Università di Cassino
Corso di Statistica 1
Esercitazione del 09/11/2006
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) Data la distribuzione del carattere **Reddito** dei soli maschi, si analizzi la mutua variabilità e si calcoli il Rapporto di Concentrazione di Gini.
b) Si calcoli l'indice di Eterogeneità di Gini per il carattere **Colore degli occhi**.
c) Calcolare l'indice di Asimmetria di Fisher per il carattere **Età** suddiviso in 4 classi equiampie.

Soluzioni

a) In presenza di caratteri trasferibili si pone un maggior interesse sullo studio della variabilità tra le singole unità statistiche, piuttosto che lo studio della variabilità rispetto ad un centro. In questo caso bisogna calcolare l'indice Delta Δ che rappresenta la media delle differenze, in valore assoluto, di tutte le possibili coppie.

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

Considerando il reddito dei soli maschi avremo:

Soggetto	Sesso	Reddito
1	M	0,7
4	M	2,5
7	M	3,8
8	M	1,3
12	M	0,8
15	M	1,9
19	M	1,6
20	M	2,2

Ordinando i valori di reddito avrà:

Soggetto	Sesso	Reddito
1	M	0,7
12	M	0,8
8	M	1,3
19	M	1,6
15	M	1,9
20	M	2,2
4	M	2,5
7	M	3,8

Si dispongono i valori ordinati in una tabella doppia sia in riga che in colonna:

	0,7	0,8	1,3	1,6	1,9	2,2	2,5	3,8
0,7								
0,8								
1,3								
1,6								
1,9								
2,2								
2,5								
3,8								

In modo da calcolare facilmente le differenze di tutte le possibili coppie

	0,7	**0,8**	**1,3**	**1,6**	**1,9**	**2,2**	**2,5**	**3,8**
0,7	**0**	0,1	0,6	0,9	1,2	1,5	1,8	3,1
0,8	0,1	**0**	0,5	0,8	1,1	1,4	1,7	3
1,3	0,6	0,5	**0**	0,3	0,6	0,9	1,2	2,5
1,6	0,9	0,8	0,3	**0**	0,3	0,6	0,9	2,2
1,9	1,2	1,1	0,6	0,3	**0**	0,3	0,6	1,9
2,2	1,5	1,4	0,9	0,6	0,3	**0**	0,3	1,6
2,5	1,8	1,7	1,2	0,9	0,6	0,3	**0**	1,3
3,8	3,1	3	2,5	2,2	1,9	1,6	1,3	**0**

Effettuando la somma degli elementi di ogni colonna (marginale di colonna) si ha:

	9,2	**8,6**	**6,6**	**6**	**6**	**6,6**	**7,8**	**15,6**

Effettuando la somma dei marginali di colonna si ottiene il valore **66,4** che corrisponde al numeratore dell'indice Delta Δ.
Adesso disponiamo di tutti i dati per poter calcolare l'indice.

$$\Delta = \frac{\sum_{i \neq j=1}^{n} |x_i - x_j|}{n(n-1)} = \frac{66,4}{56} = 1,1857$$

Quando tutte le modalità coincidono, l'indice Delta assume valore zero ($\Delta = 0$).
L'indice Delta assume il valore massimo quando tutte le modalità tranne una sono nulle ($\Delta = 2\mu$).
Il **Rapporto di Concentrazione** (**R**) di Gini si ottiene dividendo Delta per il suo valore massimo.

$$R = \frac{\Delta}{2\mu}$$

Si tratta di un indice relativo che varia tra 0 ed 1. $0 \leq R \leq 1$
R=0 si ha concentrazione minima.
R=1 si ha concentrazione massima.

Nel nostro esercizio, la media del reddito dei Maschi sarà:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{14,8}{8} = 1,85$$

Quindi il massimo valore che può assumere Delta sarà:

$$2\mu = 3,7$$

Per cui il **Rapporto di Concentrazione** (**R**) di Gini sarà:

$$R = \frac{\Delta}{2\mu} = \frac{1,1857}{3,7} = 0,3204$$

b) 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.

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

$$G = 1 - \sum_{i=1}^{k} f_i^2$$

Si tratta di un indice relativo che varia tra $0 \leq G \leq 1 - \frac{1}{k}$

G=0 si ha la minima eterogeneità.

$G = 1 - \frac{1}{k}$ si ha la massima eterogeneità.

La distribuzione di frequenza della variabile **Colore degli occhi** è:

Colore degli occhi	n_i	f_i
Nero	7	0,35
Marrone	6	0,3
Azzurro	4	0,2
Verde	3	0,15
Tot:	**20**	**1**

Colore degli occhi	f_i	$(f_i)^2$
Nero	0,35	0,1225
Marrone	0,3	0,09
Azzurro	0,2	0,04
Verde	0,15	0,0225
Tot:	**1**	**0,275**

Quindi **G** sarà:

$$G = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0{,}275 = 0{,}725$$

Volendo normalizzare G si divide il valore ottenuto per il suo massimo $1 - \frac{1}{k}$ ottenendo così G*

$$G^* = \frac{G \cdot k}{k-1} = 0{,}9667$$

Si può dire che siamo molto vicini al caso di massima eterogeneità.

c) 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 **Età**

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

si calcola prima la media aritmetica

$$\mu = \frac{1}{N}\sum_{i=1}^{C} x_i^c * n_i = 43{,}5$$

e poi lo scarto quadratico medio

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

Fatto ciò, si hanno tutti gli elementi per calcolare l'indice di asimmetria di Fisher

Età	n_i	X_i^c	$(x_i^c - \mu)$	$Z_i = {(x_i^c - \mu)}/{\sigma}$	$Z_i = \left({(x_i^c - \mu)}/{\sigma}\right)^3$	$(Z_i)^3 * n_i$
12-|30	6	21	-22,5	-1,19872	-1,72248	-10,3349
30-|48	6	39	-4,5	-0,23974	-0,01378	-0,08268
48-|66	5	57	13,5	0,719233	0,372056	1,860281
66-|84	3	75	31,5	1,67821	4,726491	14,17947
Tot:	**20**					**5,622182**

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3 n_i = \frac{5{,}622182}{20} = 0{,}2811$$

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.

$$\mu = 43{,}5 > M_e = 42$$