del modulo di segmentazione grafica consente comunque di verificare immediatamente mediante il posizionamento delle barre verticali l'effetto delle scelte effettuate e rende più facile aiutare a prendere una decisione.

L'istogramma ovviamente consente anche di individuare le fratture o discontinuità presenti nella distribuzione e di collocare in corrispondenza di questi valori le barre di suddivisione, spostandole manualmente. Si realizza così facilmente anche questa strategia che abbiamo delineato e, allo stesso modo, le barre verticali possono essere collocate manualmente in corrispondenza di valori definiti "soglia" dalle consuetudini, dalla legislazioni o da altri criteri.

Infine, una volta presa la decisione riguardo ai valori della distribuzione da utilizzare per la segmentazione, il programma consente, mediante il bottone "crea etichette" di contrassegnare automaticamente i valori della nuova variabile segmentata.