FACOLTA' DI SOCIOLOGIA – A.A. 2005-2006
ESAME DI RELAZIONI TRA VARIABILI
4/7/2006

<u>Avvertenza</u>: Fornire le formule utilizzate e tutti i passaggi dei calcoli eseguiti. Utilizzare almeno 2 cifre decimali.

Esercizio 1

Su un collettivo di 10 lavoratori sono stati osservati congiuntamente i caratteri X = "Reddito lordo annuo (in migliaia di euro)" e Y = "Genere", ottenendo i seguenti risultati:

Lavoratore	1	2	3	4	5	6	7	8	9	10
X	50-|60	>60	20-|30	>60	30-|50	20-|30	30-|50	50-|60	20-|30	30-|50
Y	M	M	F	M	F	F	M	M	F	M

(a) Organizzare i dati in una tabella a doppia entrata e fornire la distribuzione condizionata del carattere X nella sottopopolazione dei maschi. Commentare i risultati ottenuti confrontandoli con la distribuzione marginale di X.

(b) Valutare e commentare, utilizzando un opportuno indice, il grado di associazione tra le modalità "maschio" e "reddito superiore a 60 mila euro annui".

(c) Dopo aver esposto il concetto di connessione fra coppie di fenomeni, costruire l'indice Chi Quadrato discutendone la metodologia, i possibili valori e la normalizzazione.

(d) Osservando le distribuzioni costruite al punto a) e senza effettuare altri calcoli si potrebbe affermare che il reddito è statisticamente indipendente dal genere? In caso di risposta negativa, valutare e commentare il grado di connessione tra X e Y.

(e) Esporre il concetto di indipendenza in media di un fenomeno quantitativo da un altro fenomeno, definire l'indice di dipendenza η^2 e discuterne i valori.

Esercizio 2

Su un gruppo di laureati sono stati congiuntamente osservati i caratteri X = "Tempo (in anni) dalla laurea" e Y = "Reddito lordo mensile (in migliaia di euro)", ottenendo la seguente tabella a doppia entrata:

X \ Y	1-|2	2-|3	3-|4
5	8	2	0
10	6	22	0
20	0	0	16

(a) Costruire e commentare il diagramma a dispersione. Determinare la retta di regressione dei minimi quadrati che interpreta la dipendenza di Y da X.

(b) Tracciare la retta calcolata sul diagramma a dispersione e in seguito valutarne con un indice opportuno la bontà di adattamento ai dati, interpretando il risultato numerico ottenuto. Infine utilizzare la retta di regressione per prevedere il reddito lordo mensile a 15 anni dalla laurea, valutandone anche l'affidabilità sulla base della bontà di adattamento del modello utilizzato per la previsione.

(c) Esporre e discutere il criterio dei Minimi Quadrati per la determinazione della retta di regressione

(d) Sia 0.22 la varianza residua lasciata da un secondo modello, diverso dalla retta di regressione individuata al punto (a), che spiega il fenomeno Y in funzione di X. Stabilire, motivando la risposta, se tale modello è preferibile alla retta di regressione calcolata al punto (b).

(e) Definire la covarianza e discuterne anche geometricamente significato, valore, segno e legame con il coefficiente di correlazione lineare ρ.

FACOLTA' DI SOCIOLOGIA – A.A. 2005-2006
ESAME DI RELAZIONI TRA VARIABILI
Soluzioni appello del 4/7/2006

Esercizio 1

a)

X \ Y	M	F	
20-|30	0	3	3
30-|50	2	1	3
50-|60	2	0	2
>60	2	0	2
	6	4	10

Nella tabella seguente sono indicate rispettivamente la distribuzione relativa condizionata del carattere X per i maschi e la distribuzione relativa marginale del carattere X:

X	p_{i2}	$p_{i.}$
20-|30	0	0.3
30-|50	0.33	0.3
50-|60	0.33	0.2
>60	0.33	0.2

Osservando la distribuzione condizionata di X, si nota che tra i maschi non è presente alcun individuo con reddito annuo compreso tra i 20 e i 30mila euro. I redditi dei maschi si distribuiscono uniformemente tra le altre tre categorie di reddito. Dalla distribuzione marginale di X si nota invece una lieve predominanza di redditi annui inferiori a 50mila euro.

b)
Si tratta di calcolare l'indice di associazione di Edwards tra le modalità "maschio" e "reddito superiore a 60mila euro". Si costruisce pertanto una tabella 2x2 con i caratteri di interesse posizionati rispettivamente come prima modalità di riga e di colonna:

X \ Y	M	F	
>60	2	0	2
≤60	4	4	8
	6	4	

Dunque si avrà:

$$E = \frac{f_{11}f_{22}}{f_{11}f_{22} + f_{12}f_{21}} = \frac{2 \times 4}{2 \times 4 + 4 \times 0} = 1$$

Il grado di associazione tra le due modalità è massimo e positivo.

d)
Il grado di connessione viene valutato attraverso l'indice chi quadrato:

$$\chi^2 = N\left(\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{f_{ij}^2}{f_{i.}f_{.j}} - 1\right) =$$

$$= 10\left(\frac{3^2}{3\times 4} + \frac{2^2}{3\times 6} + \frac{1^2}{3\times 4} + \frac{2^2}{2\times 6} + \frac{2^2}{2\times 6} - 1\right) = 7.22$$

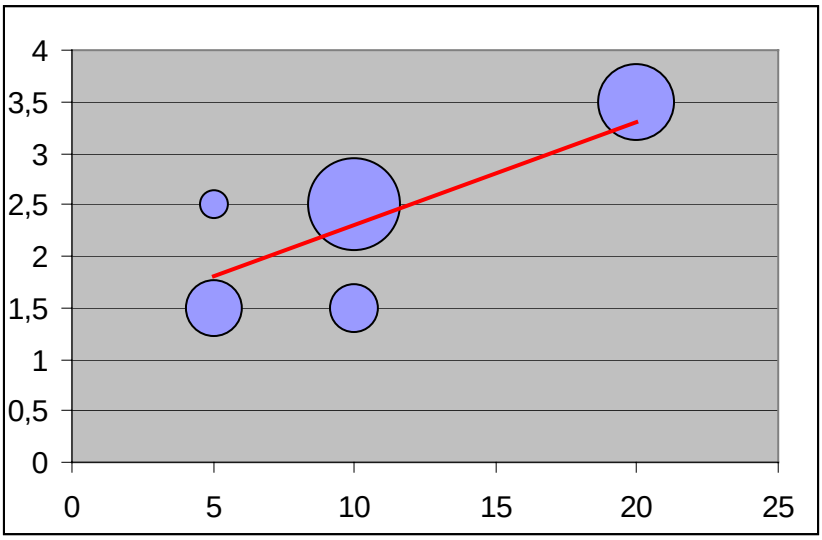
Considero l'indice normalizzato:

$$\tilde{\chi}^2 = \frac{\chi^2}{N\times \min\{(h-1),(k-1)\}} = \frac{7.22}{10\times \min\{3,1\}} = 0.72$$

Si nota un grado di connessione piuttosto elevato tra le due variabili.

Esercizio 2

a)



$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i f_{i.} = \frac{5\times 10 + 10\times 28 + 20\times 16}{54} = 12.04$$

Nei calcoli relativi alla variabile Y utilizzo i valori medi degli intervalli:

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{h} y_j f_{.j} = \frac{1.5\times 14 + 2.5\times 24 + 3.5\times 16}{54} = 2.54$$

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{k} x_i^2 f_{i.} - \bar{x}^2 = \frac{5^2\times 10 + 10^2\times 28 + 20^2\times 16}{54} - 12.04^2 = 30.04$$

$$\sigma_Y^2 = \frac{1}{N}\sum_{j=1}^{h} y_j^2 f_{.j} - \bar{y}^2 = \frac{1.5^2 \times 14 + 2.5^2 \times 24 + 3.5^2 \times 16}{54} - 2.54^2 = 0.54$$

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j f_{ij} - \bar{x} \times \bar{y} =$$

$$= \frac{5 \times 1.5 \times 8 + 5 \times 2.5 \times 2 + 10 \times 1.5 \times 6 + 10 \times 2.5 \times 22 + 20 \times 3.5 \times 16}{54} - 12.04 \times 2.54 =$$

$$= 3.59$$

I parametri della retta di regressione sono quindi dati da:

$$b = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{3.59}{30.04} = 0.12 \qquad a = \bar{y} - b\bar{x} = 2.54 - 0.12 \times 12.04 = 1.09$$

b)
Calcolo l'indice della bontà di adattamento ai dati:

$$\rho_{XY}^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \frac{3.59^2}{30.04 \times 0.54} = 0.79$$

che indica un buon adattamento della retta di regressione ai dati.

$$\hat{Y} = 1.09 + 0.12 \times 15 = 2.89$$

Il modello prevede un reddito lordo mensile di 2'890 euro a 15 anni dalla laurea. L'affidabilità di questa previsione è elevata, pari all'80%.

d)
Per scegliere tra i due modelli confronto le varianze residue. La varianza residua del primo modello (ovvero della retta dei minimi quadrati individuata in precedenza) è pari a:

$$\sigma_Y^2(1 - \rho_Y^2) = 0.54 \times 0.21 = 0.11$$

Poiché 0.11<0.22 si conclude che il modello 1 è migliore.