Esercitazione 5 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

12 Febbraio 2009

La seguente tabella riporta le informazioni relative a 25 laureati nell'anno 2005 in Economia, ad un anno dal conseguimento del titolo.

Genere	Eta	Voto*	Durata	Tipo Contratto	Utilizzo della Laurea	Efficacia della Laurea
Uomini	47,8	83	4,4	Stabile	In misura elevata	Efficace
Uomini	26,6	113	7,4	Stabile	In misura ridotta	Poco efficace
Uomini	31,5	91	12,4	Atipico	In misura ridotta	Abb. efficace
Uomini	23,6	102	4,4	Inserimento/Form	In misura elevata	Efficace
Uomini	25,9	94	6,4	Stabile	Per niente	Per nulla efficace
Donne	23,6	108	4,7	Inserimento/Form	In misura ridotta	Abb. efficace
Donne	28,6	108	5,7	Atipico	In misura ridotta	Abb. efficace
Uomini	42,1	100	3	Stabile	Per niente	Efficace
Donne	24,3	113	3,4	Atipico	Per niente	Per nulla efficace
Donne	26,3	113	3,4	Atipico	Per niente	Per nulla efficace
Uomini	24,9	106	3,7	Inserimento/Form	In misura elevata	Molto efficace
Donne	24	95	4,5	Stabile	In misura elevata	Efficace
Uomini	34,7	92	4,4	Stabile	In misura elevata	Efficace
Donne	24,7	106	5,14	Atipico	In misura ridotta	Abb. efficace
Uomini	25,9	100	5,94	Atipico	In misura ridotta	Abb. efficace
Uomini	25,4	92	6,14	Atipico	In misura elevata	Molto efficace
Donne	27,6	113	4,14	Senza contratto	Per niente	Poco efficace
Donne	23,4	113	4,4	Inserimento/Form	In misura elevata	Molto efficace
Uomini	31,3	105	3,4	Stabile	In misura elevata	Molto efficace
Uomini	29,7	110	4,14	Atipico	In misura ridotta	Abb. efficace
Uomini	27	93	7,3	Stabile	In misura elevata	Efficace
Uomini	35,6	97	15,4	Atipico	Per niente	Poco efficace
Uomini	23,1	101	3,7	Senza contratto	Per niente	Poco efficace
Uomini	25,3	91	6,5	Atipico	Per niente	Per nulla efficace
Uomini	32,6	92	13,5	Stabile	In misura ridotta	Abb. efficace

Esercizio 1.

A partire dagli stipendi percepiti da 5 dirigenti e impiegati (espressi in migliaia di euro)

a) determinare gli indici di Mutua variabilità e Concentrazione

b) costruire la curva di Lorenz;

Per costruire un indice di mutua variabilità consideriamo le differenze a coppie fra le unità statistiche $|x_i-x_j|$. 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 stipendi impiegati:

	3	3,4	3,7	3,7	4,14
3	0	-0,4	-0,7	-0,7	-1,14
3,4	0,4	0	-0,3	-0,3	-0,74
3,7	0,7	0,3	0	0	-0,44
3,7	0,7	0,3	0	0	-0,44
4,14	1,14	0,74	0,44	0,44	0

Per poi applicare la formula

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

$$\frac{2\times(0{,}4+0{,}7+0{,}3+0{,}7+0{,}3+1{,}14+0{,}74+0{,}44+0{,}44)}{5\times 4} = \frac{10{,}32}{20} = 0{,}516$$

Se i dati a disposizione sono disposti in distribuzioni di frequenza, le possibili differenze $|x_i\text{-}x_j|$ saranno in numero pari a $n_i\, n_j$. Allora, la formula della differenza semplice media (Δ) diventa quella in figura.

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

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 $\bar{x} = 3{,}588$ per cui avremo $R = \frac{0{,}516}{2*3{,}588} = 0{,}072$

Tale risultato indica che nella distribuzione osservata la concentrazione è circa il 7% della massima concentrazione che avrebbe potuto rilevarsi.

Tabella stipendi dirigenti:

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

Per poi applicare la formula

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

$$\frac{2\times(0{,}5+1{,}2+1+1{,}9+0{,}7+0{,}8+1{,}4+0{,}1+0{,}7+0{,}6)}{5\times 4} = \frac{17{,}8}{20} = 0{,}89$$

La media è pari a $\bar{x} = 4{,}48$ per cui avremo $R = \frac{0{,}89}{2*4{,}48} = 0{,}9933$

In tal caso i dirigenti hanno una concentrazione pari all'9% della massima concentrazione che avrebbe potuto rilevarsi.

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.

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{0{,}1438}{2} = 0{,}072$$

Impiegati

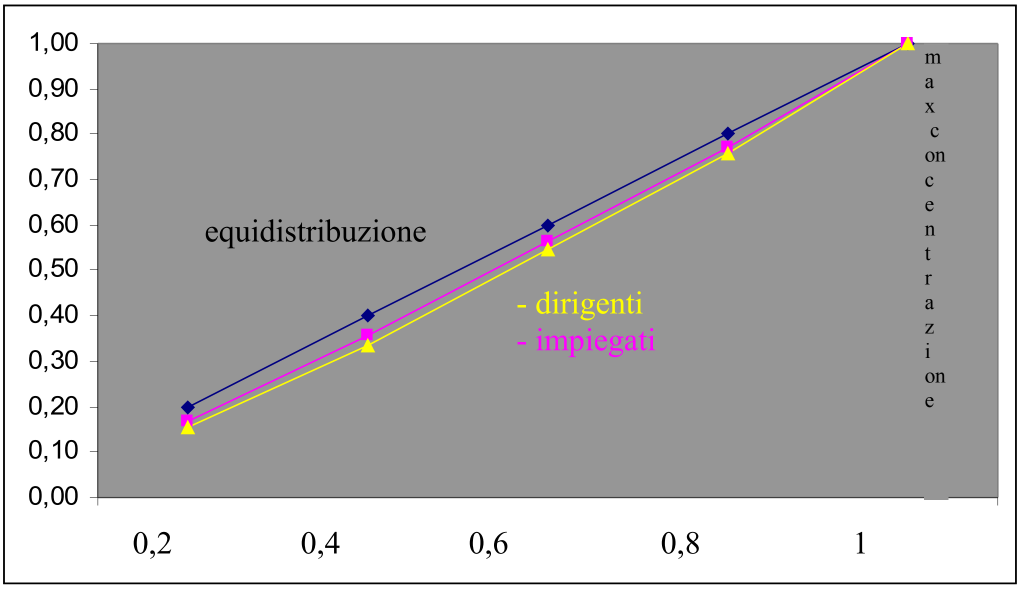
X(i)	Ai	qi	pi	pi-qi
3	3	0,167224	0,2	0,032776
3,4	6,4	0,356745	0,4	0,043255
3,7	10,1	0,562988	0,6	0,037012
3,7	13,8	0,769231	0,8	0,030769
4,14	17,94	1		0

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{0{,}2053}{2} = 0{,}10$$

dirigenti

X(i)	Ai	qi	pi	pi-qi
3,5	3,5	0,15625	0,2	0,04375
4	7,5	0,334821	0,4	0,065179
4,7	12,2	0,544643	0,6	0,055357
4,8	17	0,758929	0,8	0,041071
5,4	22,4	1	1	0

Per costruire il diagramma di Lorenz basta porre sugli assi cartesiani le quantità pi e qi determinate precedentemente (e riportate nelle tabelle accanto alle distribuzioni di frequenze):



1) Considerando il dataset Laureati, si discuta se esiste connessione tra i caratteri Tipo di contratto ed Efficacia della laurea.

Soluzione

La distribuzione doppia dei caratteri corso laurea e attività sportiva è la seguente:

Tipo di contratto Utilizzo della laurea	Stabile	Atipico	Inserimento / formazione	Senza contratto	Tot
In misura elevata	5	3	1	0	9
In misura ridotta	2	5	1	0	8
Per niente	2	4	0	2	8
totale	9	12	2	2	25

Trattandosi di due caratteri qualitativi, il loro grado di connessione si misura attraverso l'indice $\chi 2$:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

Le frequenze teoriche $\hat{n}_{ij} = \dfrac{n_{i.} \times n_{.j}}{N}$ sono raccolte nella seguente tabella:

Tipo di contratto Utilizzo della laurea	Stabile	Atipico	Inserimento/ formazione	Senza contratto
In misura elevata	3,24	4,32	0,72	0,72
In misura ridotta	2,88	3,84	0,64	0,64
Per niente	2,88	3,84	0,64	0,64

Sostituendo nella formula si ha:

$$\chi^2 = \frac{(5-3{,}24)^2}{3{,}24} + \frac{(3-4{,}32)^2}{4{,}32} + \frac{(1-0{,}72)^2}{0{,}72} + \frac{(0-0{,}72)^2}{0{,}72} + \ldots\ldots + \frac{(2-0{,}64)^2}{0{,}64} = \mathbf{7{,}45}$$

Tabella dei calcoli

0,956	0,403	0,108	0,72
0,268	0,350	0,202	0,64
0,268	0,006	0,64	2,89

Somma=7,45

L'indice
χ^2 dipende dalla numerosità del collettivo, cosicché, a parità di associazione, il suo valore aumenta all'aumentare di N. Generalmente si preferisce utilizzare degli indici "normalizzati" che diano misure non dipendenti dalla numerosità.

$$\phi^2 = \frac{\chi^2}{N} = \frac{7{,}45}{25} = 0{,}3$$

Possiamo affermare che vi è un basso grado di connessione.

Tale valore va confrontato con l'intervallo [0, 2], in quanto

$$0 \leq \varphi^2 \leq \min(r - 1;\ c - 1)$$