**Università di Cassino**

**Esercitazione di Statistica 1 del 21 novembre 2007**

**Dott.ssa Paola Costantini**

1) Considerando il DATASET DIPENDENTI, si discuta se esiste connessione tra i caratteri REGIME DI IMPIEGO e QUALIFICA FUNZIONALE.

DATASET DIPENDENTI

| ID | Stipendio percepito | Età | N. di anni di servizio | Qualifica funzionale | Regime di impiego | Genere | Stato Civile |
|---|---|---|---|---|---|---|---|
| 1 | 2650 | 40 | 15 | Operaio | Tempo pieno | M | Non coniugato |
| 2 | 2600 | 43 | 5 | Operaio | Part time | F | Vedovo |
| 3 | 2050 | 35 | 6 | Impiegato | Tempo pieno | F | Coniugato |
| 4 | 3500 | 27 | 6 | Dirigente | Part time | M | Non coniugato |
| 5 | 1400 | 36 | 3 | Dirigente | Collaboratori esterni | F | Vedovo |
| 6 | 2400 | 30 | 12 | Impiegato | Tempo pieno | M | Vedovo |
| 7 | 1900 | 41 | 13 | Operaio | Tempo pieno | F | Non coniugato |
| 8 | 2100 | 35 | 4 | Impiegato | Tempo pieno | M | Vedovo |
| 9 | 2100 | 27 | 7 | Operaio | Tempo pieno | F | Non coniugato |
| 10 | 3050 | 38 | 18 | Dirigente | Tempo pieno | F | Coniugato |
| 11 | 2800 | 38 | 20 | Operaio | Collaboratori esterni | M | Non coniugato |
| 12 | 2950 | 41 | 11 | Operaio | Collaboratori esterni | F | Non coniugato |
| 13 | 1900 | 36 | 4 | Dirigente | Collaboratori esterni | M | Vedovo |
| 14 | 1650 | 29 | 11 | Impiegato | Collaboratori esterni | F | Coniugato |
| 15 | 2550 | 40 | 4 | Impiegato | Collaboratori esterni | M | Non coniugato |
| 16 | 2000 | 23 | 10 | Impiegato | Tempo pieno | F | Coniugato |
| 17 | 2150 | 26 | 8 | Operaio | Collaboratori esterni | F | Coniugato |
| 18 | 2900 | 41 | 9 | Dirigente | Tempo pieno | M | Non coniugato |
| 19 | 2450 | 35 | 12 | Operaio | Collaboratori esterni | F | Coniugato |
| 20 | 1950 | 31 | 8 | Dirigente | Collaboratori esterni | M | Vedovo |

**Soluzione**

La distribuzione doppia dei caratteri CORSO LAUREA e ATTIVITÀ SPORTIVA è la seguente:

| Regime di impiego / Qualifica funzionale | Tempo Pieno | Part Time | Collaboratore esterno |
|---|---|---|---|
| Operaio | 3 | 1 | 4 |
| Impiegato | 4 | 0 | 2 |
| Dirigente | 2 | 1 | 3 |

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

| Regime di impiego / Qualifica funzionale | Tempo Pieno | Part Time | Collaboratore esterno |
|---|---|---|---|
| Operaio | 3,6 | 0,8 | 3,6 |
| Impiegato | 2,7 | 0,6 | 2,7 |
| Dirigente | 2,7 | 0,6 | 2,7 |

Sostituendo nella formula si ha:

$$\chi^2 = \frac{(3-3,6)^2}{3,6} + \frac{(1-0,8)^2}{0,8} + \frac{(4-3,6)^2}{3,6} + \frac{(4-2,7)^2}{2,7} + \ldots\ldots + \frac{(3-2,7)^2}{2,7} = \mathbf{2,073}$$

Quindi si ha:

$$\phi^2 = \frac{\chi^2}{N} = \frac{2,073}{20} = \mathbf{0,10365}$$

Possiamo affermare che vi è un basso grado di connessione.

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

$0 \le \varphi^2 \le \min\left(r-1;\ c-1\right)$

**Indice T di Tchuprov**

$$\mathbf{T} = \frac{\phi^2}{\min\{r-1,\ c-1\}} = \frac{\chi^2}{n \times \min\{r-1,\ c-1\}} = \mathbf{0,03455}$$

per $0 \le \mathbf{T} \le 1$

**Formula alternativa del Chi-quadro**

$$\chi^2 = n\left[\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

| $n_{ij}^2$ | Tempo Pieno | Part Time | Collaboratore esterno |
|---|---|---|---|
| Operaio | 9 | 1 | 16 |
| Impiegato | 16 | 0 | 4 |
| Dirigente | 4 | 1 | 9 |

| $n_{i.} \times n_{.j}$ | Tempo Pieno | Part Time | Collaboratore esterno |
|---|---|---|---|
| Operaio | 72 | 16 | 72 |
| Impiegato | 54 | 12 | 54 |
| Dirigente | 54 | 12 | 54 |

$\chi^2$ = 20 x [(9/72+1/16+16/72+16/54+……..+9/54)-1] = **2,073**

**2)** Considerando le classi del carattere "Stipendio Percepito"
1400 |–| 2100;
2100 –| 2800;
2800 –| 3500;
si può affermare che lo STIPENDIO PERCEPITO dipende in media dal SESSO?

L'indice per misurare il grado di indipendenza in media di un carattere quantitativo y da uno qualitativo (o quantitativo) x è il rapporto di correlazione o $\eta_{y|x}$.

| Stipendio / Sesso | 1400 \|–\| 2100 | 2100 –\| 2800 | 2800 –\| 3500 | **totale** |
|---|---|---|---|---|
| Femmine | 3 | 4 | 2 | **9** |
| Maschi | 6 | 3 | 2 | **11** |
| **Totale** | **9** | **7** | **4** | **20** |

$$\eta_{y|x} = \sqrt{\frac{\sum_{i=1}^{r} (\mu_i - \mu)^2 n_{i.}}{\sum_{j=1}^{c} (y_j - \mu)^2 n_{.j}}}$$

Considerando che:

valori centrali: $y_1 = 1750$; $y_2 = 2450$; $y_3 = 3150$

$$\mu = \frac{\sum_{j=1}^{3} y_j n_{.j}}{N} = \frac{1750*9+2450*7+3150*4}{20} = 2275 \text{ stipendio medio generale}$$

$$\mu_1 = \mu_{femine} = \frac{\sum_{j=1}^{3} y_j n_{1j}}{n_{1.}} = \frac{1750*3+2450*4+3150*2}{20} = 2195{,}45 \text{ stipendio medio donne}$$

$$\mu_2 = \mu_{maschi} = \frac{\sum_{j=1}^{3} y_j n_{2j}}{n_{2.}} = \frac{1750*6+2450*3+3150*2}{20} = 2372{,}2 \text{ stipendio medio uomini}$$

Calcolo del numeratore dell'indice

$\Sigma(\mu_{y|x} - \mu_y)^2 * n_{i.} = (2195{,}45 - 2275)^2 * 11 + (2372{,}2 - 2275)^2 * 9 = 154675{,}78$

Calcolo del denominatore dell'indice

$\Sigma(\acute{y}_j - \mu_y)^2 * n_{.j} = (1750 - 2275)^2 * 3 + (2450 - 2275)^2 * 4 + (3150 - 2275)^2 * 4 +$
$+ (1750 - 2275)^2 * 6 + (2450 - 2275)^2 * 3 + (3150 - 2275)^2 * 2 = 5757500$

Calcolo dell'indice

$$\eta_{y|x} = \sqrt{\frac{\sum_{i=1}^{r} (\mu_i - \mu)^2 n_{i.}}{\sum_{j=1}^{c} (y_j - \mu)^2 n_{.j}}} = \mathbf{0{,}16}$$

Considerando che $\mathbf{0} \leq \eta_{y/x} \leq \mathbf{1}$ abbiamo un minimo grado di dipendenza in media.

**3)** Vogliamo scoprire se esiste una qualche relazione (e se esiste di che natura è) tra l'età e il numero di contatti via internet effettuate nell'arco di una giornata dagli utenti registrati di un certo sito. Un campione di 8 utenti ha fornito i seguenti dati:

| Numero contatti | 0 | 8 | 0 | 2 | 3 | 1 | 6 | 5 |
|---|---|---|---|---|---|---|---|---|
| Eta' | 20 | 32 | 18 | 35 | 28 | 15 | 30 | 23 |

$\mu$ numero contatti = 3,125

$\mu$ età = 25,125

$\mu(x,y)$= 90

Cov = $\sigma_{x,y}$ = 90 – (3,125 * 25,125) = 11,5

$$\sigma^2_x = \frac{1}{N} \Sigma (x_i - \mu)^2 = 7{,}6 \quad \sqrt{\sigma_x} = 2{,}75$$

$$\sigma^2_y = \frac{1}{N} \Sigma (y_i - \mu)^2 = 44 \quad \sqrt{\sigma_y} = 6{,}63$$

$$Corr_{x,y} = \frac{Cov_{x,y}}{\sigma_x \sigma_y} = \frac{11{,}5}{2{,}75 * 6{,}63} = \mathbf{0{,}63}$$ **Correlazione positiva**