Università di Cassino
Corso di Statistica 1
Esercitazione del 05/11/2007
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 lo scostamento semplice mediano per il carattere **Nucleo familiare** a partire sia dalla successione di valori sia dalla distribuzione di frequenze.
b) Calcolare lo scostamento semplice medio, la varianza, lo scarto quadratico medio e il coefficiente di variazione per il carattere **Statura**, considerando la successione di valori.
c) Determinare lo scostamento semplice medio, la varianza e lo scarto quadratico medio per il carattere **Età**, utilizzando una suddivisione in 4 classi.
d) Data la distribuzione del carattere **Reddito** dei soli maschi, si analizzi la mutua variabilità e si calcoli il Rapporto di Concentrazione di Gini.

Soluzioni

a) 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**	

Lo **scostamento semplice medio dalla Mediana** è una misura di variabilità che si ottiene come media aritmetica delle differenze, in valore assoluto, tra i valori osservati e la mediana.

$$M_e = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2} = \frac{X_{10} + X_{11}}{2} = \frac{3+3}{2} = 3$$

$$S_{Me} = \frac{1}{N}\sum_{i=1}^{n} |x_i - Me| = \frac{21}{20} = 1,05$$

Soggetto	Posizione	Nucleo familiare	$\lvert X_i\text{-Me}\rvert$
11	1	1	2
12	2	1	2
17	3	1	2
3	4	2	1
8	5	2	1
15	6	2	1
19	7	2	1
1	8	3	0
5	9	3	0
9	10	3	0
16	11	3	0
20	12	3	0
2	13	4	1
6	14	4	1
7	15	4	1
14	16	4	1
18	17	4	1
4	18	5	2
10	19	5	2
13	20	5	2

Nel caso delle distribuzioni di frequenza semplici lo scostamento semplice dalla mediana viene calcolato con la seguente formula:

$$S_{Me} = \frac{1}{N}\sum_{i=1}^{k}\lvert x_k - Me\rvert \cdot n_i$$

Nucleo familiare	n_i	F_i	$\lvert X_k\text{-Me}\rvert$	$\lvert X_k\text{-Me}\rvert\ n_i$
1	3	0,15	2	6
2	4	0,35	1	4
3	5	0,60	0	0
4	5	0,85	1	5
5	3	1	2	6
Tot:	**20**			**21**

b) La successione dei valori della variabile **Statura** è la seguente:

Soggetto	Statura	$\lvert X_i - \mu \rvert$	$(X_i - \mu)$	$(X_i - \mu)^2$
1	173	4,45	4,45	19,80
2	168	0,55	-0,55	0,30
3	165	3,55	-3,55	12,60
4	180	11,45	11,45	131,10
5	163	5,55	-5,55	30,80
6	160	8,55	-8,55	73,10
7	177	8,45	8,45	71,40
8	164	4,55	-4,55	20,70
9	158	10,55	-10,55	111,30
10	170	1,45	1,45	2,10
11	167	1,55	-1,55	2,40
12	159	9,55	-9,55	91,20
13	174	5,45	5,45	29,70
14	164	4,55	-4,55	20,70
15	177	8,45	8,45	71,40
16	172	3,45	3,45	11,90
17	166	2,55	-2,55	6,50
18	160	8,55	-8,55	73,10
19	170	1,45	1,45	2,10
20	184	15,45	15,45	238,70
	3371	**120,1**		**1020,95**

Dopo il calcolo della media della variabile **Statura**

$$\mu = \frac{1}{N}\sum_{i=1}^{n} x_i = \frac{3371}{20} = 168,55$$

è possibile calcolare lo scostamento semplice medio.

Lo **scostamento semplice medio dalla media aritmetica** è una misura di variabilità che si ottiene come media aritmetica delle differenze, in valore assoluto, tra i valori osservati e la media aritmetica.

$$S_\mu = \frac{1}{N}\sum_{i=1}^{n} \lvert x_i - \mu \rvert = \frac{120,1}{20} = 6,005$$

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.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad \sigma^2 = \frac{1020,95}{20} = 51,05$$

Una difficoltà nella interpretazione della varianza deriva dal fatto che essa è espressa nell'unità di misura del fenomeno al quadrato. Per questo motivo, si calcola lo scarto quadratico medio.

Lo **scarto quadratico medio** rappresenta la media quadratica degli scarti dalla media aritmetica.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2} \qquad \sigma = \sqrt{\frac{1020,95}{20}} = \sqrt{51,05} = 7,14$$

La varianza e lo scarto quadratico medio sono indici di variabilità assoluti.

Un indice di variabilità relativo è il **coefficiente di variazione** definito da:

$$Cv = \frac{\sigma}{\mu}$$

$$Cv = \frac{\sigma}{\mu} = \frac{7,14}{168,55} = 0,0424$$

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

Età	$\mathbf{n_i}$
12|-30	6
30|-48	6
48|-66	5
66|-|84	3
Tot:	**20**

Il primo passo è quello di calcolare la media aritmetica $\mu = \frac{1}{N}\sum_{i=1}^{C} x_i^c * n_i$

Età	$\mathbf{n_i}$	X_i^c	$(\mathbf{n_i} * X_i^c)$
12|-30	6	21	126
30|-48	6	39	234
48|-66	5	57	285
66|-|84	3	75	225
Tot:	**20**		**870**

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

In questo caso, trattandosi di una distribuzione in classi, lo scostamento semplice medio si trova come: $S_\mu = \frac{1}{N}\sum_{i=1}^{C} \left| x_i^c - \mu \right| n_i$

Età	$\mathbf{n_i}$	X_i^c	$\left\| x_i^c - \mu \right\|$	$\left\| x_i^c - \mu \right\| n_i$
12|-30	6	21	22,5	135
30|-48	6	39	4,5	27
48|-66	5	57	13,5	67,5
66|-|84	3	75	31,5	94,5
Tot:	**20**			**324**

$$S_{\mu} = \frac{324}{20} = 16,2$$

In questo caso, trattandosi di una distribuzione in classi, la varianza si trova come:

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

Età	**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

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

mentre lo scarto quadratico medio è calcolabile come:

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

$$\sigma = \sqrt{\sigma^2} = \sqrt{352,4} = 18,77$$

d) In alcune circostanze si pone un maggior interesse sullo studio della variabilità tra le singole unità statistiche, piuttosto che lo studio della variabilità rispetto ad un centro.

Lo studio della mutua variabilità è possibile attraverso il calcolo dell'indice Delta (Δ) che rappresenta la media delle differenze, in valore assoluto, di tutte le possibili coppie senza ripetizioni.

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

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.

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

Quando tutte le modalità coincidono, l'indice Delta assume valore zero ($\Delta = 0$).

In presenza di caratteri trasferibili è possibile calcolare la variabilità massima e quindi il valore massimo di Delta .

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:

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