Università di Cassino
Corso di Statistica 1
Esercitazione del 28/01/2008
Dott. Alfonso Piscitelli

Esercizio 1

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

Soggetto	Età	Residenza	Reddito (Migliaia di €)	Auto Possedute	Punteggio quiz	Km giornalieri percorsi
1	22	Cantagallo	0,7	3	173	25
2	18	Cantagallo	0,2	4	168	25
3	34	Poggio a Caiano	1,6	2	165	21
4	42	Carmignano	2,5	5	180	25
5	50	Poggio a Caiano	3,2	3	163	17
6	12	Montemurlo	0,1	4	160	23
7	46	Carmignano	3,8	4	177	26
8	72	Montemurlo	1,3	2	164	35
9	27	Montemurlo	1,2	3	158	26
10	48	Carmignano	1,7	5	170	30
11	35	Montemurlo	1,9	1	167	21
12	84	Cantagallo	0,8	1	159	25
13	21	Montemurlo	0,4	5	174	26
14	44	Carmignano	1,8	4	164	33
15	56	Carmignano	1,9	2	177	24
16	58	Montemurlo	3,2	3	172	29
17	37	Cantagallo	2,1	1	166	14
18	16	Montemurlo	0,1	4	160	23
19	73	Carmignano	1,6	2	170	21
20	64	Poggio a Caiano	2,2	3	184	20

a. Determinare la moda per il carattere **Comune di Residenza**.

b. Determinare la media aritmetica per il carattere **Punteggio quiz** a partire dai dati grezzi.

c. Determinare la moda e la media per il carattere **Auto Possedute** a partire sia dalla successione di valori sia dalla distribuzione di frequenze.

d. Costruire la distribuzione di frequenza per il carattere **Punteggio quiz** suddividendo la distribuzione in 4 classi equiampie e determinare la classe modale e la media aritmetica.

e. Verificare la presenza di valori anomali per il carattere **Km giornalieri percorsi**.

f. Calcolare la media troncata al 20% per il carattere **Km giornalieri percorsi**.

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 la distribuzione di frequenza della variabile **Comune di Residenza** e solo dopo sarà possibile individuare la moda.

La Moda del carattere **Comune di Residenza** è:
"Montemurlo"

Comune di Residenza	n_i
Cantagallo	4
Carmignano	6
Montemurlo	7
Poggio a Caiano	3

b) Nella tabella sono riportati i valori della variabile **Punteggio quiz**, ordinate in modo NON DECRESCENTE.

Soggetto	**Posizione**	**Punteggio quiz**
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 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: $\mu = \frac{1}{N}(x_1 + x_2 + ... + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$.

$$\mu = 168{,}55$$

c) La successione dei valori ordinati in senso non decrescente e la corrispondente distribuzione di frequenza della variabile **Auto Possedute** sono le seguenti:

Soggetto	Posizione	Auto Possedute
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

Auto Possedute	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 e la moda della variabile **Auto Possedute** è logicamente la stessa sia nel caso della distribuzione per frequenze che nel caso della successione dei valori. Per quanto riguarda la media aritmetica, 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 $\mu = 3{,}05$

La **moda** è quel valore della variabile **Auto Possedute** associato alla frequenza più alta.
Nel nostro esercizio vi sono due intensità ("3" e "4") associate al valore di frequenza assoluta più alto ("5").
Per cui la distribuzione della variabile **Auto Possedute** è bimodale

d) La distribuzione in classi di frequenza del carattere **Punteggio quiz**, è:

Punteggio quiz	**d_i**	**n_i**	X_i^c	**f_i**	**F_i**	**h_i**
158 -|164,5	6,5	7	161,3	0,35	0,35	1,08
164,5-|171	6,5	6	167,8	0,3	0,65	0,92
171 -|177,5	6,5	5	174,3	0,25	0,9	0,77
177,5-|184	6,5	2	180,8	0,1	1	0,31
Tot:		**20**		**1**		

Per individuare la classe modale è necessario fare riferimento alla densità di frequenza (h_i). In questo caso, la classe modale è: la prima [158; 164,5], ossia quella associata alla massima densità di frequenza.

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)
7	161,3	1128,75
6	167,8	1006,5
5	174,3	871,25
2	180,8	361,5
20		3368

$$\mu = \frac{3368}{20} = 168{,}4$$

Possiamo concludere affermando che il punteggio medio del nostro collettivo è pari a 168,4.

e) Per verificare la presenza di eventuali valori anomali bisogna ordinare il carattere **Km giornalieri percorsi**.

Soggetto	Posizione	Km giornalieri percorsi
17	1	14
5	2	17
20	3	20
3	4	21
11	5	21
19	6	21
6	7	23
18	8	23
15	9	24
1	10	25
2	11	25
4	12	25
12	13	25
7	14	26
9	15	26
13	16	26
16	17	29
10	18	30
14	19	33
8	20	35

A partire dalla distribuzione ordinata del carattere **Km giornalieri percorsi**, si calcolano i valori del primo e del terzo Quartile.

$Q_1=21$

$Q_3=26$.

Il primo e il terzo Quartile ci consentono di calcolarci dei valori definiti Limite Superiore e Limite Inferiore.

$$L_I = Q_1 - 1{,}5 * (Q_3 - Q_1)$$

Nel nostro caso L_I =13,5

Dal confronto tra L_I e X_{min} sarà possibile verificare la presenza di valori troppo piccoli. Tutti i valori più piccoli di L_I sono considerati "anomali".
Nel nostro esercizio non è presente un valore più piccolo di L_I, quindi non siamo in presenza di valori troppo bassi da essere definiti anomali.

$$L_S = Q_3 + 1{,}5 * (Q_3 - Q_1)$$

Nel nostro caso L_S =33,5

Dal confronto tra L_S e X_{max}.sarà possibile verificare la presenza di valori troppo grandi. Tutti i valori che eccedono L_S sono considerati "anomali".
Nel nostro esercizio è presente un solo valore più grande di L_S (35), quindi siamo in presenza di un valore definito "anomalo".

f) La **media troncata** è la media aritmetica calcolata su una fissata percentuale di valori centrali di un insieme di dati. Per calcolare la media troncata al 20% bisogna ordinare il carattere **Km giornalieri percorsi** ed escludere il 10% dei valori più piccoli ed il 10% dei valori più grandi.

Soggetto	Posizione	Km giornalieri percorsi
17	1	14
5	2	17
20	3	20
3	4	21
11	5	21
19	6	21
6	7	23
18	8	23
15	9	24
1	10	25
2	11	25
4	12	25
12	13	25
7	14	26
9	15	26
13	16	26
16	17	29
10	18	30
14	19	33
8	20	35

Nel nostro esercizio, la media troncata al 20% sarà ottenuta escludendo i due valori più piccoli e i due più grandi.

$$Mt_{20\%} = \frac{\sum_{i=(0,1n)+1}^{n-(0,1n)} X_i}{n-(0,2n)}$$

$$Mt_{20\%} = \frac{\sum_{i=3}^{18} X_i}{16} = \frac{390}{16} = 24,375$$

La media troncata elimina l'influenza dei valori anomali.