ISTITUZIONI DI STATISTICA – A. A. 2007/2008
Marco Minozzo e Annamaria Guolo
Laurea in Economia del Commercio Internazionale
Laurea in Economia e Amministrazione delle Imprese
Università degli Studi di Verona (sede di Vicenza)

Statistica Descrittiva
Soluzioni 4. Medie lasche

Introduzione

Consideriamo X un carattere le cui modalità x_i, $i = 1, \ldots, I$, hanno frequenza assoluta f_i e relativa p_i. Date N misurazioni del carattere sulla popolazione, è possibile calcolare le medie lasche rappresentate da moda, mediana e quantili della distribuzione di X.
La **moda** (o norma) corrisponde alla modalità del carattere che si presenta con la massima frequenza. È valutabile per caratteri qualitativi sia sconnessi sia ordinali e per caratteri quantitativi, sia discreti sia continui.
La **mediana**, invece, può essere valutata per caratteri almeno ordinali. Poste le modalità di X in ordine crescente, la mediana è data dal valore osservabile per X che occupa la posizione centrale della distribuzione, cioè che lascia alla sua destra ed alla sua sinistra metà delle osservazioni effettuate.

- Si consideri una successione ordinata di osservazioni da X, con modalità non definite in classi. Se N è un numero dispari, la mediana corrisponde al valore allocato nella posizione $(N + 1)/2$. Nel caso N sia pari, invece, la mediana è un qualunque valore compreso tra i termini in posizione $N/2$ e $N/2 + 1$.

- Sia X un carattere con modalità definite per classi. Allora il calcolo della mediana procede prima di tutto dall'individuazione della classe mediana. In seguito, si cerca la mediana di interesse all'interno della classe individuata. Indicando con $[x_{i-1}, x_i)$ la classe mediana di frequenza assoluta f_i, relativa p_i e relativa cumulata P_i, la mediana è data da

$$m_e = x_{i-1} + (0,5 - P_{i-1}) * ((x_i - x_{i-1})/p_i).$$

Un'estensione del concetto di mediana è rappresentata dal calcolo dei **quantili** della distribuzione di X. Un generico quantile di ordine α è quel valore osservabile per X che lascia alla propria sinistra una quota di osservazioni pari a $\alpha\%$ ed alla propria destra una quota di osservazioni pari a $(1 - \alpha)\%$. I quantili più noti sono i **quartili**: primo quartile ($\alpha = 0.25$), secondo quartile ($\alpha = 0.50$, mediana), terzo quartile ($\alpha = 0.75$).

Esercizio A.
a) Per il calcolo delle medie richieste si consideri la tabella seguente dove con x_i si sono indicati i valori centrali delle classi e dove l'estremo inferiore e quello superiore sono stati posti pari a 12 e 50 rispettivamente,
La classe modale è la classe con massima frequenza. Essendo il carattere di interesse con modalità che sono espresse in classi di ampiezza d_i, è opportuno valutare la moda tenendo conto della densità di frequenza per classi. Tale densità è misurata da h_i, $h_i = p_i/d_i$. Quindi, la moda è quindi la classe $[25 - 29)$ a cui corrisponde una densità di 0,0636.
Consideriamo il calcolo della mediana. Individuiamo prima di tutto la classe mediana. Si tratta della classe $[25 - 30)$. Ora ricerchiamo il valore mediano all'interno della classe mediana, sotto l'ipotesi di uniforme distribuzione all'interno delle classi. La mediana è data da

$$m_e = x_{i-1} + (0,5 - P_{i-1})/p_i,$$

Classi di età	< 15	15–19	20–24	25–29	30–34	35–39	> 39	
chiusura classi	$[12-15)$	$[15-20)$	$[20-25)$	$[25-30)$	$[30-35)$	$[35-40)$	$[40-50)$	
f_i	0	4	6	7	5	0	0	22
p_i	0	0,1818	0,2727	0,3182	0,2273	0	0	1
P_i	0	0,1818	0,4545	0,7727	1	1	1	–
d_i	3	5	5	5	5	5	10	
h_i	0	0,03636	0,0545	0,0636	0,0455	0	0	
x_i	13,5	17,5	22,5	27,5	32,5	37,5	45	
$\ln x_i$	2,603	2,862	3,114	3,314	3,481	3,624	3,807	

vale a dire $m_e = 25 + (0,5 - 0,4545)/0,0636 = 25,71$.
La media aritmetica è data da

$$m = \sum_{i=1}^{7} x_i f_i/22 = \sum_{i=1}^{7} x_i p_i = 13,5 \cdot 0 + 17,5 \cdot 0,1818 + \cdots + 45 \cdot 0 = 25,45.$$

La media geometrica risulta invece

$$m = (x_1^{f_1} \cdot \cdots \cdot x_7^{f_7})^{1/22} = \exp\left(\frac{1}{22}\sum_{i=1}^{7} f_i \cdot \ln x_i\right) = 24,91.$$

b) Tra le medie calcolate per sintetizzare la distribuzione in oggetto l'unica media non appropriata è la media geometrica.
c) Per i maschi della ULSS 22 si consideri la tabella: In questo caso la media aritmentica è pari a 34,75,

Età	< 15	15–19	20–24	25–29	30–34	35–39	> 39	
chiusura	$[12-15)$	$[15-20)$	$[20-25)$	$[25-30)$	$[30-35)$	$[35-40)$	$[40-50)$	
f_i	0	12	63	83	117	108	151	534
p_i	0	0,022	0,118	0,1554	0,2191	0,2022	0,2828	1
P_i	0	0,022	0,140	0,2954	0,5145	0,7167	1	–
x_i	13,5	17,5	22,5	27,5	32,5	37,5	45	

mentre considerando che la classe mediana è $[30-35)$, la mediana è pari a $m_e = 30 + (0,5 - 0,2954) \cdot ((35-30)/0,2191) = 34,67$.

Esercizio B.
Si considera la distribuzione di frequenza seguente per il carattere di interesse

	Maschi
Dottorato	141
Diploma 4-5 anni	382
Diploma 2-3 anni	166
Licenza media	478
Licenza elementare	108
Totale	1275

La moda è data dalla modalità che si presenta con massima frequenza, vale a dire *Licenza media*.
Nel caso di carattere qualitativi la mediana può essere calcolata se il carattere in questione è almeno ordinale. In questo esempio lo è, dato che il titolo di studio segue un ordinamento logico. Per determinare la mediana disponiamo le modalità in ordine crescente e calcoliamo la distribuzione di frequenza assoluta cumulata F_j, come somma della frequenza assoluta della modalità j di riferimento e delle precedenti

	Maschi	F_j
Licenza elementare	108	108
Licenza media	478	478+108=586
Diploma 2-3 anni	166	752
Diploma 4-5 anni	382	1134
Dottorato	141	1275
Totale	1275	

Si consideri che $N = 1275$, dispari. Perciò la mediana si trova in corrispondenza della posizione $(N + 1)/2 = 638$. La mediana è *Diploma 2-3 anni*.

Esercizio C.
I dati in questione sono relativi alla distribuzione del tipo di museo nella provincia di Verona. Sulla base delle frequenze assolute si ricava che la moda è *Museo specializzato*. Essendo il carattere qualitativo sconnesso non è possibile calcolarne la mediana.

Esercizio D.
a) Ponendo l'estremo superiore dell'ultima classe pari a 100 anni non compiuti si ha

chiusura classi	$[0, 1)$	$[1, 5)$	$[5, 10)$	$[10, 15)$	$[15, 25)$	$[25, 45)$	$[45, 65)$	$[65, 100)$	Totale
f_i	6392	24696	34605	36251	94687	238052	216283	179748	830714
p_i	0,0077	0,0297	0,0416	0,0436	0,1140	0,2866	0,2604	0,2164	1
P_i	0,0077	0,0374	0,0790	0,1226	0,2366	0,5232	0,7836	1	—
d_i	1	4	5	5	10	20	20	35	—
h_i	0,0077	0,0074	0,0083	0,0087	0,0114	0,0143	0,0130	0,0061	—

L'istogramma di frequenza si ottiene rappresentando per ogni classe un rettangolo di larghezza pari a d_i e altezza pari a h_i.
b) L'intervallo $[30, 50)$ coinvolge due classi e quindi, sotto l'ipotesi di uniforme distribuzione, ad esso corrisponde una frequenza relativa teorica pari a $0,0143(45 - 30) + 0,0130(50 - 45) = 0,2795$ a cui corrisponde una frequenza assoluta di $0,2795 \cdot 830714 = 232184,6$. La mediana cade nella classe $[25, 45)$ e quindi è pari a

$$m_e = 25 + 20(0,5 - 0,2366)/0,2866 = 43,38.$$

Il primo e terzo quartile si calcolano in modo analogo, ma facendo riferimento a quei valori che lasciano alla propria sinistra un ammontare di unità statistiche pari al 25% e al 75%, rispettivamente, a differenza della mediana che ne lascia a sinistra una quota pari al 50%. Dunque, il primo quartile cade nella classe $[25, 45)$ ed è pari a

$$Q_1 = 25 + 20(0,25 - 0,2366)/0,2866 = 25,94,$$

mentre il terzo quartile cade nella classe $[45, 65)$ ed è pari a

$$Q_3 = 45 + 20(0,75 - 0,5232)/0,2604 = 62,42.$$

Esercizio E.
a) Fissando l'altezza minima e massima rispettivamente pari a 150 cm e a 199 cm, ed indicando con x_i i valori centrali di classe, per il Veneto si ha
da cui si vede che la classe modale è la classe 170–179, per cui la mediana è pari a

$$m_e(V) = 169,5 + 10(0,5 - 0,171)/0,55 = 175,48.$$

chiusura classi	[149,5–159,5)	[159,5–164,5)	[164,5–169,5)	[169,5–179,5)	[179,5–184,5)	[184,5–189,5)	[189,5–199,5)
x_i	154,5	162	167	174,5	182	187	194,5
d_i	10	5	5	10	5	5	10
p_i	0,008	0,036	0,127	0,550	0,187	0,071	0,002
$h_i \cdot 100$	0,08	0,72	2,54	5,50	3,74	1,42	0,21
$P_i \cdot 100$	0,8	4,4	17,1	72,1	90,8	97,9	100,0
$\ln x_i$	5,040	5,088	5,118	5,162	5,204	5,231	5,270

La media aritmetica è invece pari a

$$m_1(V) = \sum_{i=1}^{7} p_i x_i = 175,65,$$

mentre la media geometrica è data da

$$m_0(V) = \exp(\sum p_i \ \ln x_i) = \exp(5,0677) = 158,80.$$

b) I quantili richiesti si possono ottenere con la formula

$$x(\alpha) = c_{h-1} + d_h(\alpha - P_{h-1})/p_h,$$

dove $P_{h-1} < \alpha < P_h$. Perciò si ottiene: per il primo quartile

$$Q_1 = 169,5 + 10(0,25 - 0,171)/0,55 = 170,94;$$

per il terzo quartile

$$Q_3 = 179,5 + 5(0,75 - 0,721)/0,187 = 180,28;$$

per il primo decile

$$D_1 = 164,5 + 5(0,1 - 0,044)/0,127 = 166,71;$$

per il nono decile

$$D_9 = 179,5 + 5(0,9 - 0,721)/0,187 = 184,29.$$

c) Per sintetizzare la distribuzione della statura degli iscritti con una media opportuna si possono utilizzare sia la mediana che la media aritmetica. Non è invece appropriato usare la media geometrica.
d) Procedendo come al punto a), per la regione Sicilia si ottiene una media aritmetica pari a $m_1(S) = 171,81$. Inoltre, essendo le frequenze percentuali cumulate pari a 3,0, 13,0, 37,2, 89, 97,6, 99,6, 100,0, la classe mediana è la classe $[169,5 - 179,5)$ e la mediana è pari a

$$m_e(S) = 169,5 + 10(0,5 - 0,372)/0,518 = 171,97.$$

Si può notare che mentre la distribuzione per il Veneto ha una leggera asimmetrica positiva e si ha $m_1(V) > m_e(V)$, la distribuzione per la Sicilia ha una leggera asimmetrica negativa e si ha $m_1(S) < m_e(S)$.

Esercizio F.
Sia a il punto in cui costruire il nuovo magazzino. La distanza di ogni negozio dal punto a si valuta come $|x_i - a|$. Lo scopo è cercare il punto a in modo che la somma delle distanze dei negozi da A, cioè $|55 - a| + |90 - a| + |120 - a| + |135 - a| + |156 - a|$ sia minima. È noto che il valore che minimizza una tale quantità è la mediana. Con i dati a disposizione, la mediana è $a = 120km$.

Esercizio G.
a) Per la funzione di ripartizione del numero di esami sostenuti si consideri la seguente tabella:

Numero esami	f_i	p_i	P_i
1	58	0,3867	0,3867
2	42	0,2800	0,6667
3	12	0,0800	0,7467
4	38	0,2533	1
	150	1	–

Considerando che il carattere è discreto, la funzione di ripartizione è una funzione a gradini ed è definita da

$$F(x) = \begin{cases} 0, & x < 1, \\ 0,3867, & 1 \leq x < 2, \\ 0,6667, & 2 \leq x < 3, \\ 0,7467, & 3 \leq x < 4, \\ 1, & x \geq 4. \end{cases}$$

b) Il primo, il secondo (ovvero la mediana) ed il terzo quartile sono dati rispettivamente da $Q_1 = 1$, $Q_2 = m_e = 2$, $Q_3 = 4$.