ESERCIZIO 8.1

Un produttore di autovetture è interessato a determinare se vi sia indipendenza tra cilindrata dell'automobile acquistata e marca della stessa al fine di individuare i target di clienti cui mirare. Un campione di 1000 recenti acquirenti di automobili è selezionato casualmente, classificando ogni acquirente rispetto alla cilindrata dell'automobile e alla marca scelta. I risultati della rilevazione sono riportati nella seguente tabella:

		MARCA				
		A	B	C	D	
CILINDRATA	Piccola	157	65	181	10	**413**
	Media	126	82	142	46	**396**
	Grande	58	45	60	28	**191**
		341	**192**	**383**	**84**	**1000**

a) I due caratteri considerati possono essere considerati indipendenti ad un livello di significatività del 5%?
b) Si calcoli inoltre il livello di significatività osservato del test (p-value)

SVOLGIMENTO

a)

1) Ipotesi nulla

Ho: I due caratteri Marca e Cilindrata dell'autovettura sono indipendenti

2) Ipotesi alternativa

Ha: I due caratteri Marca e Cilindrata dell'autovettura sono dipendenti

3) Statistica test

Statistica test → $\sum_{i,j} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} \sim \chi^2_{(r-1)\times(c-1)}$

Dove:

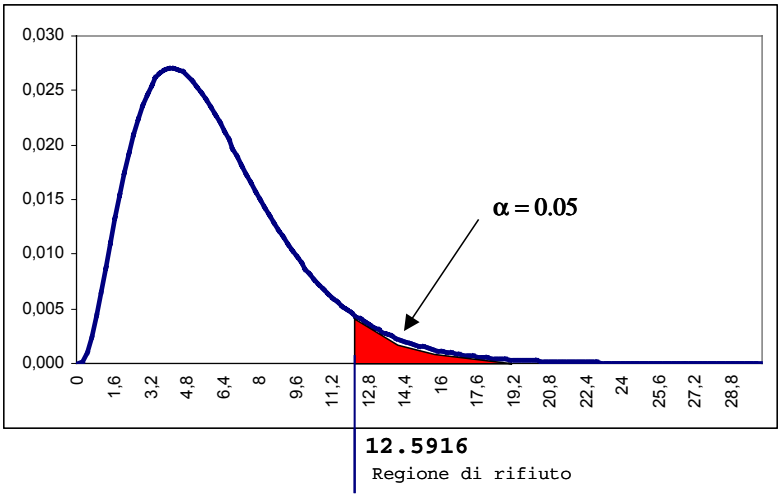
r → numero di righe della tabella a doppia entrata

c → numero di colonne della tabella a doppia entrata

$\hat{n}_{i,j} = \frac{n_{i\bullet} n_{\bullet j}}{N}$ → frequenza teorica per la cella i,j calcolata sotto l'ipotesi di indipendenza

4) Regola di decisione

α=0.05 → $\chi^2_{(r-1)\times(c-1),\alpha} = 12.5916$



⇒ Rifiuto Ho se la statistica test è maggiore di 12.5916

5) Assunzioni (ipotesi mantenute)

- Le N frequenze osservate sono un campione casuale della popolazione di interesse
- L'ampiezza del campione è sufficientemente grande in maniera che ciascuna frequenza attesa (teorica) è maggiore o uguale a 5

6) Esperimento sul campione

		MARCA				
		A	B	C	D	
CILINDRATA	Piccola	157	65	181	10	**413**
	Media	126	82	142	46	**396**
	Grande	58	45	60	28	**191**
		341	**192**	**383**	**84**	**1000**

Per calcolare la statistica test è necessario calcolare le frequenze teoriche sotto l'ipotesi di indipendenza. Ciascuna cella è ottenuta moltiplicando i rispettivi marginali di riga e colonna e dividendo per l'ampiezza del campione:

		MARCA				
		A	B	C	D	
CILINDRATA	Piccola	140,83	79,30	158,18	34,69	**413**
	Media	135,04	76,03	151,67	33,26	**396**
	Grande	65,13	36,67	73,15	16,04	**191**
		341	**192**	**383**	**84**	**1000**

Statistica test → $\sum_{i,j} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} = 45.81$

7) Conclusione

La statistica test, per il campione considerato, produce un valore nella regione di rifiuto: i dati empirici permettono di accettare l'ipotesi che vi sia dipendenza tra la cilindrata e la marca dell'autovettura scelta.

b)

Il livello di significatività osservato del test, noto anche come p-value o valore p, è la probabilità (sotto l'ipotesi che H_0 sia vera) di osservare un valore che sia almeno tanto contraddittorio rispetto all'ipotesi nulla e tale da avvalorare l'ipotesi alternativa quanto quello calcolato sui dati campionari a disposizione. E' possibile cercare il valore di χ^2=45.81 dalle tavole della distribuzione X^2 facendo riferimento alla riga relativa a 6 gradi di libertà:

$P(X^2 > \chi^2)$

	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	0,00004	0,00016	0,00098	0,00393	0,01579	2,70554	3,84146	5,02390	6,63489	7,87940
2	0,01002	0,02010	0,05064	0,10259	0,21072	4,60518	5,99148	7,37778	9,21035	10,59653
3	0,07172	0,11483	0,21579	0,35185	0,58438	6,25139	7,81472	9,34840	11,34488	12,83807
4	0,20698	0,29711	0,48442	0,71072	1,06362	7,77943	9,48773	11,14326	13,27670	14,86017
5	0,41175	0,55430	0,83121	1,14548	1,61031	9,23635	11,07048	12,83249	15,08632	16,74965
6	0,67573	0,87208	1,23734	1,63538	2,20413	10,64464	12,59158	14,44935	16,81187	18,54751

Da cui si vede come il valore che più si avvicina a 45.81 è il valore 18.54751, che lascia a destra solo 0.005: il test risulta quindi significativo, confemando quanto ottenuto con la procedura standard seguita al passo a. Per ottenere un valore più preciso per il p-value si può ricorrere ad un qualunque software con funzioni di tipo statistico (MS-Excel restituisce nel caso in questione un p-value pari a 0.00000003)

ESERCIZIO 8.2

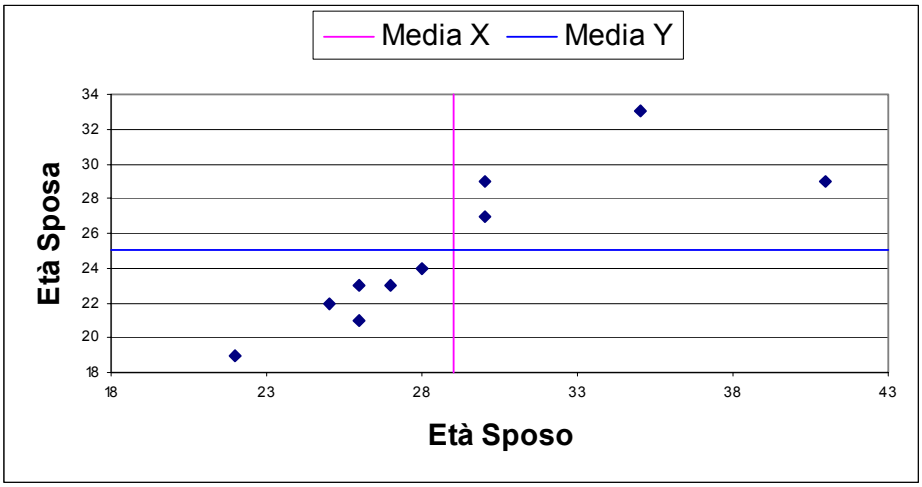
Il sindaco di un piccolo paesino è interessato a misurare la relazione esistente tra età dello sposo e età della sposa per i matrimoni celebrati tra i suoi concittadini. I dati relativi ai matrimoni celebrati nell'ultimo anno sono riportati nella seguente tabella:

Età Sposo	Età Sposa
22	19
25	22
26	21
26	23
27	23
28	24
30	29
30	27
35	33
41	29

1) Rappresentare le due serie e le medie delle due variabili usando un grafico a dispersione
2) Calcolare la covarianza tra le due variabili
3) Calcolare la correlazione tra le due variabili

SVOLGIMENTO

1)



2)
I calcoli sono riportati nella seguente tabella:

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
	X: Età Sposo	Y: Età Sposa	X-E(X)	$[X-E(X)]^2$	Y-E(Y)	$[Y-E(Y)]^2$	[X-E(X)] [Y-E(Y)]	
	22	19	-7	49	-6	36	42	
	25	22	-4	16	-3	9	12	
	26	21	-3	9	-4	16	12	
	26	23	-3	9	-2	4	6	
	27	23	-2	4	-2	4	4	
	28	24	-1	1	-1	1	1	
	30	29	1	1	4	16	4	
	30	27	1	1	2	4	2	
	35	33	6	36	8	64	48	
	41	29	12	144	4	16	48	
Tot.	290	250		270	Devianza	170	179	Codevianza
Media	29	25						

Da cui:

$$COV(X,Y)=\sigma_{XY}=\frac{\sum_i (x_i-\mu_X)(y_i-\mu_Y)}{n}=\frac{179}{10}=17.9$$

E' possibile calcolare la covarianza sfruttando la seguente formula:

$$COV(X,Y)=\frac{\sum_i x_i y_i}{n}-\frac{\sum_i x_i}{n}\frac{\sum_i y_i}{n}$$

come mostrato di seguito:

	X: Età Sposo	Y: Età Sposa	X Y
	22	19	418
	25	22	550
	26	21	546
	26	23	598
	27	23	621
	28	24	672
	30	29	870
	30	27	810
	35	33	1155
	41	29	1189
Tot.	290	250	7429
Media	29	25	742,9

$$COV(X,Y)=\frac{\sum_i x_i y_i}{n}-\frac{\sum_i x_i}{n}\frac{\sum_i y_i}{n}=742.9-29\times 25=17.9$$

3)

$$\rho(X,Y)=\frac{CODEV(X)}{\sqrt{DEV(X)DEV(Y)}}=\frac{\sum_i (x_i-\mu_X)(y_i-\mu_Y)}{\sqrt{\sum_i (x_i-\mu_X)^2\sum_i (y_i-\mu_Y)^2}}=\frac{179}{\sqrt{270\times 170}}=0.84$$