Università di Cassino
Corso di Statistica 1
Esercitazione del 24/10/2006
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) Costruire la distribuzione di frequenza per il carattere **Età** suddividendo la distribuzione in 4 classi equiampie e determinare il valore della mediana e dei quartili.
b) Costruire la distribuzione doppia di frequenza per i caratteri **Sesso** e **Età**, utilizzando per questo ultimo carattere la suddivisione in classi operata precedentemente. Verificare la proprietà associativa della media.
c) Determinare la differenza interquartile per il carattere **Nucleo familiare** a partire sia dalla successione di valori sia dalla distribuzione di frequenze.
d) Determinare la varianza e lo scarto quadratico medio per il carattere **Età**, utilizzando la suddivisione in classi operata precedentemente.

e) Data la distribuzione del carattere **Reddito** se ne misuri il grado di concentrazione.

Soluzioni

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

Età	n_i	N_i	f_i	d_i
12-|30	6	6	0,3	18
30-|48	6	12	0,3	18
48-|66	5	17	0,25	18
66-|84	3	20	0,15	18
Tot:	**20**		**1**	

Prima di passare al calcolo della mediana, per dati in classi, bisogna evidenziare la classe mediana. La classe mediana è quella associata alla prima frequenza cumulata relativa che supera lo 0,50.

. ⇒ Classe Me=30-|48 [in cui il valore della x associato alla prima frequenza relativa cumulata è maggiore di 0,50].

Quindi, la mediana sarà:

$$M_e = l_e + \frac{\frac{\sum n_i}{2} - N_{e-1}}{n_e} d_e$$

dove:

l_e =limite inferiore della classe Mediana;

N_{e-1} =frequenza cumulata associata alla classe precedente a quella Mediana;

n_e =frequenza assoluta della classe Mediana;

d_e =ampiezza della classe Mediana;

$$M_e = 30 + \frac{10-6}{6} 18 = 42$$

Le classi del primo e del terzo quartile, si individuano in corrispondenza delle rispettive frequenze cumulate relative. Avremo quindi che:

. ⇒ Classe Q_1=12-|30 [valore della x associato alla prima frequenza relativa cumulata maggiore di 0,25].

. ⇒ Classe Q_3=48-|66 [valore della x associato alla prima frequenza relativa cumulata maggiore di 0,75].

Quindi, il primo quartile sarà:

$$Q_1 = l_{Q_1} + \frac{\frac{\sum n_i}{4} - N_{Q_1-1}}{n_{Q_1}} d_{Q_1}$$

dove:

l_{Q_1} =limite inferiore della classe Q_1;

N_{Q_1-1} =frequenza cumulata associata alla classe precedente a quella di Q_1;

n_{Q_1} =frequenza assoluta della classe Q_1;

d_{Q_1} =ampiezza della classe Q_1;

$$Q_1 = 12 + \frac{5-0}{6} 18 = 27$$

Quindi, il terzo quartile sarà:

$$Q_3 = l_{Q_3} + \frac{3\frac{\sum n_i}{4} - N_{Q_3-1}}{n_{Q_3}} d_{Q_3}$$

dove:

l_{Q_3} =limite inferiore della classe Q_3;

N_{Q_3-1} =frequenza cumulata associata alla classe precedente a quella di Q_3;

n_{Q_3} =frequenza assoluta della classe Q_3;

d_{Q_3} =ampiezza della classe Q_3;

$$Q_3 = 48 + \frac{15-12}{5} 18 = 58{,}8$$

b) 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

In questa tabella, il calcolo della media aritmetica della variabile **Età** sarà:

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

Ricordando che i valori centrali delle classi sono rispettivamente: 21; 39; 57; 75.

$$\mu = \frac{(21*6)+(39*6)+(57*5)+(75*3)}{20} = \frac{870}{20} = 43{,}5$$

Per poter verificare la proprietà associativa della media bisogna ricorrere alle distribuzioni di frequenza condizionate. Le due distribuzioni di frequenza condizionate della variabile **Età** sono:

	12-|30	30-|48	48-|66	66-|84	Tot:
Maschio	1	2	2	3	**8**

	12-|30	30-|48	48-|66	66-|84	Tot:
Femmina	5	4	3	0	**12**

Per ognuna delle tabelle si calcola la media della distribuzione condizionata delle variabile **Età** rispetto alle variabile **Sesso**:

$$\mu_{Età|M} = \frac{(21*1)+(39*2)+(57*2)+(75*3)}{8} = \frac{438}{8} = 54{,}75$$

L'Età media dei Maschi è 54,75

$$\mu_{Età|F} = \frac{(21*5)+(39*4)+(57*3)+(75*0)}{12} = \frac{432}{12} = 36$$

L'Età media delle Femmine Maschi è 36

La proprietà associativa della Media afferma che la media delle medie condizionate ponderata per la numerosità del gruppo è uguale alla media generale. Essa è verificata dalla seguente uguaglianza:

$$\frac{\sum_{i=1}^{G} \mu_i * n_i}{\sum n_i} = \mu$$

dove G= numero di gruppi.

Nel nostro caso G=2 e la media delle medie condizionate è:

$$\mu = \frac{\sum_{i=1}^{2} \mu_i * n_i}{\sum n_i} = \frac{(54{,}57*8)+(36*12)}{20} = \frac{870}{20} = 43{,}5$$

c) 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**	

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 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%.

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

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

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

Si definisce differenza interquartile la differenza tra il terzo e il primo quartile.

$$IQR = Q_3 - Q_1 = 4-2=2$$

Questa quantità contiene il 50% "centrale" delle osservazioni.

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

Età	$\mathbf{n_i}$	X_i^c
12-|30	6	21
30-|48	6	39
48-|66	5	57
66-|84	3	75
Tot:	**20**	

La varianza è la media degli scarti dalla media elevati al quadrato; il rapporto tra la somma degli scarti dalla media al quadrato, e il numero delle osservazioni.
Nel nostro caso trattandosi di una distribuzione in classi la varianza si trova come:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{C}(x_i^c - \mu)^2 * n_i$$

Ricordando che la media è uguale a 43,5 avremo:

Età	$\mathbf{n_i}$	X_i^c	$(x_i^c - \mu)$	$(x_i^c - \mu)^2$	$(x_i^c - \mu)^2 * n_i$
12-|30	6	21	-22,5	506,25	3038
30-|48	6	39	-4,5	20,25	121,5
48-|66	5	57	13,5	182,25	911,3
66-|84	3	75	31,5	992,25	2977
Tot:	**20**				7047

$$\sigma^2 = \frac{7047}{20} = 352{,}4$$

$$\sigma = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{C}(x_i^c - \mu)^2 * n_i} = \sqrt[2]{\sigma^2} = \sqrt{352{,}4} = 18{,}77$$

e) La concentrazione di un carattere quantitativo è possibile solo se il carattere è trasferibile, cioè quando il carattere può passare da un'unità all'altra del collettivo. Il carattere **Reddito** è un carattere trasferibile e si parte dalla successione dei valori ordinati in senso non decrescente

X_i	**Reddito**
1	0,1
2	0,1
3	0,2
4	0,4
5	0,7
6	0,8
7	1,2
8	1,3
9	1,6
10	1,6
11	1,7
12	1,8
13	1,9
14	1,9
15	2,1
16	2,2
17	2,5
18	3,2
19	3,2
20	3,8

La concentrazione di un carattere si misura rispetto ad una condizione detta di equidistribuzione.
Si ha **concentrazione nulla** quando l'ammontare totale del carattere è ripartito in parti uguali tra le unità.
Si ha **concentrazione massima** quando tutto il carattere è posseduto da una sola unità, mentre (n-1) unità non possiedono nulla.

$\mathbf{p_i}$= la frazione cumulata dei primi *i* redditieri, $i=1,2,3,\ldots,n$ $\qquad p_i = \frac{i}{n}$

$\mathbf{q_i}$= la frazione cumulata del reddito posseduto dai primi *i* redditieri, $i=1,2,3,\ldots,n$

$$q_i = \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{n} x_j}$$

x_i	Reddito	Reddito Cumulato	p_i	q_i	$p_i - q_i$
1	0,1	0,1	0,05	0,003	0,047
2	0,1	0,2	0,1	0,006	0,094
3	0,2	0,4	0,15	0,012	0,138
4	0,4	0,8	0,2	0,025	0,175
5	0,7	1,5	0,25	0,046	0,204
6	0,8	2,3	0,3	0,071	0,229
7	1,2	3,5	0,35	0,108	0,242
8	1,3	4,8	0,4	0,149	0,251
9	1,6	6,4	0,45	0,198	0,252
10	1,6	8	0,5	0,248	0,252
11	1,7	9,7	0,55	0,300	0,25
12	1,8	11,5	0,6	0,356	0,244
13	1,9	13,4	0,65	0,415	0,235
14	1,9	15,3	0,7	0,474	0,226
15	2,1	17,4	0,75	0,539	0,211
16	2,2	19,6	0,8	0,607	0,193
17	2,5	22,1	0,85	0,684	0,166
18	3,2	25,3	0,9	0,783	0,117
19	3,2	28,5	0,95	0,882	0,068
20	3,8	32,3			
					3,594

Possiamo dire che c'è equidistribuzione quando $q_i = p_i$ per ogni $i=1,2,3,\ldots,n$

Un indice che misura la concentrazione è il **Rapporto di Concentrazione** (**R**) di Gini. Si tratta di un indice relativo che varia tra 0 ed 1.

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$

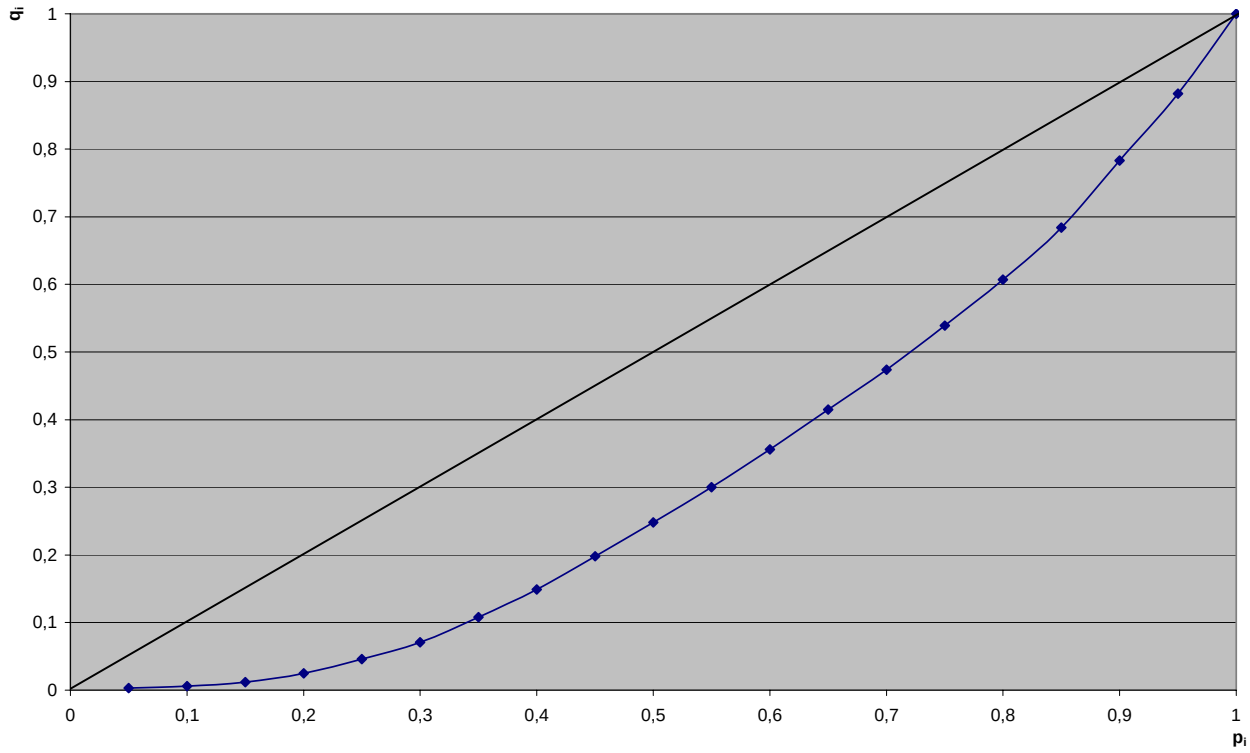
$$0 \le R \le 1$$

$$R = \frac{3{,}594}{9{,}5} = 0{,}3783$$

R=0 si ha concentrazione minima.
R=1 si ha concentrazione massima.

Una rappresentazione grafica della concentrazione può essere fatta attraverso la curva di Lorenz (curva di concentrazione), ovvero la spezzata che si ottiene

unendo i punti di coordinate (p_i ,q_i) rappresentati sul piano cartesiano.



La bisettrice rappresentata sul grafico rappresenta la situazione di equidistribuzione. L'area compresa tra la curva di concentrazione e la retta di equidistribuzione viene detta **area di concentrazione**.