Istituzioni di Statistica 1
Esercizi su tabelle di contingenza

Esercizio 1

Per stimare la percentuale di fumatori nella popolazione italiana adulta viene intervistato un campione di 60 donne e uno di 40 uomini, ottenendo le seguenti risposte:

Uomini 0 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0
0 0 0 0 1 0 0 0 0 1 1 0 0 1 0

Donne 1 0 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1 0

dove 1=fumatore e 0=non fumatore

(a) Ricavare una tabella di distribuzione doppia

	Fumo	Non Fumo	ToT
Uomini	13	27	40
Donne	14	46	60
ToT	27	73	100

(b) Determinare la percentuale di fumatori tra gli intervistati.

Si tratta di una frequenza relativa marginale:

$$\frac{27}{100} = 0,27 \Rightarrow 27\%$$

(c) Si determini la percentuale di intervistati che sono di sesso femminile e fumano.

Si tratta di una frequenza relativa congiunta:

$$\frac{14}{100} = 0,14 \Rightarrow 14\%$$

(d) Quale percentuale fuma tra gli uomini?

Si tratta di una frequenza relativa condizionata: la modalità condizionante è "Uomini", la modalità condizionata è "Fumo"

$$\frac{13}{40} = 0,325 \Rightarrow 32,\%$$

(e) Quale percentuale di fumatori è di sesso femminile?

Si tratta di una frequenza relativa condizionata: la modalità condizionante è "Fumo", la modalità condizionata è "Donne"

$$\frac{14}{27} = 0,52 \Rightarrow 52\%$$

(f) Usando un opportuno indice dire se sulla base dei dati c'è evidenza di associazione tra il fumo e il sesso.

Un indice appropriato è l'indice chi-quadrato

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

dove

$$n^*_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{N}$$

sono le frequenze teoriche sotto l'ipotesi di indipendenza.
Costruiamo la tabella delle frequenze teoriche da cui

	Fumo	Non Fumo	ToT
Uomini	$\frac{27\cdot 40}{100} = 10,8$	$\frac{73\cdot 40}{100} = 29,2$	40
Donne	$\frac{27\cdot 60}{100} = 16,2$	$\frac{73\cdot 60}{100} = 43,8$	60
ToT	27	73	100

$$\chi^2 = \frac{(13-10,8)^2}{10,8} + \frac{(27-29,2)^2}{29,2} + \frac{(14-16,2)^2}{16,2} +$$

$$+\frac{(46-43,8)^2}{43,8} = 1,023$$

La sua versione normalizzata è

$$T = \frac{1,023}{100} = 0,01$$

Le due variabili sono quindi vicine all'indipendenza.

Esercizio 2

Alla fine di una giornata di lavoro un intervistatore si accorge di aver perso i dati raccolti su un certo numero di famiglie relativamente al numero X di televisori posseduti e al numero Y di componenti della famiglia. Ricostruendo a memoria le interviste fatte, arriva alla seguente tabella

	Numero componenti Y			
Numero televisori X	1	2	3	Totale
0	0	3	1	?
1	3	?	7	16
2	1	?	?	?
Totale	?	13	11	?

1. Si completi la tabella e si dica, senza fare calcoli, se i due caratteri sono indipendenti.

	Numero componenti Y			
Numero televisori X	1	2	3	Totale
0	0	3	1	4
1	3	6	7	16
2	1	4	3	8
Totale	4	13	11	28

Non possono essere indipendenti, in quanto

$$0 \neq \frac{4 \cdot 4}{28}$$

2. Si calcolino le medie e le varianze del numero di televisori condizionatamente al numero di componenti della famiglia.

$$M(X|Y=1) = \frac{0\cdot 0 + 1\cdot 3 + 2\cdot 1}{4} = 5/4$$
$$M(X|Y=2) = \frac{0\cdot 3 + 1\cdot 6 + 2\cdot 4}{13} = 14/13$$
$$M(X|Y=3) = \frac{0\cdot 1 + 1\cdot 7 + 2\cdot 3}{11} = 13/11$$
$$M(X|Y=1) = \frac{0^2\cdot 0 + 1^2\cdot 3 + 2^2\cdot 1}{4} = 7/4$$
$$M(X|Y=2) = \frac{0^2\cdot 3 + 1^2\cdot 6 + 2^2\cdot 4}{13} = 22/13$$
$$M(X|Y=3) = \frac{0^2\cdot 1 + 1^2\cdot 7 + 2^2\cdot 3}{11} = 19/11$$
$$V(X|Y=1) = \frac{7}{4} - \frac{5^2}{4^2}$$
$$V(X|Y=2) = \frac{22}{13} - \frac{14^2}{13^2}$$
$$V(X|Y=3) = \frac{19}{11} - \frac{13^2}{11^2}$$

Esercizio 3

Si consideri la seguente tabella relativa alla distribuzione di un gruppo di studenti che hanno superato la prova scritta di un esame a quiz secondo il tempo impiegato (in minuti) e il voto conseguito:

	Voto				
Tempo	$18 \vdash 21$	$22 \vdash 24$	$25 \vdash 27$	$28 \vdash 30$	ToT
$30 \dashv 60$		3	2	3	10
$60 \dashv 90$	4	6		6	20
$90 \dashv 120$	4		4	6	20
$120 \dashv 150$	6			9	30
$150 \dashv 180$	4	6	4	6	20
ToT		30			

(a) Completare la tabella.

	Voto				
Tempo	$18 \vdash 21$	$22 \vdash 24$	$25 \vdash 27$	$28 \vdash 30$	ToT
$30 \dashv 60$	2	3	2	3	10
$60 \dashv 90$	4	6	4	6	20
$90 \dashv 120$	4	6	4	6	20
$120 \dashv 150$	6	9	6	9	30
$150 \dashv 180$	4	6	4	6	20
ToT	20	30	20	30	100

(b) Calcolare la precentuale di studenti con un voto maggiore o uguale a 28 che hanno impiegato meno di 90 minuti.

Si tratta di una frequenza relativa condizionata: l'evento condizionante è "voto$\geq$ 28", l'evento condizionato è "tempo $<$ 90"

$$\frac{3+6}{30} = 0,3 \Rightarrow 30\%$$

(c) Calcolare la percentuale di studenti tra quelli che hanno impiegato un tempo compreso tra 90 e 150 minuti che hanno preso un voto maggiore o uguale a 25 e minore o uguale a 27.

Si tratta di una frequenza condizionata: l'evento condizionante è "tempo compreso tra 90 e 150 minuti", la modalità condizionata è "$25 \leq$ voto ≤ 27"

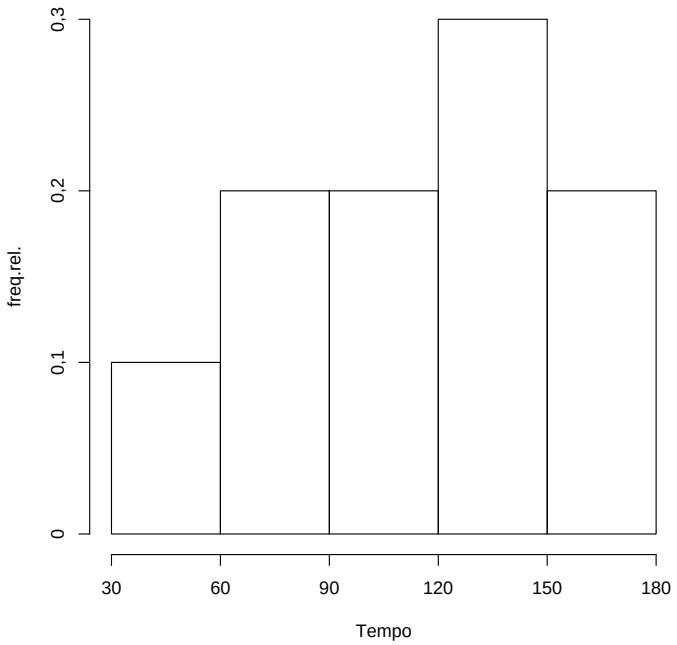
$$\frac{4+6}{20+30} = 0,2 \Rightarrow 20\%$$

(d) Scrivere la distribuzione di frequenza marginale della variabile "Tempo" e disegnarne l'istogramma.

Poiché le classi hanno uguale ampiezza, l'istogram-

Tempo	Freq. Assolute	Freq. Relative
$30 \dashv 60$	10	10/100=0,1
$60 \dashv 90$	20	20/100=0,2
$90 \dashv 120$	20	20/100=0,2
$120 \dashv 150$	30	30/100=0,3
$150 \dashv 180$	20	20/100=0,2

ma può essere costruito mettendo in altezza le frequenze (assolute o relative). Nel grafico sottostante nelle ordinate sono riportate le frequenze relative.



Dal grafico osserviamo una certa asimmetria negativa.

(e) Calcolare moda e mediana della variabile "Tempo".

Dal grafico vediamo che la classe modale è la classe $120 \dashv 150$.
Sommando le frequenze relative, notiamo che al valore 120 corrisponde la frequenza relativa cumulata pari a 0,5; pertanto, la mediana è 120 minuti.

(f) Scrivere le distribuzioni di frequenza relativa dei voti condizionate al tempo impiegato.

	Voto				
Tempo	$18 \vdash\dashv 21$	$22 \vdash\dashv 22$	$25 \vdash\dashv 27$	$28 \vdash\dashv 30$	
$30 \dashv 60$	2/10=0,2	3/10=0,3	2/10=0,2	3/10=0,3	1
$60 \dashv 90$	4/20=0,2	6/20=0,3	4/20=0,2	6/20=0,3	1
$90 \dashv 120$	4/20=0,2	6/20=0,3	4/20=0,2	6/20=0,3	1
$120 \dashv 150$	6/30=0,2	9/30=0,3	6/30=0,2	9/30=0,3	1
$150 \dashv 180$	4/20=0,2	6/20=0,3	4/20=0,2	6/20=0,3	1

Notiamo che tutte le distribuzioni delle frequenze relative di "Voto" condizionate alle modalità della variabile "Tempo" sono uguali (e pari alla distribuzione delle frequenze relative marginale di "Voto"), da cui concludiamo che le due variabili sono indipendenti.

(g) Calcolare il voto medio e la sua deviazione standard.

$$\text{Voto medio} = \frac{1}{100}(19,5 \cdot 20 + 23 \cdot 30 + 26 \cdot 20 + 29 \cdot 30) = 24,7$$

$$\text{Momento secondo} = \frac{1}{100}(19,5^2 \cdot 20 + 23^2 \cdot 30 + 26^2 \cdot 20 + \\ + 29^2 \cdot 30) = 622,25$$

varianza del voto=622,25-$24,7^2$=12,16

deviazione standard= $\sqrt{12,16} = 3,49$

Esercizio 4

Il proprietario di un negozio di computer vuole sapere quanto velocemente vengono saldate le fatture relative ai PC per tre diverse tipologie di clienti (A=Enti Pubblici, B=Aziende, C=Privati). A questo fine, riporta, per le fatture saldate negli ultimi mesi, la tipologia del cliente e i giorni intercorsi tra la consegna dei PC e il saldo della fattura, ottenendo i risultati riassunti nella tabella seguente:

Giorni trascorsi consegna-saldo	Tipologia del cliente A	B	C
$0 \dashv 10$	26	52	40
$10 \dashv 20$	42	60	46
$20 \dashv 30$	12	18	14

(a) Si rappresenti graficamente la distribuzione di frequenza marginale della variabile "Tipologia del cliente".

(b) Si rappresenti graficamente la distribuzione di frequenza marginale della variabile "Numero di giorni intercorsi tra la consegna e il saldo".

(c) Quale è la percentuale di Enti Pubblici tra i clienti che hanno saldato la fattura più di 10 giorni dopo la consegna del PC?

(d) Quale è la percentuale di fatture emesse ad Enti Pubblici che sono state saldate più di 10 giorni dopo la consegna del PC?

(e) Attraverso un opportuno indice si valuti se esiste una relazione tra la tipologia del cliente e il numero di giorni trascorsi tra la consegna e il saldo.

Soluzioni Esercizio 4

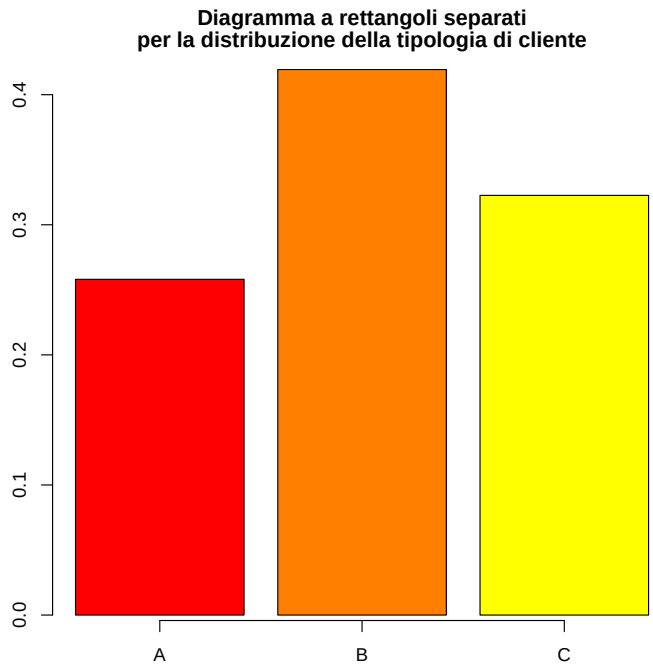
(a) Per costruire le distribuzioni marginali delle due variabili rappresentate in tabella, calcoliamo i totali di riga e di colonna, ottenendo:

Giorni trascorsi consegna-saldo	Tipologia del cliente A	B	C	ToT
$0 \dashv 10$	26	52	40	118
$10 \dashv 20$	42	60	46	148
$20 \dashv 30$	12	18	14	44
ToT	80	130	100	310

La distribuzione di frequenza relativa marginale per la variabile "Tipologia del cliente" è allora

Tipologia del cliente	Frequenza relativa
A	80/310=0,26
B	130/310=0,42
C	100/310=0,32
ToT	1

Trattandosi di una variabile qualitativa, uno strumento grafico adatto a rappresentare la sua distribuzione di frequenza è il diagramma a rettangoli separati, quale quello contenuto nella figura sottostante, in cui viene evidenziata la prevalenza delle Aziende sulle altre tipologie di clienti.

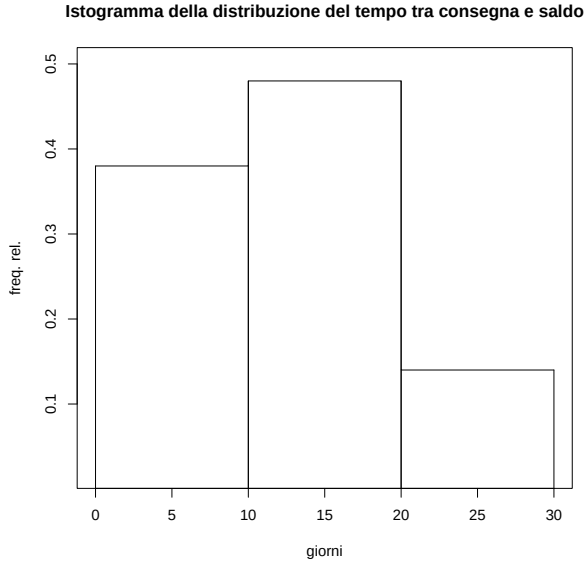


(b) La distribuzione delle frequenze relative della variabile "Numero di giorni intercorsi tra la consegna e il saldo" è

Giorni trascorsi consegna-saldo	Frequenze relative
$0 \dashv 10$	118/310=0,38
$10 \dashv 20$	148/310=0,48
$20 \dashv 30$	44/310=0,14
ToT	1

Trattandosi di una variabile quantitativa discreta, ma con numerose modalità, uno strumento grafico appropriato per rappresentarne la distribuzione di frequenza è l'istogramma. Poiché le classi hanno tutte uguale ampiezza, le altezze dei rettangoli che compongono l'istogramma possono essere prese pari alle frequenze relative, come nel grafico nel seguito riportato.

Il basso numero di classi in cui è stata suddivisa la



variabile rende difficile interpretare il grafico, e quindi la distribuzione della variabile, oltre all'osservazione che la classe $10 \dashv 20$ è la classe modale.

(c) Viene richiesta una frequenza relativa condizionata, in cui l'evento condizionante è "il cliente salda la fattura più di 10 giorni dopo la consegna del PC" e l'evento condizionato è "il cliente è un Ente Pubblico". Per calcolare questa frequenza condizionata, utilizziamo l'espressione frequenza congiunta su frequenza marginale dell'evento condizionate. La frequenza assoluta congiunta dei due eventi è 42+12=54. La frequenza assoluta marginale dell'evento condizionante è 148+44=192. Pertanto, la frequenza relativa richiesta è 54/192=0,28, ossia 28%.

(d) Anche in questo caso viene richiesta una frequenza relativa condizionata. L'evento condizionante è "il cliente è un Ente Pubblico" e l'evento condizionato è "il cliente salda la fattura più di 10 giorni dopo la consegna del PC". La frequenza assoluta congiunta è 42+12=54, come al punto precedente, mentre la fre-

quenza assoluta marginale dell'evento condizionante è 80. Pertanto, la frequenza richiesta è 54/80=0,675, ossia 67,5%.

(e) Un indice attraverso il quale si può valutare il grado di dipendenza tra le due variabili è l'indiche chi-quadro o una sua versione normalizzata. Per costruire l'indice chi-quadro abbiamo bisogno della tabella delle frequenze teoriche sotto l'ipotesi di indipendenza stocastica tra le due variabili, che viene di seguito riportata.

Giorni trascorsi consegna-saldo	Tipologia del cliente A	B	C
$0 \dashv 10$	$\frac{118 \cdot 80}{310} = 30,45$	$\frac{118 \cdot 130}{310} = 49,48$	$\frac{118 \cdot 100}{310} = 38,06$
$10 \dashv 20$	38,19	62,06	47,74
$20 \dashv 30$	11,35	18,45	14,19

L'indiche chi-quadro è

$$\chi^2 = \frac{(26-30,45)^2}{30,45} + \frac{(52-49,48)^2}{49,48} + \ldots + \frac{(14-14,19)^2}{14,19} = 1,44$$

e una sua versione normalizzata è

$$\Phi = \frac{\chi^2}{N \min(r-1, c-1)} = \frac{1,44}{310 \cdot 2} = 0,0023,$$

da cui si deduce che le due variabili sono molto vicine all'indipendenza, vale a dire, la tipologia del cliente non influisce sulla velocità con cui le fatture vengono saldate.

Esercizio 5

Un'azienda vuole conoscere se la soddisfazione nel lavoro può essere determinata anche dallo stipendio del dipendente. A questo scopo ha intervistato tutti i suoi dipendenti e li ha classificati in base al salario lordo mensile, come riportato nella tabella seguente.

Salario (migliaia di euro)	Grado di soddisfazione	
	Insoddisfatto	Soddisfatto
$1 \vdash 2$	36	4
$2 \vdash 3{,}5$	45	38
$3{,}5 \vdash 5{,}5$	22	54
$5{,}5 \vdash 8$	0	15

a) Attraverso due diagrammi a scatola (*boxplot*) si confrontino le distribuzioni del salario dei dipendenti che si dichiarano insoddisfatti e dei dipendenti che si dichiarano soddisfatti; si commenti quanto evidenziato dal confronto grafico.

b) Si può dire che le due variabili "Salario" e "Grado di Soddisfazione" sono indipendenti? (Si giustifichi la risposta.)

c) Si calcoli la percentuale di dipendenti, tra quelli che si dichiarano insoddisfatti, con un salario lordo mensile inferiore a 3 mila euro.

d) Si calcoli la percentuale di insoddisfatti tra coloro che hanno un salario lordo mensile inferiore a 3 mila euro.

e) Si suppongano fissate, e pari a quelle deducibili dalla tabella, le distribuzioni di frequenza marginale delle due variabili. Scrivere la distribuzione di frequenza relativa della variabile "Salario" condizionata alla modalità "Insoddisfatto" della variabile "Grado di soddisfazione", sotto l'ipotesi di indipendenza delle due variabili.

f) Alla luce dei risultati dell'indagine, la direzione dell'azienda decide di intervenire, aumentando di 200 euro il salario lordo mensile dei dipendenti che percepiscono un salario inferiore a 2 mila euro. Si calcoli la media e la varianza dello stipendio lordo mensile dei dipendenti dell'azienda in seguito a questo intervento.

Soluzioni Esercizio 5

a) Per costruire i due *boxplot* dobbiamo calcolare i quartili e i valori minimo e massimo delle distribuzioni del salario lordo mensile condizionate alle due modalità del grado di soddisfazione, lavorando separatamente sulle due colonne centrali della tabella

Salario (migliaia di euro)	Grado di soddisfazione Insoddisfatto	Soddisfatto	Tot
$1 \vdash 2$	36	4	40
$2 \vdash 3{,}5$	45	38	83
$3{,}5 \vdash 5{,}5$	22	54	76
$5{,}5 \vdash 8$	0	15	15
Tot	103	111	214

Calcolando le frequenze relative cumulate, si ottiene

Salario	Freq. rel. cumulate insoddisfatti	Freq. rel. cumulate soddisfatti
$1 \vdash 2$	0,35	0,04
$2 \vdash 3{,}5$	0,79	0,38
$3{,}5 \vdash 5{,}5$	1	0,86
$5{,}5 \vdash 8$	1	1

- Per i dipendenti che si dichiarano insoddisfatti:

Min=1

Max=5,5

Il primo quartile (il quantile 0,25) cade nella classe $1 \vdash 2$ e, sotto ipotesi di uniformità all'interno della classe, è

$$Q_I = 1 + (2-1)\frac{0,25-0}{0,35} = 1,71 \text{ migliaia di euro}$$

La mediana cade nella classe $2 \vdash 3,5$ e, sotto ipotesi di uniformità all'interno della classe, è

$$Me = 2+(3,5-2)\frac{0,5-0,35}{0,79-0,35} = 2,51 \text{ migliaia di euro}$$

Il terzo quartile (il quantile 0,75) cade anch'esso nella classe $2 \vdash 3,5$ e, sotto ipotesi di uniformità all'interno della classe, è

$$Q_{III} = 2+(3,5-2)\frac{0,75-0,35}{0,79-0,35} = 3,36 \text{ migliaia di euro}$$

- Per i dipendenti che si dichiarano soddisfatti:

Min=1
Max=8

Il primo quartile cade nella classe $2 \vdash 3{,}5$ e, sotto ipotesi di uniformità all'interno della classe, è

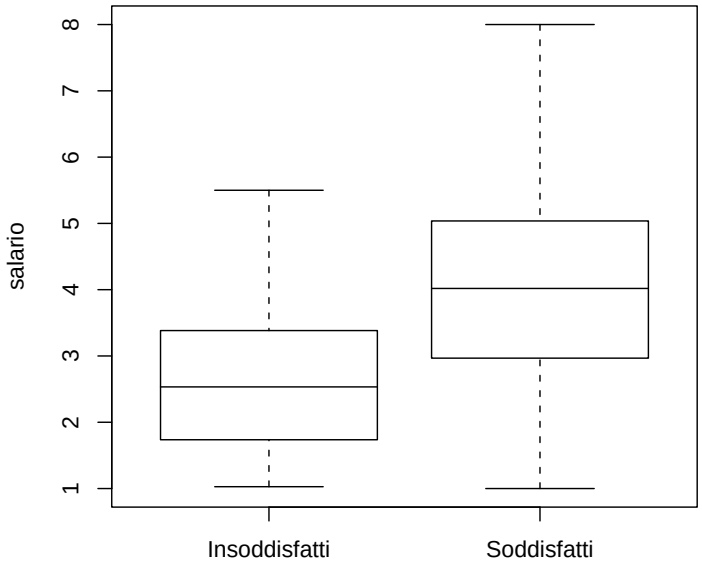
$$Q_I = 2+(3,5-2)\frac{0,25-0,04}{0,38-0,04} = 2,93 \text{ migliaia di euro}$$

La mediana cade nella classe $3,5 \vdash 5,5$ e, sotto ipotesi di uniformità all'interno della classe, è

$$Me = 3,5+(5,5-3,5)\frac{0,5-0,38}{0,86-0,38} = 4 \text{ migliaia di euro}$$

Il terzo quartile cade anch'esso nella classe $3,5 \vdash 5,5$ e, sotto ipotesi di uniformità all'interno della classe, è

$$Q_{III} = 3,5+(5,5-3,5)\frac{0,75-0,38}{0,86-0,38} = 5,04 \text{ migliaia di euro}$$



Dai due *boxplot* si nota che vi è una leggera asimmetria positiva della distribuzione del salario per i dipendenti che si dichiarano insoddisfatti, mentre la distribuzione del salario dei dipendenti che si dichiarano soddisfatti è simmetrica. Si nota dalla posizione dei due *boxplot* sull'asse verticale che il salario dei dipendenti soddisfatti tende ad essere più alto del salario dei dipendenti insoddisfatti. Possiamo essere ancora più precisi, notando che tutti i quantili dei dipendenti soddisfatti sono maggiori o uguali dei corrispondenti quantili per i dipendenti insoddisfatti; pertanto possiamo concludere che il salario dei dipendenti soddisfatti è statisticamente superiore al salario dei dipendenti insoddisfatti.

b) Se le due variabili fossero indipendenti, le distribuzioni condizionate della variabile "Salario" date le modalità della variabile "Grado di soddisfazione" dovreb-

bero essere uguali, ossia i due *boxplot* dovrebbero essere identici. Questo evidentemente non si verifica e possiamo concludere che esiste una forma di dipendenza tra le due variabili.

c) Si tratta di una frequenza relativa condizionata alla modalità "insoddisfatto". La frequenza assoluta congiunta di "salario inferiore a 3 mila euro" e "insoddisfatto", sotto ipotesi di uniformità all'interno della classe $2 \vdash 3,5$ è

$$36 + 45 \times \frac{3-2}{3,5-2} = 66$$

La percentuale richiesta è quindi

$$\frac{66}{103} \cdot 100\% = 64\%$$

d) Si tratta di una frequenza relativa condizionata a "salario inferiore a 3 mila euro". Dal punto c) la frequenza assoluta congiunta di "salario inferiore a 3 mila euro" e "insoddisfatto" è 66; la frequenza assoluta marginale di "salario inferiore a 3 mila euro" si deduce dalla distribuzione delle frequenze assolute marginali del "Salario", e, sotto ipotesi di uniformità nella classe $2 \vdash 3,5$, risulta pari a

$$40 + 83 \times \frac{3-2}{3,5-2} = 95 \qquad \text{approssimato}$$

Pertanto, la percentuale cercata è

$$\frac{66}{95} \times 100\% = 69\%$$

e) Se le due variabili fossero indipendenti, le distribuzioni di frequenza relativa della variabile "Salario" condizionate alle modalità della variabile "Grado di soddisfazione" dovrebbero essere entrambe uguali alla distribuzione delle frequenze relative marginali della variabile "condizionata", ossia del "Salario". Dalla tabella del punto (a) possiamo derivare la distribuzione delle frequenze relative marginali del salario

Salario	Frequenze relative
$1 \vdash 2$	0,19
$2 \vdash 3{,}5$	0,39
$3{,}5 \vdash 5{,}5$	0,36
$5{,}5 \vdash 8$	0,07

che, sotto ipotesi di indipendenza, rappresenta anche la distribuzione delle frequenze relative del "Salario" condizionata alla modalità "Insoddisfatto" (e "Soddisfatto") della variabile "Grado di soddisfazione".

f) Riprendiamo la tabella della distribuzione marginale della variabile "Salario".

Salario	Frequenze relative
$1 \vdash 2$	0,19
$2 \vdash 3{,}5$	0,39
$3{,}5 \vdash 5{,}5$	0,36
$5{,}5 \vdash 8$	0,07

Se coloro che cadono nella classe $1 \vdash 2$ mila di euro percepiscono un aumento di 200 euro, la classe diventa $1,2 \vdash 2,2$ mila euro, con valore centrale 1,7

mila euro. Pertanto, lo stipendio medio in seguito alla decisione aziendale è

$$M(\text{Salario}) = 1,7 \cdot 0,19 + 2,75 \cdot 0,39 + 4,5 \cdot 0,36 +$$

$$+6,75 \cdot 0,07 = 3,49 \text{ migliaia di euro}$$

Per calcolare la varianza, determiniamo il momento secondo

$$M(\text{Salario}^2) = 1,7^2 \cdot 0,19 + 2,75^2 \cdot 0,39 + 4,5^2 \cdot 0,36 +$$

$$+6,75^2 \cdot 0,07 = 13,98$$

La varianza è

$$S^2(\text{Salario}) = 13,98 - 3,49^2 = 1,8$$