Università di Cassino
Corso di Statistica 1
Esercitazione del 17/10/2006
Dott. Alfonso Piscitelli

Esercizio 1

Il seguente *data set* riporta la rilevazione di alcuni caratteri su un collettivo di 20 soggetti.

Soggetto	Sesso	Età	Reddito (Migliaia di €)	Titolo di studio	Nucleo familiare	Statura (cm)	Colore degli occhi
1	M	22	0,7	Diploma	3	173	NERO
2	F	18	0,2	Lic. Media	4	168	MARRONE
3	F	34	1,6	Diploma	2	165	MARRONE
4	M	42	2,5	Laurea	5	180	NERO
5	F	50	3,2	Diploma	3	163	AZZURRO
6	F	12	0,1	Lic. Elementare	4	160	NERO
7	M	46	3,8	Lic. Media	4	177	MARRONE
8	M	72	1,3	Nessun Titolo	2	164	VERDE
9	F	27	1,2	Laurea	3	158	AZZURRO
10	F	48	1,7	Lic. Media	5	170	NERO
11	F	35	1,9	Laurea	1	167	NERO
12	M	84	0,8	Nessun Titolo	1	159	MARRONE
13	F	21	0,4	Diploma	5	174	AZZURRO
14	F	44	1,8	Diploma	4	164	VERDE
15	M	56	1,9	Lic. Media	2	177	NERO
16	F	58	3,2	Lic. Media	3	172	NERO
17	F	37	2,1	Diploma	1	166	MARRONE
18	F	16	0,1	Lic. Media	4	160	MARRONE
19	M	73	1,6	Lic. Elementare	2	170	AZZURRO
20	M	64	2,2	Lic. Elementare	3	184	VERDE

a) Determinare la moda per i caratteri **Colore degli occhi** e **Titolo di studio**.
b) Determinare la media aritmetica e la mediana per il carattere **Statura**.
c) Esprimere il carattere **Reddito** in 3 modalità ordinali (basso, medio e alto) secondo la seguente corrispondenza:

basso	(fino a 1,3 Migliaia di €)
medio	(da 1,4 a 2,6 Migliaia di €)
alto	(da 2,7 a 3,8 Migliaia di €)

costruire per il nuovo carattere ottenuto la distribuzione di frequenza e determinarne la moda e la mediana.

d) Determinare la media, la mediana ed i quartili per il carattere **Nucleo familiare** a partire sia dalla successione di valori sia dalla distribuzione di frequenze.

e) Costruire la distribuzione di frequenza per il carattere **Età** suddividendo la distribuzione in 4 classi equiampie e determinare la classe modale, la classe mediana e dei quartili e la media aritmetica.

f) Costruire la distribuzione doppia di frequenza per i caratteri **Sesso** e **Età**, utilizzando per questo ultimo carattere la suddivisione in classi operata precedentemente.

Soluzioni

a) La **moda** è quel valore della variabile X associato alla frequenza più alta, in altre parole è l'intensità (o la modalità, nel caso di variabili qualitative) che si presenta il maggior numero di volte. Nel nostro caso è necessario da prima calcolarsi le distribuzioni di frequenza delle variabili **Colore degli occhi** e **Titolo di studio** e solo dopo sarà possibile individuare le rispettive mode.

La Moda del carattere **Colore degli occhi** è:
"Nero"

Colore degli occhi	n_i
Azzurro	4
Marrone	6
Nero	7
Verde	3
	Tot: 20

Titolo di studio	n_i
Nessun Titolo	2
Lic. Elementare	3
Lic. Media	6
Diploma	6
Laurea	3
	Tot: 20

La Moda del carattere **Titolo di studio** è:
"Lic. Media e Diploma"
(Bimodale)

b) Nella tabella sono riportati i valori della variabile **Statura**, ordinate in modo NON DECRESCENTE.
La media aritmetica di un insieme di N valori osservati $x_1, x_2, ..., x_N$ di un carattere quantitativo X è pari alla somma dei valori osservati divisa per il loro numero: $\mu = \frac{1}{N}(x_1 + x_2 + ... + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$.

$\mu = 168{,}55$

Soggetto	Posizione	Statura (cm)
9	1	158
12	2	159
6	3	160
18	4	160
5	5	163
8	6	164
14	7	164
3	8	165
17	9	166
11	10	167
2	11	168
10	12	170
19	13	170
16	14	172
1	15	173
13	16	174
7	17	177
15	18	177
4	19	180
20	20	184

La mediana è quel valore che suddivide la distribuzione, ordinata in modo decrescente o non decrescente, in due parti uguali lasciando quindi il 50% delle osservazioni alla sua destra e il restante 50% alla sua sinistra.
Nel caso di N pari, cioè di un numero di osservazioni pari, la mediana è data dalla media aritmetica del valore che occupa la (N/2)-esima posizione e il successivo, cioè quello che occupa la (N/2)+1-esima. Nel nostro caso 20/2=10 ⇒ media tra la 10° e l'11°posizione.

Me = 167,5

Nel caso in cui i valori siano espressi in una distribuzione di frequenze, per calcolare la mediana bisogna considerare la frequenza relativa cumulata. Il valore mediano sarà il primo che supera il valore 0,50.

c) La distribuzione di frequenza dalla variabile **Reddito** suddivisa nelle 3 modalità ordinali indicate dal testo dell'esercizio è la seguente:

Reddito	n_i	f_i	F_i
Basso	8	0,4	0,4
Medio	9	0,45	0,85
Alto	3	0,15	1
Tot	**20**		

da cui si nota che la moda è la modalità "Medio".
Quindi, la maggior parte del nostro collettivo guadagna un reddito compreso tra 1,4 e 2,6 Migliaia di €.
Per individuare la mediana si fa riferimento alla distribuzione di frequenze cumulate relative. Individueremo il valore mediano come quello associato alla prima frequenza cumulata relativa che supera lo 0,50. Nel nostro caso la mediana è la classe di reddito [1,4 e 2,6 Migliaia di €], in altre parole, la modalità "Medio".

d) La successione dei valori ordinati in senso non decrescente e la corrispondente distribuzione di frequenza della variabile **Nucleo familiare** sono le seguenti:

Soggetto	Posizione	Nucleo familiare
11	1	1
12	2	1
17	3	1
3	4	2
8	5	2
15	6	2
19	7	2
1	8	3
5	9	3
9	10	3
16	11	3
20	12	3
2	13	4
6	14	4
7	15	4
14	16	4
18	17	4
4	18	5
10	19	5
13	20	5

Nucleo familiare	n_i	f_i	F_i
1	3	0,15	0,15
2	4	0,2	0,35
3	5	0,25	0,60
4	5	0,25	0,85
5	3	0,15	1
Tot:	**20**	**1**	

La media delle variabile **Nucleo familiare** è logicamente la stessa sia nel caso della distribuzione per frequenze che nel caso della successione dei valori. La differenza è nel modo di calcolarla. Nel caso della distribuzione di frequenza:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i * n_i$$

In entrambi i casi la media è $\mu = 3{,}05$

Il primo quartile corrisponde a quel valore del carattere X che lascia alla sua sinistra il 25% delle osservazioni e alla sua destra il 75%.
Il secondo quartile o mediana è quel valore che suddivide la distribuzione ordinata (in modo crescente o decrescente) in due parti lasciando il 50% delle osservazioni alla sua destra e il restante 50% alla sua sinistra.

Il terzo quartile corrisponde a quel valore del carattere X che lascia alla sua sinistra il 75% delle osservazioni e alla sua destra il rimanente 25%.

$$Q_1 = \frac{X_{\frac{N}{4}} + X_{\frac{N}{4}+1}}{2} = \frac{X_5 + X_6}{2} = \frac{2+2}{2} = 2$$

$$Q_2 = M_e = 3$$

$$Q_3 = \frac{X_{3*\frac{N}{4}} + X_{3*\frac{N'}{4}+1}}{2} = \frac{X_{15} + X_{16}}{2} = \frac{4+4}{2} = 4$$

Nel caso delle distribuzioni di frequenza semplice, invece, i quartili vengono individuati facendo riferimento alle frequenze cumulate o alle frequenze relative cumulate. In questo caso:

. ⇒ il primo quartile è quel valore della x associato alla prima frequenza relativa cumulata maggiore di 0,25. [Q_1=2]

. ⇒ il secondo quartile o mediana è quel valore della x associato alla prima frequenza relativa cumulata maggiore di 0,50. [Q_2=Me=3]

. ⇒ il terzo quartile è quel valore della x associato alla prima frequenza relativa cumulata maggiore di 0,75. [Q_3=4].

e) La distribuzione in classi di frequenza del carattere **Età**, è:

Età	d_i	n_i	X_i^c	f_i	F_i	h_i
12-|30	18	6	21	0,3	0,3	0,33
30-|48	18	6	39	0,3	0,6	0,33
48-|66	18	5	57	0,25	0,85	0,28
66-|84	18	3	75	0,15	1,0	0,17
Tot:		**20**		**1**		

Per individuare la classe modale è necessario fare riferimento alle frequenze assolute n_i. In questo caso, vi sono due classi modali: la prima è [12; 30] mentre la seconda è [30; 48].

La media di una distribuzione in classi si trova dividendo la somma dei prodotti tra il valore centrale di ogni classe e la frequenza della classe per il numero totale di osservazioni. Nel nostro caso:

$$\mu = \frac{1}{N}\sum_{i=1}^{C} x_i^c * n_i$$

n_i	X_i^c	(n_i * X_i^c)
6	21	126
6	39	234
5	57	285
3	75	225
20		870

$$\mu = \frac{870}{20} = 43{,}5$$

Possiamo concludere affermando che l'età media del nostro collettivo è pari a 43,5 anni.
La classe Mediana, così come pure quelle dei quartili, si individuano esattamente come fatto nel punto precedente (*cfr* punto **d**). Avremo quindi che:
. ⇒ Classe Q_1=12-|30 [valore della x associato alla prima frequenza relativa cumulata maggiore di 0,25].
. ⇒ Classe Q_2=Me=30-|48 [valore della x associato alla prima frequenza relativa cumulata maggiore di 0,50].
. ⇒ Classe Q_3=48-|66 [valore della x associato alla prima frequenza relativa cumulata maggiore di 0,75].

f) Per rappresentare la distribuzione doppia di frequenze dei due caratteri **Sesso** e **Età** (suddiviso in classi) è necessario costruirsi una tabella a doppia entrata che ha per righe le due modalità della variabile **Sesso** {Maschio, Femmina} e per colonne le 4 classi in cui è stata suddivisa la variabile **Età** {12-|30; 30-|48; 48-|66; 66-|84 }.

Si precisa che la stessa informazione si avrebbe da una tabella che ha per righe le classi della variabile **Età** e per colonne le modalità della variabile **Sesso**.

La distribuzione doppia di frequenza delle due variabili è:

	12-|30	30-|48	48-|66	66-|84	Tot:
Maschio	1	2	2	3	8
Femmina	5	4	3	0	12
Tot:	6	6	5	3	20