

STATISTICA PER LE SCIENZE SOCIALI



il Mulino

**GEORGE W. BOHRNSTEDT
DAVID KNOKE**

TAB. 2.4. Numero di morti per incidenti stradali registrati nel 1986 negli Stati Uniti

Stato	Morti	Stato	Morti
Distretto di Columbia	86	Washington	763
Alaska	98	Wisconsin	769
Vermont	105	Massachusetts	773
North Dakota	120	Mississippi	785
Hawaii	124	Maryland	796
South Dakota	152	Kentucky	829
Rhode Island	155	Louisiana	987
Delaware	160	New Jersey	1.059
Wyoming	163	Arizona	1.073
New Hampshire	181	South Carolina	1.077
Maine	218	Indiana	1.077
Montana	240	Virginia	1.141
Idaho	273	Alabama	1.180
Nevada	274	Missouri	1.205
Nebraska	311	Tennessee	1.372
Utah	370	Georgia	1.604
Iowa	446	Ohio	1.609
Connecticut	448	Illinois	1.633
West Virginia	470	Michigan	1.666
Kansas	519	North Carolina	1.727
New Mexico	539	Pennsylvania	1.968
Minnesota	608	New York	2.183
Arkansas	624	Florida	2.925
Oregon	649	Texas	3.701
Colorado	657	California	5.523
Oklahoma	725		

(maggioranza relativa). Alcune distribuzioni sono dette *bimodali*, cioè possiedono due mode. In senso stretto, in una distribuzione bimodale le due categorie più numerose devono avere esattamente lo stesso numero di osservazioni. Tuttavia, in pratica questa perfetta uguaglianza si verifica solo raramente e, pertanto, gli scienziati usano il termine *bimodale* per denotare qualsiasi distribuzione nella quale esistono due categorie che contengono il maggior numero di casi e hanno una numerosità molto simile.

In generale, qualsiasi statistica applicabile alle variabili discrete è valida anche nell'analisi delle variabili continue. Pertanto, la moda è una misura di tendenza centrale applicabile a entrambi i tipi di variabili.

4.2. La mediana

La **mediana** si applica solo alle variabili le cui categorie possono essere ordinate in modo crescente. Espressamente, la mediana è quel valore che divide una distribuzione ordinata esattamente in due metà, in modo tale che la metà dei casi osservati abbiano valori inferiori a quello mediano, e l'altra metà abbiano valori superiori. L'esistenza di un caso che corrisponda direttamente al valore mediano dipende dal numero di casi osservati. Per fare un esempio, si considerino le seguenti distribuzioni:

6. Percentili e quantili

Un'utile statistica che può essere ricavata dalle distribuzioni cumulate è il **percentile**, inteso come il valore al di sotto del quale è compresa una data percentuale di casi. Per esempio, nella tabella 2.3 il valore «14 anni di istruzione formale» corrisponde al 73° percentile; ciò significa che il 73% degli intervistati sono rimasti a scuola al massimo 14 anni. La mediana equivale al 50° percentile e, nella tabella 2.3, corrisponde al valore «12 anni di istruzione formale».

Quando i dati sono aggregati in categorie, il calcolo dei percentili diventa piuttosto complicato e richiede una conoscenza dei limiti reali e dei punti centrali, argomenti che discuteremo nel prossimo paragrafo.

6.1. Limiti reali e punti centrali

Quando i valori relativi a una determinata distribuzione vengono aggregati in categorie di varia ampiezza, sorge il problema di determinare i limiti reali e i punti centrali delle categorie stesse. A fini illustrativi, i valori della distribuzione riportata nella tabella 2.4 sono stati aggregati in una serie di intervalli di una certa ampiezza (cfr. tab. 2.9). Precisamente, come si può desumere osservando la prima colonna della tabella 2.9, ogni intervallo ha un'ampiezza pari a 800 decessi. I limiti di ciascun intervallo sono definiti da numeri interi (ad esempio 1-800), tuttavia i **limiti reali** dell'intervallo sono (nell'esempio specifico) 0,5-800,5. Sono questi i numeri che definiscono esattamente il limite inferiore e quello superiore dei valori che potrebbero rientrare nella categoria (o intervallo) in questione. Si noti che i limiti reali coprono interamente ciascun intervallo, in modo tale da garantire che nella distribuzione non ci siano buchi o aree scoperte. La seconda colonna della tabella 2.9 riporta i limiti reali dei sette intervalli creati per questo esempio.

Il **punto centrale** di un intervallo è ottenuto sommando i limiti reali di quell'intervallo e dividendo la somma per due. Ad esempio, il punto centrale dell'intervallo 0,5 – 800,5 è pari a $(0,5 + 800,5)/2 = 400,5$.

TAB. 2.9. Distribuzione dei morti per incidenti stradali registrati nel 1986 negli Stati Uniti

Numero di morti	Limiti reali	Punto centrale	<i>f</i>	<i>p</i>	<i>fc</i>
1-800	0,5-800,5	400,5	31	0,61	31
801-1.600	800,5-1.600,5	1.200,5	10	0,20	41
1.601-2.400	1.600,5-2.400,5	2.000,5	7	0,14	48
2.401-3.200	2.400,5-3.200,5	2.800,5	1	0,02	49
3.201-4.000	3.200,5-4.000,5	3.600,5	1	0,02	50
4.001-4.800	4.000,5-4.800,5	4.400,5	0	0,00	50
4.801-5.600	4.800,5-5.600,5	5.200,5	1	0,02	51

I punti centrali di tutti gli altri intervalli sono riportati nella terza colonna della tabella 2.9. In sostanza, *il punto centrale è il numero singolo che rappresenta in modo migliore un dato intervallo*.

Una volta che sono stati stabiliti gli intervalli per l'aggregazione, si conta il numero di casi che appartengono a ciascuna categoria (quarta colonna della tab. 2.9). Infine, le frequenze così ottenute sono trasformate in proporzioni (colonna 5) e in frequenze cumulate (colonna 6).

6.2. Il calcolo dei percentili

Quando i dati sono aggregati (come nella tab. 2.9), i percentili vengono calcolati utilizzando la seguente formula:

$$P_i = L_p + \left(\frac{(p_i)(N) - fc_p}{f_p} \right) (W_i)$$

dove:

P_i = valore della distribuzione corrispondente all' i^{imo} percentile

L_p = limite reale inferiore dell'intervallo che contiene l' i^{imo} percentile

p_i = l' i^{imo} percentile in forma di proporzione

N = numero totale di osservazioni

fc_p = frequenza cumulata fino all'intervallo che contiene P_i (escluso)

f_p = frequenza dell'intervallo che contiene l' i^{imo} percentile

W_i = ampiezza dell'intervallo che contiene P_i ($W_i = U_p - L_p$)

U_p = limite reale superiore dell'intervallo che contiene l' i^{imo} percentile.

Nella tabella 2.9 il 90° percentile è calcolato come segue:

$$P_{90} = 1.600,5 + \left(\frac{(0,90)(51) - 41}{7} \right) (800) = 2.160,5$$

in quanto $i = 90$, $L_p = 1.600,5$, $N = 51$, $fc_p = 41$, $f_p = 7$ e $W_i = 800$.

La mediana di questa stessa distribuzione, calcolata sui dati individuali (cfr. tab. 2.4), risulta pari a 725, valore corrispondente al 26° caso. Se calcolata sui dati aggregati riportati nella tabella 2.9, invece, la mediana – corrispondente al 50° percentile – è pari a:

$$P_{50} = 0,5 + \left(\frac{(0,50)(51) - 0}{31} \right) (800) = 658,6$$

Questo esempio mostra chiaramente che l'aggregazione dei dati

può dare luogo a errori di descrizione non trascurabili, specialmente se la distribuzione originale è fortemente asimmetrica. Dunque, quando si calcolano statistiche come la media, la mediana o la varianza è sempre meglio utilizzare i dati nella forma più disaggregata possibile.

6.3. I quantili

I percentili sono casi speciali dei **quantili**, che dividono una data distribuzione in categorie formate da un'uguale proporzione di casi. I quantili più utilizzati sono i quartili, i quintili e i decili.

I **quartili** sono i valori di una distribuzione continua che dividono le osservazioni in *quattro categorie di uguale numerosità*. Q_1 è il valore al di sotto del quale si trovano un quarto delle osservazioni, Q_2 è il valore al di sotto del quale si trovano la metà delle osservazioni, e così via. I **quintili** dividono le osservazioni in *cinque categorie di uguale numerosità*, mentre i **decili** le dividono in *dieci categorie di uguale numerosità*. Come abbiamo già detto, i percentili dividono le osservazioni in *cento categorie di uguale numerosità*.

Come abbiamo già visto, il simbolo P_i è utilizzato per indicare l' i^{mo} percentile. Analogamente, D_i denota l' i^{mo} decile, K_i l' i^{mo} quintile e Q_i l' i^{mo} quartile. Utilizzando questa notazione, si ha che $Q_1 = P_{25}$, cioè il primo quartile equivale esattamente al 25° percentile. Analogamente, $K_1 = D_2$, cioè il primo quintile è uguale al secondo decile. Il quadro 2.3 riassume in una tabella le relazioni esistenti fra i vari quantili e i percentili.

QUADRO 2.3. Relazioni fra quartili, quintili, decili e percentili

Le relazioni esistenti fra quartili, quintili e decili da un lato, e percentili dall'altro, possono essere espresse in forma sinottica nella seguente tavola:

Quartili	Percentili	Quintili	Percentili	Decili	Percentili
Q_1	P_{25}	K_1	P_{20}	D_1	P_{10}
Q_2	P_{50}	K_2	P_{40}	D_2	P_{20}
Q_3	P_{75}	K_3	P_{60}	D_3	P_{30}
		K_4	P_{80}	D_4	P_{40}
				D_5	P_{50}
				D_6	P_{60}
				D_7	P_{70}
				D_8	P_{80}
				D_9	P_{90}

quanto non esistono decine) e la cifra secondaria, dopo l'arrotondamento del decimale, è pari a 7. Nel caso del Minnesota e del Montana, caratterizzati da un tasso di partecipazione elettorale del 62,3%, la cifra primaria è invece pari a 6, mentre quella secondaria equivale a 2.

Un diagramma rami e foglie assomiglia, in una certa misura a un istogramma appoggiato di lato. I «rami» della figura 2.4 corrispondono alle varie cifre primarie, ordinate in colonna dalla più bassa alla più alta; a destra di ciascun «ramo» sono situate le rispettive «foglie», ognuna delle quali rappresenta la cifra secondaria corrispondente a una data osservazione. Il «ramo» zero possiede solo una «foglia» perché vi è un solo stato (la Louisiana) che esibisce un tasso di partecipazione elettorale inferiore al 10%. I «rami» 1 e 2 non possiedono, invece, alcuna foglia, indicando che in nessuno stato si è registrato un tasso di partecipazione elettorale compreso fra 10 e 29%. I «rami» compresi fra 3 e 6, infine, possiedono un numero variabile di «foglie». Al «ramo» 6, ad esempio, sono attaccate cinque «foglie», una per ciascuno degli stati caratterizzati da un tasso di partecipazione elettorale uguale o superiore al 60%. La figura 2.4 mostra chiaramente che la distribuzione dei tassi di partecipazione elettorale relativi al 1988 possiede una leggera asimmetria negativa. Rispetto all'istogramma, il diagramma rami e foglie offre un vantaggio: consente al lettore di cogliere con maggiore precisione le informazioni sulle singole osservazioni che compongono la distribuzione.

Una seconda tecnica esplorativa per l'analisi e la rappresentazione dei dati continui è costituita dal **diagramma box-and-whisker**, più semplicemente denominato **boxplot**. Il *boxplot* mostra contemporaneamente la tendenza centrale, la variabilità e la forma di una data distribuzione. Per costruire un *boxplot* è necessario conoscere l'ampiezza della distribuzione, la sua mediana e i suoi «cardini» (*hinges*). A scopo illustrativo, utilizzeremo ancora una volta i dati sui decessi per incidenti stradali (cfr. tab. 2.4) al fine di chiarire questi concetti.

Il **cardine inferiore**, indicato con il simbolo H_1 , è il valore associato all'osservazione che divide il quartile inferiore di una distribuzione ordinata in senso crescente dai tre quarti superiori. Nel nostro caso, il quartile inferiore corrisponde alla 13^{ima} osservazione, lo stato dell'Idaho, al quale è associato un valore pari a 273 decessi. Dunque, $H_1 = 273$. Il **cardine superiore**, indicato con il simbolo H_2 , corrisponde invece al valore dell'osservazione che divide il quartile superiore di una distribuzione ordinata in senso crescente dai tre quarti inferiori. Nel nostro caso, il quartile superiore corrisponde alla 39^{ima} osservazione, lo stato dell'Alabama, al quale è associato un valore pari a 1.180 decessi. Dunque, $H_2 = 1.180$. Nelle distribuzioni con un ampio numero di casi il primo e il terzo quartile (Q_1 e Q_3) possono essere utilizzati per identificare rispettivamente H_1 e H_2 .

Per costruire il *boxplot* bisogna innanzitutto disegnare un rettangolo i cui lati inferiore e superiore devono essere tracciati all'altezza

dei due «cardini», come mostrato nella figura 2.5. Quindi, bisogna stabilire se nella distribuzione sono presenti **valori estremi** (*outliers*). Questi valori sono quelli che, perché troppo elevati o troppo bassi, si collocano ai margini della distribuzione, in un certo senso separati da essa. Un modo per identificare i valori estremi consiste nel determinare, per ciascuna osservazione, se si colloca al di sopra del **margine interno superiore** (MIS) o al di sotto del **margine interno inferiore** (MII). Per calcolare questi due margini bisogna prima determinare lo **spazio fra i cardini** (HS), che non è altro che la differenza fra il cardine superiore e quello inferiore: $HS = H_2 - H_1$. Nel nostro esempio, $HS = 1.180 - 273 = 907$. Disponendo di questa misura possiamo ora

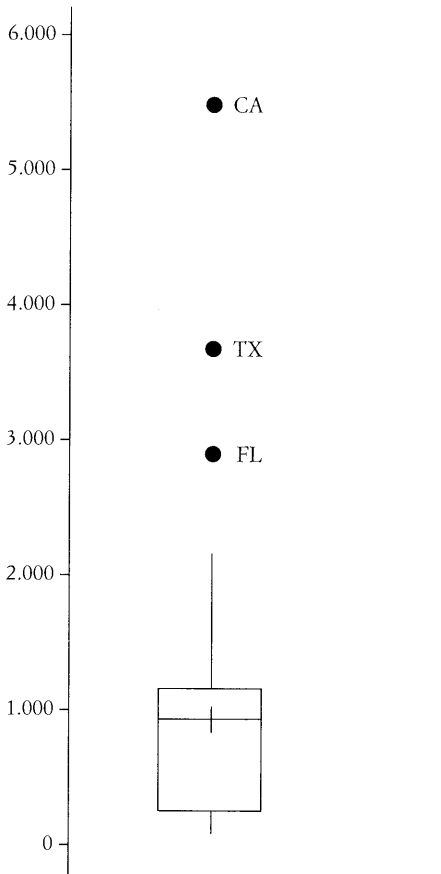


FIG. 2.5. *Boxplot* raffigurante la distribuzione dei morti per incidenti stradali registrati nel 1986 negli Stati Uniti.

calcolare i due margini interni e determinare quindi se la distribuzione contiene dei valori estremi:

$$MII = H_1 - (1,5)(HS)$$

e

$$MIS = H_2 + (1,5)(HS)$$

Tornando al nostro esempio, $MI = 273 - (1,5)(907) = -1.807,5$ e $MIS = 1.180 + (1,5)(907) = 2.540,5$. Chiaramente, il margine interno inferiore è minore di qualsiasi valore possibile, in quanto nessuno stato può avere un numero negativo di incidenti stradali mortali! Dunque, non esistono valori estremi nella parte inferiore della distribuzione. Tuttavia, esistono tre *outliers* nella parte superiore della distribuzione: la Florida, il Texas e la California. In un *boxplot*, i valori estremi sono indicati mediante simboli accompagnati da etichette che riportano i nomi dei casi corrispondenti.

Il *boxplot* è completato da due *whiskers* (linee): la prima unisce il cardine superiore (H_2) al valore più grande della distribuzione che non è un *outlier*; la seconda, invece, unisce il cardine inferiore (H_1) al valore più piccolo della distribuzione che non è un *outlier*.

Il *boxplot* relativo ai decessi per incidenti stradali mostra chiaramente che *a)* la distribuzione presenta un certo grado di asimmetria positiva; *b)* gli stati della Florida, del Texas e della California rappresentano in modo evidente dei valori estremi.

Nel quarto capitolo vedremo come è possibile utilizzare i diagrammi a rami e foglie e i *boxplot* per confrontare due distribuzioni.

Problemi

1. Costruisci una distribuzione di frequenze a partire dal seguente insieme di valori:

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

2. Costruisci una tavola di frequenze assolute, relative e percentuali utilizzando il seguente insieme di nazionalità europee, dove F = Francia, G = Germania, S = Spagna, I = Italia, R = Russia e A = Altri paesi:

F S F S I G I R A R I F S A S F S G I R