Esercitazione 4 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

11 Febbraio 2011

Dataset Studenti

N	*SESSO*	*ETA'*	*PESO*	*ALTEZZA*	*DIPLOMAI*	*COMPONENTI*	*OCCHIALI*	*FUMO*
1	0	20,6	65	180	Ist.Tecnico	6	0	1
2	0	20,2	75	180	Liceo	4	0	0
3	0	20,3	60	173	Ist.Tecnico	4	1	0
4	0	23,9	93	187	Liceo	8	0	1
5	0	21,4	66	164	Ist.Tecnico	5	0	0
6	0	25	84	186	Ist.Tecnico	4	0	0
7	0	20,8	67	175	Altro dipl.	4	0	1
8	0	20,6	89	170	Liceo	3	1	0
9	0	27,1	71	180	Liceo	1	0	1
10	0	23,3	63	170	Liceo	4	0	0
11	1	20,5	51	161	Ist.Tecnico	4	0	1
12	1	19,1	58	167	Ist.Tecnico	5	1	1
13	1	22,1	67	165	Altro dipl.	5	1	1
14	1	21,8	51	156	Ist.Tecnico	4	0	0
15	1	19,2	60	170	Ist.Tecnico	5	1	1
16	1	20,8	55	165	Liceo	4	1	1
17	1	21	55	158	Liceo	5	1	0
18	1	20,9	58	170	Liceo	5	1	1
19	1	22,7	76	170	Liceo	6	1	0
20	1	21	55	165	Liceo	7	0	0

Esercizio 1

Calcolare la mediana, la varianza e l'indice di Asimmetria di Fisher per il carattere Altezza suddiviso in 3 classi equifrequenti.

C_i	n_i	f_i	N_i	F_i	$\hat{C}_i$	a_i	d_i
C_1 = [156; 165]	7	0,35	7	0,35	160,5	9	0,039
C_2 =] 165; 173]	7	0,35	14	0,70	169	8	0,043
C_3 =] 173; 187]	6	0,30	20	1	180	14	0,021
Totali	**20**	**1,00**					

Media = $\mu_X = \dfrac{\sum_{i=1}^{k} \hat{x}_i \cdot n_i}{n}$

$$\bar{x} = \frac{\sum_{i=1}^{k}(160{,}5 \times 7 + 169 \times 7 + 180 \times 6)}{20} = 169{,}3$$

Mediana

$$\frac{F_{Me-1}}{F_{Me-1}}$$

$$Me = 165 + (173 - 165) \cdot \frac{0{,}5 - 0{,}35}{0{,}7 - 0{,}35} = 168{,}42$$

Varianza

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{c}_i - \bar{x}\right)^2 * n_i = \frac{(160{,}5 - 169{,}3)^2 \cdot 7 + (169 - 169{,}3)^2 \cdot 7 + (180 - 169{,}3)^2 \cdot 6}{20} = 46{,}33$$

Scarto quadratico medio = $\sigma = \sqrt{\sigma^2} = \sqrt{7{,}8475} =$ **2,8**

L'indice di Fisher, è un indice di forma basato sui momenti terzi standardizzati:

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3 n_i$$

$$\gamma > 0 \rightarrow AsimmetricaPositiva;$$
$$\gamma = 0 \rightarrow Simmetrica;$$
$$\gamma < 0 \rightarrow AsimmetricaNegativa;$$

Partendo dalla distribuzione in classi del carattere Durata,

c_i	n_i	$\hat{C}_i$
C_1 = [156; 165]	7	160,5
C_2 =] 165; 173]	7	169
C_3 =] 173; 187]	6	180
Totali	**20**	

Calcoliamo dapprima la media aritmetica

$$\bar{x} = \frac{\sum_{i=1}^{k}(160{,}5 \times 7 + 169 \times 7 + 180 \times 6)}{20} = 169{,}3$$

Poi lo scarto quadratico medio

$\sigma = \sqrt{\sigma^2} = \sqrt{7,8475} = \mathbf{2,8}$

A questo punto abbiamo tutti gli elementi utili per calcolare l'indice di asimmetria di Fisher.

x_i	n_i	$\hat{c}_i$	$\hat{c}_i - \bar{x}$	$Z_i = \frac{(\hat{c}_i - \bar{x})}{\sigma}$	$Z_i = \left(\frac{(\hat{c}_i - \bar{x})}{\sigma}\right)^3$	$(Z_i)^3 \cdot n_i$
C_1 = [156; 165]	7	160,5	-8,8	-1,294	-2,16731	-15,17118
C_2 =] 165; 173]	7	169	-0,3	-0,044	-0,00009	-0,00060
C_3 =] 173; 187]	6	180	10,7	1,574	3,89605	23,37630
Totali	**20**					**8,20452**

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{c}_i - \bar{x}}{\sigma}\right)^3 \cdot n_i = \frac{8,20452}{20} = 0,41$$

Possiamo concludere che la distribuzione è caratterizzata da un'asimmetria positiva (indice maggiore di zero).
Tale risultato è confermato dal confronto tra la mediana e la media aritmetica.

$$\bar{x} = 169,5 > Me = 168,42$$

Esercizio 3
Calcolare l'indice di Curtosi di Pearson per il carattere Altezza.

Soluzione
La Curtosi riguarda un maggiore o minore appiattimento della forma della distribuzione.
L'indice γ_c è un indice di forma basato sui momenti quarti standardizzati.

$$\gamma_C = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4 - 3$$

$$\gamma_C > 0 \rightarrow Leptocurtica;$$
$$\gamma_C = 0 \rightarrow Normocurtica;$$
$$\gamma_C < 0 \rightarrow Platicurtica;$$

Conosciamo la media

$$\bar{x} = \frac{\sum_{i=1}^{k}(160{,}5 \times 7 + 169 \times 7 + 180 \times 6)}{20} = 169{,}3$$

E lo scarto quadratico medio

$$\sigma = \sqrt{\sigma^2} = 6{,}8$$

Dati ordinati	**n_i**	$\hat{c}_i - \bar{x}$	$Z_i = \frac{(\hat{c}_i - \bar{x})}{\sigma}$	$Z_i = \left(\frac{(\hat{c}_i - \bar{x})}{\sigma}\right)^4$	$(Z_i)^4 \cdot n_i$
C_1 = [156; 165]	7	-8,8	-1,294	2,803736	19,626151
C_2 =] 165; 173]	7	-0,3	-0,044	0,000004	0,000026
C_3 =] 173; 187]	6	10,7	1,574	6,137887	36,827324
					56,45

$$\gamma_c = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{c}_i - \bar{x}}{\sigma}\right)^4 - 3 = \left(\frac{56{,}45}{20}\right) - 3 = -0{,}18$$

La distribuzione è Platicurtica.

Esercizio 4

Si calcoli l'indice di Eterogeneità di Gini per il carattere *Tipo di Diploma*.

Nel caso di variabili qualitative la variabilità del carattere è espressa in termini di mutabilità, definita come l'attitudine di un carattere ad assumere differenti modalità qualitative. Quando tutte le unità statistiche assumono la stessa modalità, si ha una perfetta omogeneità. (minima eterogeneità) Quando le modalità del carattere hanno tutte la stessa frequenza assoluta o relativa, si ha la massima disomogeneità. L'Eterogeneità misura la variabilità delle frequenze delle k modalità del carattere.

Soluzione

L'Indice di Eterogeneità (G) di Gini si basa sulle frequenze relative.

Efficacia della Laurea	ni	fi
Liceo	10	0,5
Ist. Tecnici	8	0,4
Altro diploma	2	0,1
Totale	20	1

$$G = 1 - \sum_{i=1}^{k} f_i^2 = 1 - \left(0{,}5^2 + 0{,}4^2 + 0{,}1^2\right) = 1\text{-}0{,}42 = 0{,}58$$

Dividiamo l'indice per il suo massimo:

Gmax=1-1/k=1 = 1 -0,33=0,67

G*=G/Gmax= 0,58/0,67=0,86

Conclusione

G* prossimo ad 1, la distribuzione è eterogenea.

Esercizio 5.

A partire dal numero di componenti della famiglia dei primi 5 uomini e donne del dataset determinare un indice di Mutua variabilità

Per costruire un indice di mutua variabilità consideriamo le differenze a coppie fra le unità statistiche |xi-xj|. Tali differenze vanno calcolate per tutte le possibili coppie. Come indice sintetico si può considerare la differenza media (ovvero sommare tutte le differenze e dividere per il numero totale di coppie).
L'indice che misura la mutua variabilità è, quindi, la differenza semplice media (Δ).
La differenza semplice media è interpretabile come la distanza media tra le unità statistiche prese a coppie.

Per il calcolo del Δ si può costruire la tabella delle differenze semplici , come segue:

Tabella numero di componenti donne:

	4	4	5	6	8
4	0	0	-1	-2	-4
4	0	0	-1	-2	-4
5	1	1	0	-1	-3
6	2	2	1	0	-2
8	4	4	3	2	0

Per poi applicare la formula

$$\Delta = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{n(n-1)} =$$

$$\frac{2\times(1+2+4+1+2+4+1+3+2)}{5\times 4} = \frac{40}{20} = 2$$

Un indice normalizzato in [0, 1] di mutua variabilità si ottiene dividendo Δ per il suo massimo teorico 2μ. Tale indice si indica con R.

R=Δ/2μ, noto come rapporto di concentrazione.

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.
La media è pari a $\bar{x} = 5{,}4$ per cui avremo $R = \frac{2}{2 * 5{,}4} = 0{,}18$

In tal caso le donne hanno una concentrazione pari al 18% della massima concentrazione che avrebbe potuto rilevarsi.

Tabella numero di componenti uomini:

	4	4	5	5	5
4	0	0	-1	-1	-1
4	0	0	-1	-1	-1
5	1	1	0	0	0
5	1	1	0	0	0
5	1	1	0	0	0

Per poi applicare la formula

$$\Delta = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{n(n-1)} =$$

$$\frac{2 \times (1+1+1+1+1+1)}{5 \times 4} = \frac{12}{20} = 0{,}6$$

La media è pari a $\bar{x} = 4{,}4$ per cui avremo $R = \frac{0{,}6}{2 * 4{,}4} = 0{,}068$

In tal caso le donne hanno una concentrazione pari al 7% della massima concentrazione che avrebbe potuto rilevarsi.