Distribuzioni Doppie e Relazioni tra Variabili — Esercitazione n° 04

ESERCIZIO 1

Tra le famiglie del Comune di Vigata sono stati rilevati congiuntamente la PROFESSIONE DEL CAPOFAMIGLIA (X) e l'AMMONTARE DELLA SPESA SETTIMANALE PER MEZZI DI TRASPORTO (Y). I dati espressi in € sono riportati nella seguente tabella:

Y / X	0 -|20	20 -|50	50 -|100	100 -|150	Totale
Libero Professionista	20	15	8	26	69
Lav. Dipendente	15	5	14	13	47
Disoccupato	5	40	3	8	56
Pensionato	10	16	2	4	32
Totale	50	76	27	51	204

1) costruire la tabella dei profili riga (distribuzioni condizionate Y|X)
2) tra le famiglie che spendono più di 50 € determinare la percentuale di quelle che hanno il capofamiglia in attività
3) stabilire se tra i due caratteri esiste dipendenza e in caso positivo misurarne l'intensità

---

1) Per costruire la tabella dei profili riga è necessario dividere ogni frequenza congiunta per la corrispondente frequenza marginale di riga:

Y / X	0 -|20	20 -|50	50 -|100	100 -|150	Totale
Libero Professionista	20/69	15/69	8/69	26/69	1
Lav. Dipendente	15/47	5/47	14/47	13/47	1
Disoccupato	5/56	40/56	3/56	8/56	1
Pensionato	10/32	16/32	2/32	4/32	1
Totale	50/204	76/204	27/204	51/204	1

In tal modo su ogni riga si ottiene la distribuzione condizionata $Y|x_i$

Y / X	0 -|20	20 -|50	50 -|100	100 -|150	Totale
Libero Professionista	29%	22%	12%	38%	100%
Lav. Dipendente	32%	11%	30%	28%	100%
Disoccupato	9%	71%	5%	14%	100%
Pensionato	31%	50%	6%	13%	100%
Totale	25%	37%	13%	25%	100%

Dalla tabella dei profili riga si evince come ad esempio il 38% delle famiglie con capofamiglia libero professionista sostenga ogni settimana una spesa per trasporti tra 100 e 150 €, mentre solo il 28% delle famiglie con capofamiglia lavoratore dipendente ha lo stesso livello di spesa.

2) Le famiglie che hanno una spesa settimanale per trasporti maggiore di 50€ sono quelle con un ammontare di spesa tra 50 e 100 € (27 famiglie) e quelle con un ammontare di spesa tra 100 e 150 € (51 famiglie), per un totale di 78 famiglie. Consideriamo in attività i lavoratori dipendenti ed i liberi professionisti, escludendo i disoccupati e coloro che non sono più in attività, cioè i pensionati. Tra le famiglie con capofamiglia libero professionista 34 spendono più di 50 € (8 + 26), mentre tra le famiglie con capofamiglia dipendente 27 spendono più di 50 € (14 + 13). La percentuale di famiglie con capofamiglia in attività e ammontare di spesa per trasporti superiore ai 50 € sarà allora pari a (27 + 34)/78 = 0,78 => 78%.

(3) E' possibile valutare per i due caratteri sia la dipendenza in distribuzione che la dipendenza in media. Valutiamo innanzi tutto la dipendenza in distribuzione. Si verifica la condizione di indipendenza calcolando una frequenza teorica e quindi confrontando tale quantità con la corrispondente frequenza osservata: possiamo vedere che $n_{11} = 20 \neq n_{11}^{*} = (n_{1.} \times n_{.1})/n = (50 \times 69) / 204 = 16,91$ e quindi cade l'ipotesi di indipendenza. Procediamo quindi con lo studio della connessione tra le due variabili calcolando, sotto l'ipotesi di indipendenza, le frequenze teoriche:

Y / X	0 -|20	20 -|50	50 -|100	100 -|150	Totale
Libero Professionista	16,91	25,71	9,13	17,25	69
Lav. Dipendente	11,52	17,51	6,22	11,75	47
Disoccupato	13,73	20,86	7,41	14,00	56
Pensionato	7,84	11,92	4,24	8,00	32
Totale	50	76	27	51	204

Possiamo a questo punto calcolare le contingenze per stabilire se c'è connessione e di che tipo

Y / X	0 -|20	20 -|50	50 -|100	100 -|150	Totale
Libero Professionista	3,09	-10,71	-1,13	8,75	0
Lav. Dipendente	3,48	-12,51	7,78	1,25	0
Disoccupato	-8,73	19,14	-4,41	-6,00	0
Pensionato	2,16	4,08	-2,24	-4,00	0
Totale	0	0	0	0	0

Dall'analisi delle contingenze si rileva come ci sia un forte addensamento di casi osservati per le famiglie di disoccupati con un ammontare di spesa tra 20 e 50 € rispetto all'ipotesi di indipendenza. Analogamente si rileva nella stessa classe di spesa un numero di casi osservati decisamente inferiore all'ipotesi di indipendenza per le famiglie con capofamiglia professionista o dipendente.
Essendo le contingenze tutte diverse da 0 dobbiamo concludere che i due caratteri non sono sconnessi e quindi procedere alla valutazione del grado di dipendenza. Calcoliamo innanzi tutto l'indice Chi-quadro:

χ^2 = 0,56 + 4,46 + 0,14 + 4,44 + 1,05 + 8,94 + 9,73 + 0,13 + 5,55 + 17,56 + 2,62 + 2,57 + 0,60 + 1,40 + 1,18 + 2,00 = 62,94

Essendo l'indice Chi-quadro diverso da zero abbiamo la conferma che i due caratteri non sono indipendenti. Ricorriamo all'indice V di Cramer per misurare l'intensità della dipendenza.
Il massimo del Chi-quadro rispetto alla tabella analizzata è (204 x 3) = 612 (numerosità del collettivo per il valore più piccolo tra dimensione di riga meno uno e di colonna meno uno). L'indice V è allora:

$V = (62,94/612)^{1/2} = 0,10^{1/2} = 0,32$ => 32% (della massima connessione osservabile)

Dalla lettura dell'indice di Cramer possiamo concludere che tra professione del capofamiglia e ammontare settimanale di spesa per trasporti c'è un livello di dipendenza statistica medio-basso.

Sapendo che l'indipendenza in distribuzione implica l'indipendenza in media, possiamo escludere che l'ammontare di spesa sia indipendente dalla professione del capofamiglia e procedere quindi alla misura del livello di dipendenza anche con il rapporto di correlazione di Pearson.

Calcoliamo innanzi tutto la media generale e le medie condizionate:

Y / X	0 -|20	20 -|50	50 -|100	100 -|150	Totale
Lav. Dipendente	20	15	8	26	69
Libero Professionista	15	5	14	13	47
Disoccupato	5	40	3	8	56
Pensionato	10	16	2	4	32
Totale	50	76	27	51	204

M(spesa|l.dip.) = [(10 x 20) + (35 x 15) + (75 x 8) + (125 x 26)]/69 = 4575/69 = 66,30 €
M(spesa|l.pro.) = [(10 x 15) + (35 x 5) + (75 x 14) + (125 x 13)]/47 = 3000/47 = 63,83 €
M(spesa|dis.) = [(10 x 5) + (35 x 40) + (75 x 3) + (125 x 8)]/56 = 2675/56 = 48,00 €
M(spesa|pens.) = [(10 x 10) + (35 x 16) + (75 x 2) + (125 x 4)]/32 = 1310/32 = 40,94 €
M(spesa) = [(10 x 50) + (35 x 76) + (75 x 27) + (125 x 51)]/204 = 11560/204 = 56,67 €

Potevamo ottenere lo stesso risultato considerando la proprietà dell'associatività della media:
M(spesa) = [(66,30 x 69) + (63,83 x 47) + (48 x 56) + (40,94 x 32)]/204 = 11560/204 = 56,67 €

Per calcolare il rapporto di correlazione è necessario a questo punto calcolare la DEV(T), che misura la variabilità totale, e la DEV(B), che invece misura la parte di variabilità tra i gruppi:

DEV(T) = $[(10 - 56,67)^2 \times 50] + [(25 - 56,67)^2 \times 76] + [(75 - 56,67)^2 \times 27] + [(125 - 56,67)^2 \times 51]$ =
= 108888,89 + 35677,78 + 9075,00 + 238141,67 = 391783

DEV(B) = $[(66,30 - 56,67)^2 \times 69] + [(63,83 - 56,67)^2 \times 47] + [(48 - 56,67)^2 \times 56]$ +
+ $[(40,94 - 56,67)^2 \times 32]$ = 6409,06 + 2411,58 + 4434,57 + 7917,01 = 21172,23

L'indice Eta-quadro risulta allora pari a:

$\eta^2_{Y|X}$ = 21172/391783 = 0,054 => 5,4% (della massima dipendenza in media osservabile)

Anche dall'analisi della dipendenza in media abbiamo quindi la conferma c'è una relazione tra ammontare di spesa e professione del capofamiglia, ma l'intensità è bassa.

ESERCIZIO 2

Nel corso del 2006 sono stati rilevati giornalmente il LIVELLO DEL TRAFFICO CITTADINO e la CONDIZIONE METEOROLOGICA:

	BASSO	MEDIO	ALTO	TOT
SERENO	100	13	5	118
VARIABILE	21	105	10	136
PIOGGIA	6	15	90	111
TOT	127	133	105	365

1) Qual è il collettivo di riferimento?
2) Fornire una spiegazione per le frequenze $\mathbf{n_{23}}$ e $\mathbf{n_{1.}}$
3) Stabilire se tra i due caratteri vi è dipendenza statistica

---

1) Il collettivo oggetto di studio in questo caso è l'insieme dei diversi giorni dell'anno 2006. La frequenza congiunta $\mathbf{n_{23}}$ rappresenta il numero di giorni in cui il tempo è stato *variabile* ed il livello di traffico *alto*; la frequenza marginale $\mathbf{n_{1.}}$ rappresenta invece il numero di giorni del 2006 in cui il livello di traffico è stato *basso*.

2) Per stabilire se c'è dipendenza tra i due caratteri e quindi nel caso, valutarne l'intensità, procediamo al calcolo delle frequenze teoriche sotto l'ipotesi di indipendenza in distribuzione (dopo aver verificato che cade l'ipotesi di indipendenza):

	BASSO	MEDIO	ALTO	TOT
SERENO	41,06	43,00	33,95	118
VARIABILE	47,32	49,56	39,12	136
PIOGGIA	38,62	40,45	31,93	111
TOT	127	133	105	365

Calcoliamo quindi la tabella delle contingenze:

	BASSO	MEDIO	ALTO	TOT
SERENO	58,94	-30,00	-28,95	0,00
VARIABILE	-26,32	55,44	-29,12	0,00
PIOGGIA	-32,62	-25,45	58,07	0,00
TOT	0,00	0,00	0,00	0,00

Si nota come sulla diagonale principale della tabella ci sia un addensamento di casi osservati superiore alla situazione teorica di indipendenza, mentre in tutte le altre celle i casi osservati sono stati inferiori a quelli che ci saremmo aspettati sotto l'ipotesi di indipendenza statistica.
Essendo tutte le contingenze diverse da zero escludiamo l'ipotesi di indipendenza e procediamo al calcolo dell'indice Chi quadro

χ^2 = 84,62 + 20,93 + 24,68 + 14,64 + 62,03 + 21,68 + 27,55 + 16,01 + 105,60 = 377,74

Il valore dell'indice ci conferma che i due caratteri sono connessi. Misuriamo il livello di dipendenza calcolando l'indice V di Cramer:

max χ^2 = 365 x min [(3-1);(3-1)] = 730

V = $(377,74/730)^{1/2}$ = $0,52^{1/2}$ = 0,72 => 72% (della max connessione osservabile)

Dalla lettura dell'indice di Cramer concludiamo che vi è un alto grado di dipendenza tra la condizione atmosferica e il livello di traffico.

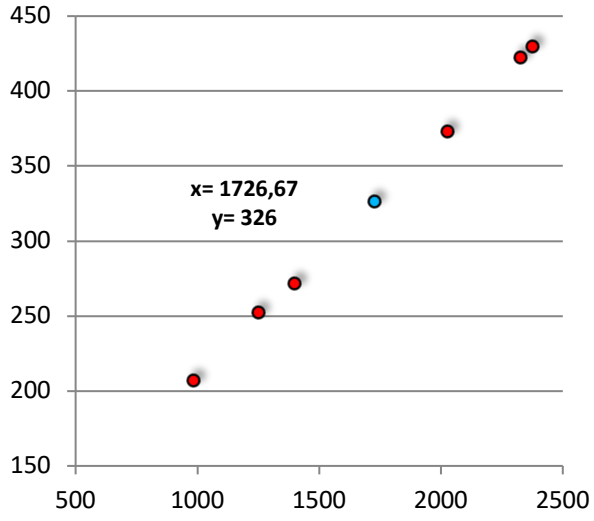
ESERCIZIO 3

Un'agenzia di viaggi è interessata a sapere se le tariffe aeree sono in relazione alla durata (in miglia) dei voli. L'agenzia ipotizza che più lungo è il volo, più costoso è il biglietto. Vengono allora raccolti i seguenti dati:

Distanza	2375	1400	1250	2325	985	2025
Tariffa (€)	430	272	252	422	207	373

Verificare se c'è concordanza tra tariffa e distanza percorsa e valutarne l'intensità.

Innanzi tutto costruiamo il diagramma di dispersione per la Distanza (X) e la Tariffa (Y)



Osservando il diagramma è plausibile affermare che i due caratteri sono concordi e che c'è una forte relazione di dipendenza lineare.

Possiamo calcolare la Tariffa e la Distanza media:

TARIFFA MEDIA = (430+272+252+422+207+373)/6=326

DISTANZA MEDIA = 1726,67

Pur senza disegnare il grafico degli scostamenti si osserva in generale la presenza di scostamenti concordi

Per valutare la concordanza media tra Distanza e Tariffa calcoliamo la COVARIANZA:

$$COV(X,Y) = \Sigma[x_i - M(X)][y_j - M(Y)]/n = \Sigma\, x_i y_j/n - M(X)M(Y)$$

Distanza	**2375**	**1400**	**1250**	**2325**	**985**	**2025**	
Tariffa (€)	**430**	**272**	**252**	**422**	**207**	**373**	
D x T	1021250	380800	315000	981150	203895	755325	3657420

COV (X,Y) = 3657420/6 - 1726,67 x 326 = 46676,67

Tra i due caratteri c'è una relazione di dipendenza e sono concordi. Per valutare l'intensità della concordanza calcoliamo il coefficiente di correlazione

r(X,Y) = COV(X,Y)/[s.q.m(X) x s.q.m.(Y)]

s.q.m(X) = 540,26 s.q.m.(Y) = 86,40 r(X,Y) = 46676,67/[540,26 x 86,40] = 0,999

Tra i due caratteri c'è una altissima correlazione positiva. L'ipotesi dell'agenzia è confermata.

ESERCIZIO 4

Una grande università americana è interessata a sapere se sussiste una qualche relazione tra lo stipendio iniziale di un laureato (in migliaia di dollari) e il numero di corsi di statistica che ha seguito come studente. I dati analizzati sono i seguenti:

corsi seguiti	1	1	2	3	3	4
retribuzione (migliaia $)	26	25	26	27	28	30

Verificare se c'è una relazione tra le variabili e valutarne l'intensità.

ESERCIZIO 5

Una ditta è interessata a comprendere la relazione esistente tra il numero di giorni dedicati alla formazione degli impiegati per svolgere un particolare compito e i risultati ottenuti sulla base di un test. L'ufficio risorse umane raccoglie i seguenti dati:

giorni di formazione	1,0	1,5	2,0	2,5	2,0	3,0
punteggio test	41	60	72	91	80	95

Verificare se c'è una relazione tra le variabili e valutarne l'intensità.

ESERCIZIO 6

Nei paesi in via di sviluppo l'acqua inquinata è spesso connessa alle malattie. Nella tabella di seguito sono riportati i dati della speranza di vita nel 1994 in 10 paesi in via di sviluppo, insieme con la percentuale della loro popolazione che usava acqua inquinata nel periodo 1990-1996.

Durata media di vita	% di popolazione con accesso solo a fonti idriche inquinate
79,0	5
72,4	29
70,1	15
69,3	35
68,2	20
55,9	43
57,2	71
54,4	72
45,6	52
33,6	66

C'è una relazione tra l'uso di acqua inquinata e la speranza di vita?

ESERCIZIO 7

È stato studiato il voto di laurea degli studenti di Economia Aziendale e il reddito annuale (in €) del padre, riportando i seguenti dati:

Voto di Laurea	Reddito (in €)
98	20000
102	25000
95	21500
90	27500
110	32000
104	20500
94	40000
85	25500
100	28000
95	20000

Analizzare la relazione tra i due caratteri

SOLUZIONI

ESERCIZIO 4 => COV(X,Y) = 1,667 r(X,Y) = 0,923
ESERCIZIO 5 => COV(X,Y) = 11,583 r(X,Y) = 0,970
ESERCIZIO 6 => COV(X,Y) = -240,646 r(X,Y) = -0,802
ESERCIZIO 7 => COV(X,Y) = 900 r(X,Y) = 0,02