PRIMA PROVA INTERMEDIA DI STATISTICA
CLEA, CLEFIN (COD. 377/371/5047/4038)
3 Novembre 2004

Cognome ***Nome***

Numero di matricola

COMPITO C1

Ai fini della valutazione si terrà conto solo ed esclusivamente di quanto riportato negli appositi spazi. Al termine della prova, è OBBLIGATORIO consegnare il presente foglio ed il foglio di brutta (DI CUI NON SI TERRÀ CONTO AI FINI DELLA VALUTAZIONE).

APPROSSIMARE TUTTI I CALCOLI ALLA QUARTA CIFRA DECIMALE

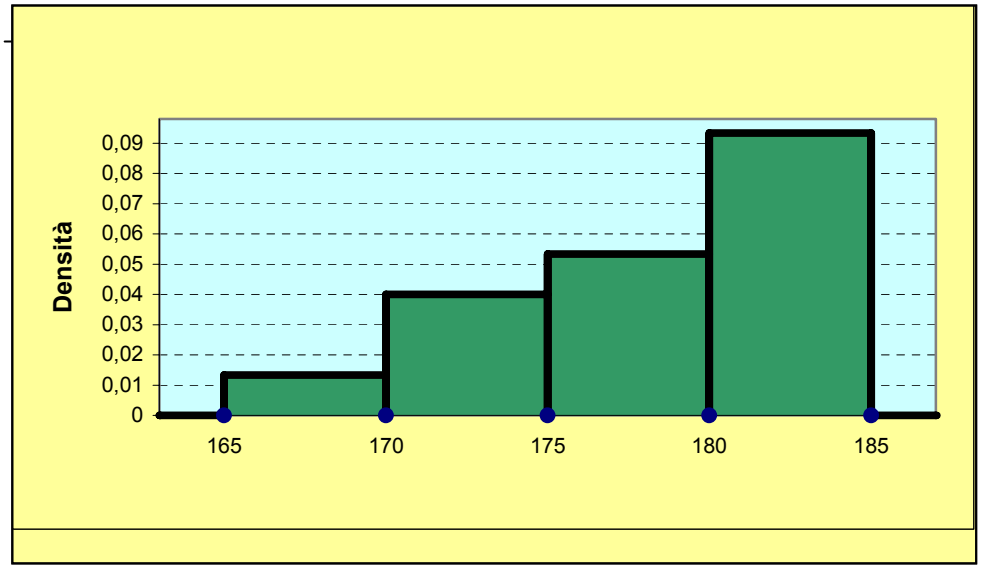
Ogni anno nel mese di Ottobre si disputa la gara "Ironman" delle Hawaii, una competizione di triathlon che consiste nel percorrere senza interruzione 3,8 km a nuoto, 180 km in bicicletta e 42,195 km a piedi. Per i primi 15 atleti nella classifica maschile dell'edizione 2003 vengono rilevati i seguenti caratteri

NUOTO	tempo impiegato nella frazione a nuoto, in minuti
BICI	tempo impiegato nella frazione in bicicletta, in minuti
CORSA	tempo impiegato nella frazione a piedi, in minuti
DISTACCO	minuti di distacco rispetto al tempo finale del vincitore
NAZIONE	paese di origine
FRAZIONE	miglior frazione percorsa (1=nuoto; 2=bici; 3=corsa)
ANNI	età al momento della gara

Di seguito sono riportati i dati, con alcuni calcoli utili.

Atleta	**Nuoto** (*N*)	**Bici** (*B*)	**Corsa** (*C*)	**Distacco** (*D*)	**Nazione** (*S*)	**Frazione** ***(F)***	**Anni** *(A)*	D^2	N^2	D×N
P. Reid	50.60	280.07	167.63	0	Canada	3	34	0	2560.36	0
R. Beke	52.47	277.98	174.20	6	Belgio	2	26	36	2753.10	314.82
C. Brown	50.63	279.95	175.57	8	NZelanda	1	31	64	2563.40	405.04
N. Stadler	52.73	273.67	182.83	10	Germania	2	30	100	2780.45	527.30
L. Bell	50.55	279.70	180.32	12	Australia	1	24	144	2555.30	606.60
J. Zack	51.70	278.82	181.03	13	Germania	2	38	169	2672.89	672.10
F. Al-Sultan	48.95	282.02	180.48	13	Germania	1	25	169	2396.10	636.35
C. Widoff	50.65	279.72	181.67	13	USA	1	34	169	2565.42	658.45
M. Lovato	52.55	284.07	176.22	14	USA	3	30	196	2761.50	735.70
M. Luoto	51.73	289.58	173.07	15	Finlandia	3	30	225	2676.00	775.95
T. Hellriegel	50.78	278.88	182.97	15	Germania	2	32	225	2578.61	761.70
J. Shortis	52.90	286.60	175.43	17	Australia	3	33	289	2798.41	899.30
C. Lieto	50.72	278.55	184.00	17	USA	2	31	289	2572.52	862.24
X. Le Floch	50.68	290.20	172.78	19	Francia	3	29	361	2568.46	962.92
L. Leder	52.23	288.88	176.05	19	Germania	3	32	361	2728.00	992.37
TOTALI	**769.87**	**4228.69**	**2664.25**	**191**			**459**	**2797**	**39530.52**	**9810.84**

1. (3 punti) Si ricavi la distribuzione di frequenza per il carattere CORSA riclassificato negli intervalli [165,170), [170,175), [175,180), [180,185) e se ne disegni l'istogramma. Se ne calcolino inoltre media e mediana, operando sui dati riclassificati.



C	n_i	p_i	c_i	$F(x)$
[165,170)	1	0.0667	0.0133	0.0667
[170,175)	3	0.2	0.04	0.2667
[175,180)	4	0.2667	0.0533	0.5334
[180,185)	7	0.4667	0.0933	1

M(C) = (167.5×0.0667)+(172.5×0.2)+(177.5×0.2667)+(182.5×0.4667) = 178.1842
La classe mediana è [175,180).
Me(C) = x si ricava risolvendo
0.0667+0.2+(x-175)0.0533 = 0.5 ==> x = 179.3771

2. (3 punti) Si calcolino il primo ed il terzo quartile per il carattere CORSA, riclassificato come nell'esercizio 1. Ci sono degli outliers nei dati? Si giustifichi la risposta.

$Q_1 = x$ e $Q_3 = y$, dove x e y si ricavano risolvendo le equazioni

0.0667+0.04(x-170) = 0.25
0.0667+0.2+0.2667+0.0933(y -180) = 0.75
==> x = 174.5825, y = 182.3215

Q_3-Q_1 = 9.8827

Il minimo e il massimo dei valori osservati sul carattere CORSA sono distanti rispettivamente da Q_1 e Q_3 meno di una volta e mezza il range interquartile (ossia 14.824) e pertanto non ci sono outliers.

3. (2 punti) Si considerino i due campioni seguenti
a) 1, 2, 30, 1, 1
b) 6, 7, 6, 8, 8.
La media è una buona misura di sintesi per entrambi? E' preferibile riassumere i dati con la mediana? Si giustifichino le risposte.

I due campioni hanno la stessa media, pari a 7, ma sono molto diversi tra di loro. Per il campione (a) la media non è un buon indicatore di posizione, poiché è molto sensibile alla presenza di un solo valore estremo, 30. In questo caso è opportuno utilizzare la mediana, che non è influenzata da osservazioni estreme. Nel caso (b) invece i dati sono omogenei e quindi la media è un buon indicatore. Inoltre, non c'è differenza tra media e mediana, in quanto sono entrambe pari a 7.

4. (2 punti) Si calcoli lo scarto quadratico medio per il carattere DISTACCO e si determinino gli estremi di un intervallo contenente almeno il 75% delle osservazioni.

$Var(D) = M(D^2) - M^2(D) = 186.4667 - 162.1369 = 24.3298$
$SQM(D) = 4.9325$

Uno dei possibili intervalli che contengono almeno il 75% delle osservazioni si può determinare con la disuguaglianza di Chebyshev ed è dato da $M(D) \pm 2SQM(D) = 12.7333 \pm 9.865 = (2.8683, 22.5983)$. Poiché è nota la distribuzione di frequenza del carattere DISTACCO, qualunque intervallo che contenga almeno il 75% delle osservazioni è accettabile, ad esempio l'intervallo (x_{MIN}, x_{MAX}).

5. (4 punti) Si consideri il carattere NAZIONE ricodificato nel modo seguente: il carattere assume valore 1 per atleti europei e 0 per gli altri.
a) (2 punti) Si costruisca e si riporti la tabella a doppia entrata per NAZIONE (ricodificato) e FRAZIONE e quindi si determini $Fr\{FRAZIONE \geq 2; NAZIONE = 1\}$

N \ F	1	2	3	
0	3	1	3	7
1	1	4	3	8
	4	5	6	

N \ F	1	2	3	
0	0.2	0.0667	0.2	0.4667
1	0.0667	0.2667	0.2	0.5334
	0.2667	0.3334	0.4	

$Fr\{FRAZIONE \geq 2; NAZIONE = 1\} = 0.2667 + 0.2 = 0.4667$

b) (1 punto) Si calcoli la moda per la distribuzione marginale di FRAZIONE.

F	p_i
1	0.2667
2	0.3334
3	0.4

$Mo(F) = 3$

c) (1 punto) Gli atleti che hanno come miglior frazione quella a piedi (F=3) sono più numerosi proporzionalmente tra gli europei o tra i non europei? Si giustifichi la risposta.

P(F=3|N=0) = 0.2/0.4667 = 0.4285
P(F=3|N=1) = 0.2/0.5334 = 0.3749
La frequenza di atleti con F=3 è maggiore nella sottopopolazione degli atleti non europei.

6. (2 punti) Si valuti l'intensità del legame lineare tra DISTACCO e NUOTO, utilizzando un opportuno indice. Si può concludere che i due caratteri *non* sono statisticamente indipendenti? Si giustifichi la risposta.

$M(D) = 12.7333$
$M(N) = 51.3247$
$Var(D) = 24.3298$
$Var(N) = M(N^2) - M^2(N) = 2635.368 - 2634.2214 = 1.1466$
$Cov(D,N) = M(D \times N) - M(D)M(N) = 654.056 - 653.5328 = 0.5232$

$\rho(D,N) = Cov(D,N) / SQM(D)\, SQM(N) = 0.099$

Poichè il coefficiente di correlazione lineare è diverso da zero non c'è indipendenza correlativa, e quindi non c'è nemmeno indipendenza statistica.

7. (2 punti) Si descriva il metodo dei minimi quadrati e si scriva l'espressione analitica dei parametri dell'interpolante lineare ottenuta con tale metodo.

Se si osservano coppie di valori (x_i, y_i) , $i=1,...,n$ relative a due variabili X e Y, l'interpolante lineare di Y ottenuta con il metodo dei minimi quadrati è la retta che minimizza, rispetto ad a e b, la quantità

$$\sum_{i=1}^{n} [y_i - (a + b x_i)]^2$$

I parametri dell'interpolante lineare risultano:

$b = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$

$a = M(Y) - b\, M(X)$

8. (3 punti) Siano A, B e C tre eventi definiti sullo stesso spazio campionario Ω: si sa che P(A)=2/5, P(B|A)=1/3, P(B|Ā)=4/5.
a) (1 punto) Si calcoli P(A|B).

Per il teorema delle probabilità totali
$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = (1/3 \times 2/5) + (4/5 \times 3/5) = 0.6133$

Per il teorema di Bayes
$P(A|B) = P(B|A)P(A)/P(B) = 0.1333/0.6133 = 0.2173$

b) (2 punti) Sapendo che $P(A\cap B)=2P(B\cap C)$, si calcoli $P(C|B)$.

$P(A\cap B) = P(B|A)P(A) = 0.1333$

$P(C|B) = P(B\cap C)/P(B) = P(A\cap B)/2P(B) = 0.1333/1.2266 = 0.1087$

9. (3 punti) Un esame scritto è composto da 2 esercizi, (a) e (b). L'esame è superato solo se si risolvono correttamente entrambi gli esercizi. Ipotizzando che la probabilità di risolvere correttamente un singolo esercizio sia pari a 1/2:
a) (1 punto) Qual è la probabilità di superare l'esame?

Siano A_1 = esercizio (a) svolto correttamente, A_0 = esercizio (a) sbagliato, e similmente per B_1 e B_0. Gli esiti dell'esame possibili sono $\{(A_1, B_1),(A_1, B_0),(A_0, B_1),(A_0, B_0)\}$.
Sia E l'evento "l'esame è superato". Allora $E=\{(A_1, B_1)\}$ e
$P(E) = 1/4 = 0.25$

b) (2 punti) 3 studenti devono sostenere l'esame. Si calcoli la probabilità che uno di loro non superi l'esame.

Sia X = numero di studenti che superano l'esame. Allora X è una variabile binomiale di parametri 3 e 0.25, e la probabilità che uno degli studenti non superi l'esame è
$P(X=2) = 3\times 0.25^2 \times 0.75 = 0.1406$

PRIMA PROVA INTERMEDIA DI STATISTICA
CLEA, CLEFIN (COD. 377/371/5047/4038)
3 Novembre 2004

Cognome ***Nome***

Numero di matricola

COMPITO C2

Ai fini della valutazione si terrà conto solo ed esclusivamente di quanto riportato negli appositi spazi. Al termine della prova, è OBBLIGATORIO consegnare il presente foglio ed il foglio di brutta (DI CUI NON SI TERRÀ CONTO AI FINI DELLA VALUTAZIONE).

APPROSSIMARE TUTTI I CALCOLI ALLA QUARTA CIFRA DECIMALE

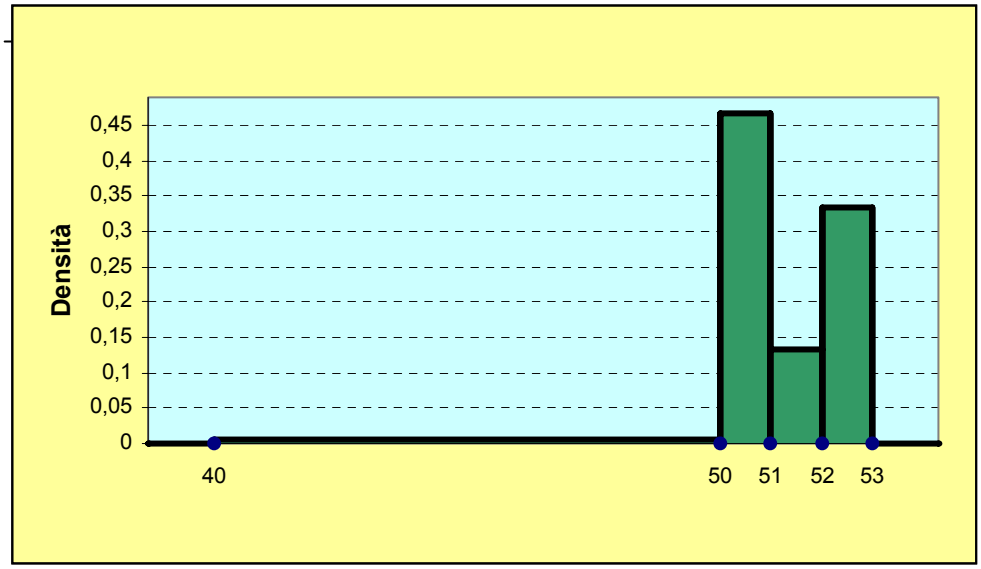
Ogni anno nel mese di Ottobre si disputa la gara "Ironman" delle Hawaii, una competizione di triathlon che consiste nel percorrere senza interruzione 3,8 km a nuoto, 180 km in bicicletta e 42,195 km a piedi. Per i primi 15 atleti nella classifica maschile dell'edizione 2003 vengono rilevati i seguenti caratteri

NUOTO	tempo impiegato nella frazione a nuoto, in minuti
BICI	tempo impiegato nella frazione in bicicletta, in minuti
CORSA	tempo impiegato nella frazione a piedi, in minuti
DISTACCO	minuti di distacco rispetto al tempo finale del vincitore
NAZIONE	paese di origine
FRAZIONE	miglior frazione percorsa (1=nuoto; 2=bici; 3=corsa)
ANNI	età al momento della gara

Di seguito sono riportati i dati, con alcuni calcoli utili.

Atleta	**Nuoto** (*N*)	**Bici** (*B*)	**Corsa** (*C*)	**Distacco** (*D*)	**Nazione** (*S*)	**Frazione** ***(F)***	**Anni** *(A)*	D^2	A^2	B^2	$D \times B$
P. Reid	50.60	280.07	167.63	0	Canada	3	34	0	1156	78439.20	0
R. Beke	52.47	277.98	174.20	6	Belgio	2	26	36	676	77272.88	1667.88
C. Brown	50.63	279.95	175.57	8	NZelanda	1	31	64	961	78372	2239.6
N. Stadler	52.73	273.67	182.83	10	Germania	2	30	100	900	74895.27	2736.7
L. Bell	50.55	279.70	180.32	12	Australia	1	24	144	576	78232.09	3356.4
J. Zack	51.70	278.82	181.03	13	Germania	2	38	169	1444	77740.59	3624.66
F. Al-Sultan	48.95	282.02	180.48	13	Germania	1	25	169	625	79535.28	3666.26
C. Widoff	50.65	279.72	181.67	13	USA	1	34	169	1156	78243.28	3636.36
M. Lovato	52.55	284.07	176.22	14	USA	3	30	196	900	80695.77	3976.98
M. Luoto	51.73	289.58	173.07	15	Finlandia	3	30	225	900	83856.58	4343.7
T. Hellriegel	50.78	278.88	182.97	15	Germania	2	32	225	1024	77774.05	4183.2
J. Shortis	52.90	286.60	175.43	17	Australia	3	33	289	1089	82139.56	4872.2
C. Lieto	50.72	278.55	184.00	17	USA	2	31	289	961	77590.10	4735.35
X. Le Floch	50.68	290.20	172.78	19	Francia	3	29	361	841	84216.04	5513.8
L. Leder	52.23	288.88	176.05	19	Germania	3	32	361	1024	83451.65	5488.72
TOTALI	**769.87**	**4228.69**	**2664.25**	**191**			**459**	**2797**	**14233**	**1192454,34**	**54041.81**

1. (3 punti) Si ricavi la distribuzione di frequenza per il carattere NUOTO riclassificato negli intervalli [40,50), [50,51), [51,52), [52,53) e se ne disegni l'istogramma. Se ne calcolino inoltre media e mediana, operando sui dati riclassificati.



N	n_i	p_i	c_i	$F(x)$
[40,50)	1	0.0667	0.0067	0.0667
[50,51)	7	0.4667	0.4667	0.5334
[51,52)	2	0.1333	0.1333	0.6667
[52,53)	5	0.3333	0.3333	1

M(N) = (45×0.0667)+(50.5×0.4667)+(51.5×0.1333)+(52.5×0.3333) = 50.933
La classe mediana è [50,51).
Me(N) = x si ricava risolvendo
0.0667+(x-50)0.4667 = 0.5 ==> x = 50.9284

2. (3 punti) Si calcolino il primo ed il terzo quartile per il carattere NUOTO, riclassificato come nell'esercizio 1. Ci sono degli outliers nei dati? Si giustifichi la risposta.

$Q_1 = x$ e $Q_3 = y$, dove x e y si ricavano risolvendo le equazioni

0.0667+0.4667(x-50) = 0.25
0.0667+0.4667+0.1333+0.3333(x-52) = 0.75
==> x = 50.3928, y = 52.25
Q_3-Q_1 = 1.8572

Il minimo e il massimo dei valori osservati sul carattere CORSA sono distanti rispettivamente da Q_1 e Q_3 meno di una volta e mezza il range interquartile (ossia 2.7858) e pertanto non ci sono outliers.

3. (2 punti) Si considerino i due campioni seguenti
a) 11, 13, 11, 10, 10
b) 1, 2, 1, 50, 1.
La media è una buona misura di sintesi per entrambi? E' preferibile riassumere i dati con la mediana? Si giustifichino le risposte.

I due campioni hanno la stessa media, pari a 11, ma sono molto diversi tra di loro. Per il campione (b) la media non è un buon indicatore di posizione, poiché è molto sensibile alla presenza di un solo valore estremo, 50. In questo caso è opportuno utilizzare la mediana, che non è influenzata da osservazioni estreme. Nel caso (a) invece i dati sono omogenei e quindi la media è un buon indicatore. Inoltre, non c'è differenza tra media e mediana, in quanto sono entrambe pari a 11.

4. (2 punti) Si calcoli lo scarto quadratico medio per il carattere ANNI e si determinino gli estremi di un intervallo contenente almeno il 75% delle osservazioni.

$Var(A) = M(A^2) - M^2(A) = 948.8667 - 936.36 = 12.5067$
$SQM(A) = 3.5365$

Uno dei possibili intervalli che contengono almeno il 75% delle osservazioni si può determinare con la disuguaglianza di Chebyshev ed è dato da $M(A) \pm 2SQM(A) = 30.6 \pm 7.073 =$ (23.527,37.673). Poiché è nota la distribuzione di frequenza del carattereANNI, qualunque intervallo che contenga almeno il 75% delle osservazioni è accettabile, ad esempio l'intervallo (x_{MIN}, x_{MAX}).

5. (4 punti) Si consideri il carattere ANNI ricodificato nel modo seguente: il carattere assume valore 1 per atleti con meno di 30 anni e 0 per gli altri.
a) (2 punti) Si costruisca e si riporti la tabella a doppia entrata per ANNI (ricodificato) e FRAZIONE e quindi si determini $Fr\{FRAZIONE \geq 2; ANNI = 1\}$.

A \ F	1	2	3	
0	2	4	5	11
1	2	1	1	4
	4	5	6	

A \ F	1	2	3	
0	0.1333	0.2667	0.3333	0.7333
1	0.1333	0.0667	0.0667	0.2667
	0.2666	0.3334	0.4	

$Fr\{FRAZIONE \geq 2; ANNI = 1\} = 0.0667 + 0.0667 = 0.1334$

b) (1 punto) Si calcoli la moda per la distribuzione di FRAZIONE subordinata ad ANNI=1.

F	$p_{i\,\|A=1}$
1	0.1333/0.2667 = 0.4998
2	0.0667/0.2667 = 0.25
3	0.0667/0.2667 = 0.25

$Mo(F) = 1$

c) (1 punto) Gli atleti che hanno come miglior frazione quella a piedi (F =3) sono più numerosi proporzionalmente tra gli atleti con meno di 30 anni o tra gli altri? Si giustifichi la risposta.

P(F=3|A=0) = 0.3333/0.7333 = 0.4545
P(F=3|A=1) = 0.0667/0.2667 = 0.25
La frequenza di atleti con F=3 è maggiore nella sottopopolazione degli atleti con almeno 30 anni.

6. (2 punti) Data *Var*(D) = 24.3298, si verifichi l'intensità del legame lineare tra DISTACCO e BICI, utilizzando un opportuno indice. Si può concludere che i due caratteri *non* sono statisticamente indipendenti? Si giustifichi la risposta.

$M(D) = 12.7333$
$M(B) = 281.9127$
$Var(B) = M(B^2) - M^2(B) = 79496.956 - 79474.7704 = 22.1856$
$Cov(D,B) = M(D \times B) - M(D)M(B) = 3602.7873 - 3589.678 = 13.1093$

$\rho(D,B) = Cov(D,B) / SQM(D)\, SQM(B) = 0.5642$

Poiché il coefficiente di correlazione lineare è diverso da zero non c'è indipendenza correlativa, e quindi non c'è nemmeno indipendenza statistica.

7. (2 punti) Si descriva il metodo dei minimi quadrati e si scriva l'espressione analitica dei parametri dell'interpolante lineare ottenuta con tale metodo.

Se si osservano coppie di valori (x_i, y_i) , $i=1,...,n$ relative a due variabili X e Y, l'interpolante lineare di Y ottenuta con il metodo dei minimi quadrati è la retta che minimizza, rispetto ad a e b, la quantità

$$\sum_{i=1}^{n} [y_i - (a + b x_i)]^2$$

I parametri dell'interpolante lineare risultano:

$b = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$

$a = M(Y) - b\, M(X)$

8. (3 punti) Siano A, B e C tre eventi definiti sullo stesso spazio campionario Ω: si sa che P(A)=3/10, P(B|A)=1/6, P(B|Ā)=7/10.

a) (1 punto) Si calcoli P(A|B).

Per il teorema delle probabilità totali
P(B) = P(B|A)P(A)+P(B|Ā)P(Ā) = (1/6×3/10)+(7/10×7/10) = 0.54

Per il teorema di Bayes
P(A|B) = P(B|A)P(A)/P(B) = 0.05/0.54 = 0.0926

b) (2 punti) Sapendo che $P(A \cap B)=1/2P(B \cap C)$, si calcoli $P(C|B)$.

$P(A \cap B) = P(B|A)P(A) = 0.05$

$P(C|B) = P(B \cap C)/P(B) = 2P(A \cap B)/P(B) = 0.1/0.54 = 0.1852$

9. (3 punti) Un esame scritto è composto da 2 esercizi, (a) e (b). L'esame è superato se se ne risolve correttamente almeno uno. Ipotizzando che la probabilità di risolvere correttamente un singolo esercizio sia pari a 1/2:

a) (1 punto) Qual è la probabilità di superare l'esame?

L'esame è superato nei seguenti casi: si risolve correttamente solo l'esercizio (a), si risolve correttamente solo l'esercizio (b), si risolvono correttamente entrambi gli esercizi.
Siano A_1 = esercizio (a) svolto correttamente, A_0 = esercizio (a) sbagliato, e similmente per B_1 e B_0. Gli esiti dell'esame possibili sono $\{(A_1, B_1),(A_1, B_0),(A_0, B_1),(A_0, B_0)\}$.
Sia E l'evento "l'esame è superato". Allora $E=\{(A_1, B_1),(A_1, B_0),(A_0, B_1)\}$ e
$P(E) = 1/4+1/4+1/4 = 0.75$

b) (2 punti) 5 studenti devono sostenere l'esame. Si calcoli la probabilità che almeno due studenti superino l'esame.

Sia X = numero di studenti che superano l'esame. Allora X è una variabile binomiale di parametri 5 e 0.75, e la probabilità che siano almeno in due a superare l'esame è
$1-P(X=0)-P(X=1) = 1-(0.25^5 + 5 \times 0.75 \times 0.25^4) = 1-0.0156 = 0.9844$