Esercitazione 6 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

15 Novembre 2011

Esercizio 1

Data la seguente tabella

Settore Reddito	Agricoltura	Industria	Altre attività	Totale
Fino a 15	50	116	160	326
15 ⊣ 30	90	140	241	471
30 ⊣ 45	20	200	260	480
45 ⊣ 65	1	280	200	481
Totale	161	736	861	1758

esiste indipendenza in media di Y da X? In caso di risposta negativa fornire una misura del grado di dipendenza e commentare.

Reddito (valori centrali)	Agricoltura		Industria		Altre attività	
y_i	n_{i1}	$y_i^2 \cdot n_{i1}$	n_{i2}	$y_i^2 \cdot n_{i2}$	n_{i3}	$y_i^2 \cdot n_{i3}$
10	50	5000	116	11600	160	16000
22,5	90	45562,5	140	70875	241	122006,25
37,5	20	28125	200	281250	260	365625
55	1	3025	280	847000	200	605000
Totale	161	81712,5	736	1210725	861	1108631,25

Le medie parziali del reddito annuo risultano essere:

$$\bar{y}_1 = \frac{3330}{161} = 20{,}68 \quad \bar{y}_2 = \frac{27210}{736} = 36{,}97 \quad \bar{y}_3 = \frac{27210}{736} = \frac{27772{,}5}{861} = 32{,}26$$

Le varianze parziali risultano essere:

$$\sigma_1^2 = \frac{81712{,}5}{161} - 20{,}68^2 = 79{,}74$$

$$\sigma_2^2 = \frac{1210725}{736} - 36{,}97^2 = 278{,}22$$

$$\sigma_3^2 = \frac{1108631,25}{861} - 32,26^2 = 247,15$$

Il reddito medio aritmetico per l'intera popolazione risulta essere:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{4} y_i \cdot n_i = \frac{58312,5}{1758} = 33,17$$

La varianza della popolazione totale (calcolata con il metodo indiretto) risulta essere:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{4} y_i^2 \cdot n_{i..} - \bar{y}^2 = \frac{2401068,75}{1758} - 33,17^2 = 256,57$$

Al fine di verificare la scomposizione della varianza, calcoliamo la varianza *interna* ai gruppi:

$$\sigma^2{}_{INT} = \frac{1}{N}\sum_{j=1}^{3} \sigma_j^2 \cdot n._j = \frac{1}{1758}\left[(79,74 \cdot 161) + (278,22 \cdot 736) + (247,15 \cdot 861)\right] = 244,83$$

La varianza *esterna* risulta essere:

$$\sigma^2{}_{EXT} = \frac{1}{N}\sum_{j=1}^{3} (\bar{y}_j - \bar{y})^2 \cdot n. = \frac{1}{1758} \cdot$$

$$\cdot\left[(20,68 - 33,17) \cdot 161 + (278,22 - 33,17) \cdot 736 + (247,15 - 33,17) \cdot 861\right] =$$

$$= \frac{36456,85}{1758} = 20,74$$

Come già osservato, dato che le medie parziali del reddito variano al mutare del settore di attività economica, possiamo concludere che il "Reddito Annuo" non è **indipendente in media** dal "Settore di Attività Economica".

Dato che nel caso di indipendenza in media si ha $\bar{y_1} = \bar{y_2} = \bar{y_3} = \bar{y}$, al fine di quantificare l'allontanamento dall'indipendenza in media è del tutto naturale utilizzare un indice basato sugli scarti $|\bar{yi} - \bar{y}|$. Uno di questi indici è:

$$\eta_{Y|X} = \frac{\sigma^2{}_{\mathrm{EXT_Y}}}{\sigma_Y^2} = \frac{\frac{1}{N}\sum_{j=1}^{3} (\bar{y}_j - \bar{y})^2 \cdot n.}{\frac{1}{N}\sum_{i=1}^{4} y_i^2 \cdot n_{i..} - \bar{y}^2} = \frac{20,74}{265,57} = 0,078$$

Il valore dell'indice informa che la variabilità fra le medie parziali, rappresenta il 7.8% della variabilità totale. L'indice è pari al 7.8% del suo massimo valore assumibile (corrispondente al caso di massima connessione) e ci permette di concludere che il

carattere “Reddito Annuo” è debolmente dipendente in media dal carattere “Settore di Attività Economica”.

Esercizio n 2

Data la seguente tabella, si discuta se esiste connessione tra i caratteri GENERE e QUALIFICA FUNZIONALE.

La distribuzione doppia dei caratteri GENERE e QUALIFICA FUNZIONALE 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 χ^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

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

Indice T di Tchuprov

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

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

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

Esercizio 3

Vogliamo scoprire se esiste una qualche relazione (e se esiste di che natura è) tra il prezzo di vendita (espresso in migliaia di dollari) e la superficie abitabile di 10 abitazioni, espressa in piedi quadrati (0,0929 metri quadrati).
Poiché sospettiamo che i prezzi di vendita possano dipendere dalla superficie abitabile, il prezzo di vendita diventa la variabile dipendente e la superficie abitabile la variabile esplicativa.

SqFt (x)	Prezzo (y)	x_i^2	y_i^2	$x \cdot y$
521	26	271441	676	13546
661	31	436921	961	20491
694	37,4	481636	1398,76	25955,6
743	34,8	552049	1211,04	25856,4
787	39,2	619369	1536,64	30850,4
825	38	680625	1444	31350
883	39,6	779689	1568,16	34966,8
920	31,2	846400	973,44	28704
965	37,2	931225	1383,84	35898
1011	38,4	1022121	1474,56	38822,4
8010	**352,8**	**6621476**	**12627,44**	**286440,6**

$$\mu_y = \frac{352,8}{10} = 35,28 \qquad \mu_x = \frac{8010}{10} = 801 \qquad \sum_{1=1}^{n} x_i^{\,2} = 6621476 \qquad \sum_{1=1}^{n} y_i^{\,2} = 12627,44$$

$$\sigma^2{}_x = VAR(X) = E[X^2] - \mu^2 = \frac{6621476}{10} - 801^2 = 20546,6 \rightarrow \sigma = \sqrt{20546,6} = 143,34$$

$$\sigma^2{}_y = VAR(YX) = E[Y^2] - \mu^2 = \frac{12627,44}{10} - 35,28^2 = 18,06 \rightarrow \sigma = \sqrt{18,06} = 4,25$$

$$\mu(x \cdot y) = 286440,6/10 = 28644,06$$

$$Cov_{x,y} = \mu(x \cdot y) - (\mu_x \cdot \mu_y) = 28644,06 - (801 \cdot 35,28) = 384,78$$

$$Corr_{x,y} = \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{384,78}{143,34 \cdot 4,25} = 0,63$$ **Correlazione positiva**

$$\rho^2_{x,y} = R^2 = 0,63^2 = 0,3989$$

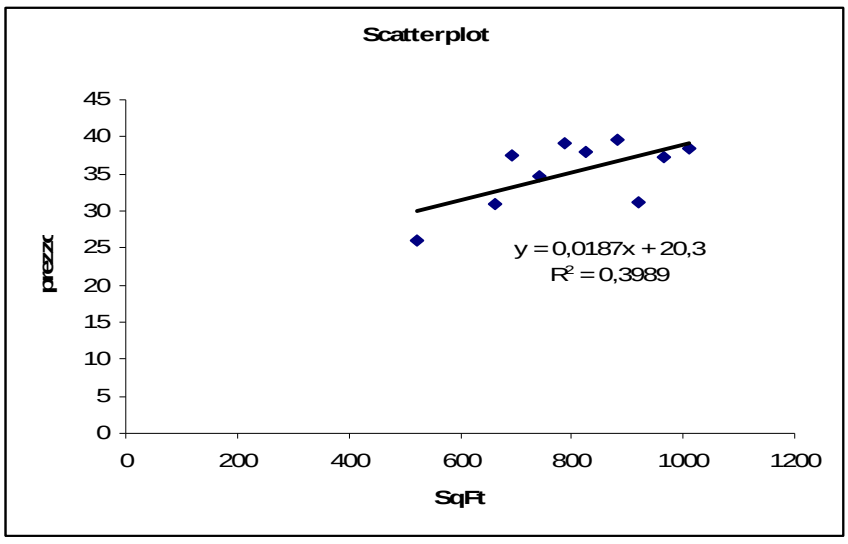
STIMA DELLA RETTA DI REGRESSIONE

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{\beta}_1 = \frac{Cov_{x,y}}{\sigma^2{}_x} = \frac{384,78}{20546,6} = 0,0187$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 35,28 - (0,0187 \cdot 801) = 20,3$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i = 20 + 0,0187 \cdot x_i$$



L'intercetta di y o termine costante è 20,3, espresso nella stessa unità della variabile y. L'inclinazione (o coefficiente di regressione) è 0,0187 e indica la variazione media della variabile y corrispondente alla variazione di una unità della variabile x. L'unità di misura per questo esempio è 0,0187 migliaia di dollari per piede quadrato quadrato,(1 piede q. = 0,0929 metri quadrati) o circa19 euro per metro quadrato. Se due proprietà differiscono in superficie di 100 piedi quadrati ci m che i prezzi previsti di vendita di 0,019 *100 = 1,9 migliaia di dollari ovvero 1.900 dollari.

Prezzo y	SqFt x	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
521	26	-280	-9,28	2598,4	86,12	78400
661	31	-140	-4,28	599,2	18,32	19600
694	37,4	-107	2,12	-226,84	4,49	11449
743	34,8	-58	-0,48	27,84	0,23	3364
787	39,2	-14	3,92	-54,88	15,37	196
825	38	24	2,72	65,28	7,40	576
883	39,6	82	4,32	354,24	18,66	6724
920	31,2	119	-4,08	-485,52	16,65	14161
965	37,2	164	1,92	314,88	3,69	26896
1011	38,4	210	3,12	655,2	9,73	
8010	352,8	0	0	3847,8	180,656	161366

Quando si costruisce un modello di regressione l'obiettivo è quello di spiegare le variazioni della variabile dipendente (Y) mediante le variazioni della variabile esplicativa (X). Maggiore è la percentuale della varianza della Y che si riesce a spiegare con la variabile X, più soddisfacente sarà il modello. L'informazione della percentuale della varianza di Y spiegata dal modello di regressione è fornita dall'indice di determinazione R2, che varia tra 0 e 1. Esso è dato dal rapporto tra devianza spiegata e devianza totale del modello.

$$R^2 = \frac{DevSpiegata}{DevTotale} = \frac{D(R)}{D(Y)}$$

$$DevTotale = D(Y) = \sum (y_i - \bar{y})^2 = 161366$$

$$Dev\,\mathrm{Re}\,sidua = D(E) = \sum (y_i - \bar{y})^2 * (1 - \rho^2) = 161366 * (1 - 0{,}3989) = 96997$$

$$DevSpiegata = D(R) = D(Y) - D(E) = 161366 - 96997 = 64369$$

$$R^2 = \frac{D(R)}{D(Y)} = \frac{64369}{161366} = 0{,}3989$$

Dal risultato di può notare come $R^2 = \rho^2$.

Inoltre, è possibile ottenere $R^2 = 1 - \frac{D(E)}{D(Y)} = 1 - \frac{96997}{161366} = 0{,}3989$