

- Prova in itinere del 19 Dicembre 2012 - Fila A

Esercizio 1)

Nella tabella seguente viene riportata la statistica relativa al numero di esami sostenuti con successo dalle matricole di Ingegneria industriale iscritte all'ateneo mantovano nel A.A. 2010/2011 alla fine della prima sessione di esami. Per esempio 6 matricole non hanno superato nessun esame nell'A.A. 2010/2011.

Esami superati	0	1	2	3	4	5	6
Frequenza Assoluta	6	6	11	15	9	7	6

Determinare

- La tipologia del carattere.
- Un indice sintetico di posizione.
- Se possibile, un indice sintetico di variabilità.
- Una rappresentazione grafica adeguata.
- L'eventuale presenza di outlier.

Esercizio 2)

Il dato riportato al punto precedente è stato ulteriormente caratterizzato nella seguente tabella a doppia entrata (da completare vedi punto b) aggiungendo il genere.

		Y: stato civile		Totale
		Maschi	Femmine	
X: esami superati	meno di 3	18		
	3 o 4	14		
	Piu di 4			
Totale			25	

Il candidato

- Indichi il tipo di carattere della serie bivariata.
- Completi la tabella con i dati mancanti.
- Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di posizione.
- Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di variabilità.
- Verifichi, ad un opportuno livello di significatività, se gli uomini dimostrano una maggiore predisposizione allo studio rispetto alle donne.

Esercizio 3)

Utilizzando i dati relativi alla facoltà di ingegneria industriale descritti all'esercizio 1 come campione dell'intera università mantovana, stimare per intervallo il valore atteso del numero di esami sostenuti con successo da una matricola alla fine della prima sessione di esami. Il candidato indichi e verifichi le ipotesi richieste per effettuare la stima, procedendo al calcolo anche qualora queste non siano soddisfatte.

Esercizio 4)

Si considerino i seguenti eventi dichiarati incompatibili.

E_1 : estraendo un'unità a caso dalla statistica descritta all'esercizio 2, questa sia di genere femminile
 E_2 : si ottenga $x \leq 5$ dove x è estratto da una v.c. distribuita come $N(6,1)$

- Il candidato calcoli le seguenti Probabilità $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$ $P(E_1 | E_2)$.
- Il candidato indichi se i due eventi E_1 ed E_2 sono indipendenti.

- Prova in itinere del 19 Dicembre 2012 -
Svolgimento - Fila A

Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo quantitativo (in quanto espresso da numeri) discreto (in quanto le modalità sono numeri naturali e concettualmente limitate al un numero massimo pari al numero di corsi elargiti al primo semestre)

b) Un indice sintetico di posizione.

Un carattere di tipo quantitativo ammette tre indici sintetici di posizione: la moda, la mediana e media. Un indice idoneo in questo caso è la mediana, in quanto risulta poco affetto dalla presenza di eventuali outlier.

Per calcolare la mediana si deve valutare la numerosità della popolazione ($N = 60$) facilmente ottenibile cumulando le frequenze assolute

Esami sostenuti (m_i)	0	1	2	3	4	5	6
Frequenze assolute (n_i)	6	6	11	15	9	7	6
Frequenze. ass. cumulate	6	12	23	38	47	54	60

Dopo di che, la mediana è il valore che bipartisce la popolazione, ovvero, una volta ordinate le osservazioni si ricerca quella che lascia alla sua destra $(N-1)/2 = 29,5$ elementi. Poiché non esiste l'osservazione di posto 30,5 viene preso come mediana la media fra il 30° ed 31° valore. Analizzando le frequenze cumulate si ottiene che ambo le osservazioni mostrano la modalità 3. Pertanto la mediana (q_2) è 3

c) Se possibile, un indice sintetico di variabilità.

Un carattere di tipo quantitativo ammette quattro indici sintetici di variabilità: il range (o campo di variazione) la distanza interquartile, la varianza e la deviazione standard (o scarto quadratico medio). Avendo illustrato la mediana come indice di posizione la scelta più logica per l'indice di variabilità connesso è quella di utilizzare la distanza interquartile che si basa sullo stesso concetto. Infatti essa rappresenta la differenza fra il primo (q_1) ed il terzo (q_3) quartile. Dato q_1 , una volta ordinate le osservazioni, lascia alla propria sinistra $(N-1)/4 = 14,75$ osservazioni mentre q_3 lascia alla propria destra $(N-1)/4 = 14,75$ osservazioni. Anche in questo caso non ottenendo numeri interi dovremo mediare le posizioni intere più vicine. Si ha dunque che

$$q_1 = \text{media } 15^\circ \text{ e } 16^\circ \text{ valore} = 2$$

$$q_3 = \text{media } 45^\circ \text{ e } 46^\circ \text{ valore} = 4$$

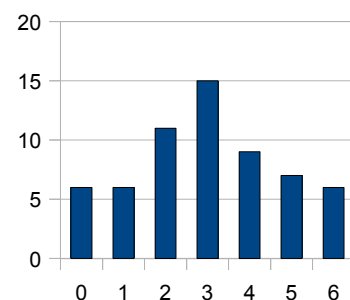
Si ha che la distanza interquartile $D = q_3 - q_1 = 4 - 2 = 2$

d) Una rappresentazione grafica adeguata.

Un carattere di tipo quantitativo le cui le modalità abbiano frequenze superiori all'unità viene solitamente rappresentato mediante un diagramma a barre.

Questo diagramma è composto da barre orizzontali (o verticali) inserite in un piano cartesiano. Il grafico riposta una barra per ogni modalità, la cui base (o altezza) viene fissata e centrata nel valore della modalità corrispondente mentre la sua altezza (o base) raggiunge la relativa frequenza assoluta.

A lato si riporta il diagramma a barre ricavato dalla distribuzione in esame



e) L'eventuale presenza di outlier.

Un modo per individuare gli outlier (ovvero valori troppo distanti dalla statistica e probabilmente erronei) è quello di ricorrere alla definizione di Valore Adiacente Superiore e di Valore Adiacente Inferiore, per individuare i valori rispettivamente troppo alti o troppo bassi. Questi limiti vengono calcolati sottraendo al primo quartile K volte la distanza interquartile (VAI) e sommando al terzo quartile K volte la distanza interquartile (VAS). I valori esterni all'intervallo VAI-VAS vengono considerati outlier. Tipici valori di K sono 1, 1.5 e 2. Utilizzando $K = 1$ si ha che

$$VAI = 2 - 1 \cdot 2 = 0$$

$$VAS = 4 + 1 \cdot 2 = 6$$

Non esistendo alcuna osservazioni esterna all'intervallo [0 ; 6] possiamo concludere che la popolazione presumibilmente non presenta outlier.

Esercizio 2)

a) *Indichi il tipo di carattere della serie bivariata.*

L'esercizio verte sull'analisi di una serie bivariata, ottenuta combinando un carattere quantitativo discreto con un carattere qualitativo non ordinabile. Poiché i due caratteri componenti la serie sono diversi si ha che la serie non ha un carattere definito.

b) *Completate la tabella con i dati mancanti.*

Per completare la tabella di contingenza basta ricordarsi che le marginali del carattere X sono ottenibili dalla statistica riportata nell'Esercizio 1.

		Y: genere		Totale
		Maschi	Femmine	
X: Esami superati	meno di 3	18	5	23
	3 o 4	14	10	24
	Piu di 4	3	10	13
Totale		35	25	60

c) *Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di posizione*

Una serie bivariata ottenuta misurando due caratteri qualitativi non ordinabili ammette un solo indice sintetico di posizione: la moda. La moda di una bi-variata si ottiene valutando la modalità della serie corrispondente alla frequenza (assoluta o relativa) maggiore. Nel caso in esame la frequenza assoluta maggiore è 18 a cui corrisponde la seguente moda

(meno di 3 ; Maschio)

d) *Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di variabilità*

Una serie bivariata con un carattere qualitativo non ammette indice sintetico di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

e) *Verifichi, ad un opportuno livello di significatività, se gli uomini dimostrano una maggiore predisposizione allo studio rispetto alle donne.*

Se l'asserzione fosse vera i due caratteri che compongono la bivariata in oggetto dovrebbero essere dipendenti (infatti l'appartenza ad un genere influenza il numero di esami superati con successo). Per provare questa dipendenza si può effettuare un test di ipotesi volto a verificare se le frequenze delle osservazioni rilevate nel campione sono sufficientemente vicine (ad un determinato livello di significatività) a quelle teoriche ottenute dall'ipotesi di indipendenza e confidare che esso dia esito negativo. Il test in oggetto viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadrato avente gradi di libertà pari al numero di parametri liberi della distribuzione teorica.

Il primo punto di questa procedura consiste nel calcolo delle frequenze teoriche ricavate dalle frequenze marginali ottenute orlando la tabella delle frequenze .

$$\hat{n}_{i,j} = n \hat{p}_{i,j} = \frac{n_{i,+} n_{+,j}}{n} \quad \forall i, j$$

nella tabella si riportano le frequenze osservate e quelle teoriche fra parentesi

		Y: genere		Totale
		Maschi	Femmine	
X: Esami superati	meno di 3	18 (13.42)	5 (9.58)	23
	3 o 4	14 (14)	10 (10)	24
	Piu di 4	3 (7.58)	10 (5.42)	13
Totale		35	25	60

A questo punto è possibile valutare la convergenza dello stimatore di Pizzetti Pearson, possibile solo se tutte le

frequenze teoriche sono superiori a 5. Constatato che la condizione è verificata si può procedere al calcolo della regione di accettazione fissato il livello di significatività al 5%.

$$A = [0; \chi^2_{1-\alpha}((M_x - 1)(M_y - 1))] = [0; \chi^2_{1-0.05}((3-1)(2-1))] = [0; \chi^2_{0.95}(2)] = [0; 5.991]$$

Si può ora procedere al calcolo dello stimatore vero e proprio

$$\frac{\sum_{i=1}^3 \sum_{j=1}^2 (n_{i,j} - \hat{n}_{i,j})^2}{\hat{n}_{i,j}} = \frac{(18-13.42)^2}{13.42} + \frac{(5-9.58)^2}{9.58} + \frac{(14-14)^2}{14} + \frac{(10-10)^2}{10} + \frac{(3-7.58)^2}{7.58} + \frac{(10-5.42)^2}{5.42} = 10.41$$

Poichè il valore dello stimatore è esterno all'intervallo di accettazione posso dire che i due caratteri non sono indipendenti ad un livello di significatività del 5 %.

Il test fornisce una evidenza sperimentale di una dipendenza fra i due caratteri. Confrontando i valori assoluti possiamo vedere che al genere femminile corrisponde la maggior frequenza relativa di esami passati (10/25 contro i 3/35 del "sesso forte") pertanto si può concludere che l'ipotesi formulata nella consegna sia smentita dai dati ad un livello di significatività del 5%.

Esercizio 3)

Nel testo viene richiesto di stimare il valore atteso della variabile casuale.

X: numero di esami superati da una matricola dell'università di Mantova nel primo semestre

Questa stima si appoggia allo stimatore media campionaria e si basa su due ipotesi

1. richiede un campione la cui dimensione sia di almeno 30 elementi (soddisfatta)
2. che ogni estrazione sia i.i.d.. La variabile riguarda tutta l'università, mentre il campione riguarda la sola facoltà di ingegneria. Per poter considerare le prove identicamente distribuite si dovrebbe poter dire che il bacino degli studenti di ingegneria sia estremamente simile a quello di tutti gli studenti. Ipotesi difficilmente sostenibile.

Il testo prevede di fare la stima in ogni caso, per tanto si procede nei calcoli. Come primo punto risulta necessario definire un livello di confidenza, ad esempio il 98% ($\alpha = 2\%$) da cui si ha che in ogni coda resta 1%. A questo punto si ricava che la stima è data da

$$E[X] \in \left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} ; \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]$$

Dopo aver effettuato i conti in tabella riusciamo a ricavare

$$\bar{x} = \sum_{i=1}^6 m_i * f_i = 3 \quad \sigma^2 = (\sum_{i=1}^6 m_i^2 * f_i) - \bar{x}^2 = 12 - 9 = 3 \quad s = \sqrt{\frac{\sigma^2 * n}{n-1}} = \sqrt{\frac{3 * 60}{59}} = 1.47$$

$$z_{1-\frac{0.02}{2}} = z_{0.99} = 2.33$$

da cui si ha l'intervallo richiesto

$$E[X] \in \left[3 - z_{1-\frac{0.02}{2}} \frac{1.47}{\sqrt{60}} ; 3 + z_{1-\frac{0.02}{2}} \frac{1.47}{\sqrt{60}} \right] = \left[3 - 2.33 \frac{1.47}{\sqrt{60}} ; 3 + 2.33 \frac{1.47}{\sqrt{60}} \right] = [2.56 ; 3.44]$$

Esami sostenuti (m_i)	0	1	2	3	4	5	6	Totali
Frequenze relative (f_i)	1/10	1/10	11/60	1/4	3/20	7/60	1/10	
$m_i * f_i$	0	1/10	11/30	3/4	3/5	7/12	3/5	3
$m_i^2 * f_i$	0	1/10	11/15	9/4	12/5	35/12	18/5	12

Esercizio 4)

a) Il candidato calcoli le seguenti Probabilità: $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$ $P(E_1 | E_2)$.

Per determinare $P(E_i)$ possiamo considerare come insieme universo 60 unità statistiche che compongono il campione descritto nell'esercizio 1. Poiché esse sono equiprobabili, possiamo utilizzare la definizione classica, ricordando che in questo insieme universo l'evento E_i è dato dalle 25 femmine si ha che $P(E_i)$ essere ricavate utilizzando la definizione classica; secondo la quale la probabilità è il rapporto dei casi favorevoli sui casi totali.

Pertanto si ha che: $P(E_1) = 25/60 = 41.67\%$

Per determinare $P(E_2)$ dobbiamo ricondurci alla variabile standardizzata. Ricordando la trasformazione di standardizzazione $Z = \frac{X - E[X]}{\sqrt{Var[X]}}$ si ha che

$$z = \frac{5-6}{\sqrt{1}} = -1 \quad \text{quindi} \quad P(E_2) = P(x < 5) = P(z < -1) = P(z < 1) = 0.1587$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica ricordando che nel caso di eventi incompatibili la probabilità dell'evento intersezione è nulla.

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.4167 + 0.1587 - 0 = 0.5754 \quad P(E_1 | E_2) = P \frac{(E_1 \cap E_2)}{P(E_2)} = \frac{0}{0.1587} = 0$$

b) Il candidato indichi se i due eventi E_1 ed E_2 sono indipendenti.

Due eventi incompatibili sono per forza dipendenti. Infatti sapere che un vento si è verificato annulla la probabilità che si verifichi il secondo.

- Prova in itinere del 19 Dicembre 2012 - Fila B

Esercizio 1)

Nella tabella seguente viene riportato il tempo di analisi di esami ematologici riscontrato nell'ultimo anno (365 gg) in un laboratorio di analisi veronese. Per esempio 240 esami hanno richiesto un tempo compreso tra 0 e 12 ore.

Tempo impiegato	0-12 ore	12- 24 ore	24 - 48 ore	48 - 96 ore
Frequenza Assoluta	240	140	80	40

Determinare

- La tipologia del carattere.
- Tutti gli indici sintetici di posizione possibili da calcolare.
- Se possibile, un indice sintetico di variabilità.
- Una rappresentazione grafica adeguata.

Esercizio 2)

Da un'indagine si sono rilevati in 5 facoltà italiane di biotecnologie (indicate con lettere A-E) il numero di ore di tutorato e le percentuali di superamento dell'esame di matematica & statistica. I dati ottenuti sono rappresentati nella forma: *Facoltà*(ore; percentuale).

$A(0; 65) B(0; 55) C(6; 60) D(6; 50) E(8; 70)$

Il candidato,

- Indichi e fornisca una rappresentazione grafica adeguata alla serie ottenuta.
- Se possibile, indichi e calcoli un opportuno indice di variabilità
- Ipotizzando un legame di tipo lineare,
 - Calcoli l'opportuna regressione
 - Ipotizzi quale sarebbe la percentuale prevista nel caso si facessero 20 ore di tutorato.
 - Il legame ipotizzato è attendibile? Motivare numericamente la risposta.

Esercizio 3)

Si vuole verificare la bontà di un dado a 6 facce. Il candidato:

- determini il numero di osservazioni necessarie affinché si possa procedere a tale verifica
- supposto di aver eseguito 120 lanci, indicare se il dado è onesto basandosi sulle seguenti frequenze assolute

Esito	1	2	3	4	5	6
Frequenza	22	12	20	30	18	18

Esercizio 4)

Si considerino i seguenti eventi dichiarati indipendenti.

E_1 : estraendo una facoltà a caso dal campione descritto all'Esercizio 2, questa non abbia svolto tutorato

E_2 : si ottenga $x=1$ dove x è estratto da una v.c. distribuita come $Ber(0.2)$

- Il candidato calcoli le seguenti Probabilità $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$ $P(E_1 | E_2)$.
- Il candidato indichi se i due eventi E_1 ed E_2 sono incompatibili.

- Prova in itinere del 19 Dicembre 2012 -
Svolgimento Fila B

Esercizio 1)

a) Determini la tipologia del carattere.

Il carattere è di tipo quantitativo (in quanto espresso da numeri) continuo (in quanto si vuole monitorare un tempo che concettualmente è continuo).

b) Determinare tutti gli indici sintetici di posizione possibili da calcolare.

Per caratteri qualitativi continui questo tipo di caratteri è possibile calcolare tutti gli indici di posizione visti nel corso (media, moda e mediana). Sebbene la definizione degli indici sia data utilizzando le osservazioni (o le modalità nel caso siano presenti frequenze) esse sono state estese di modo da considerare anche le statistiche espresse mediante classi di modalità.

Media: somma delle osservazioni diviso il numero delle osservazioni. Nel caso di dati in classi si sostituisce al valore della modalità il valore di centro classe della classe di modalità interessata. Usando i calcoli presenti in Tabella 1 si ha:

$$media = \sum_{i=1}^M c_i * f_i = 19.44$$

Moda: modalità corrispondente alla frequenza maggiore. Nel caso di dati in classi si sostituisce al valore della modalità la classe di modalità interessata e si parla di classe modale.

$$moda = classe \text{ con frequenza maggiore } (240) = 0 - 12 \text{ ore}$$

Mediana: osservazione che bipartisce le osservazioni. Nel caso di dati espressi in modalità essa la mediana è il valore che bipartisce la popolazione, ovvero, una volta ordinate le osservazioni si ricerca quella che lascia alla sua destra (o sinistra) la metà delle osservazioni meno una. Nel caso in esame non vi sono le osservazioni, in quanto queste sono raccolte in classi, pertanto la mediana si indica come il valore che bipartisce l'area dell'istogramma (si ricorda che l'area dell'istogramma è sempre unitaria). Dal calcolo delle frequenze cumulative (F_i) si vede come la mediana cada nella 2 classe (prima classe a superare lo 0.5). Per determinare l'esatto valore basta imporre che l'area presente nella classe 2 sia sufficiente a raggiungere il valore di 0.5. Si ottiene quindi:

$$\begin{aligned} Area \text{ residua} &= 0.5 - 0.48 = 0.02 \\ base \text{ rettangolo} &= 0.02 / 0.02333 = 0.85 \\ mediana &= limite \text{ classe } (i-1) + base = 12.85 \end{aligned}$$

c) Determinare, se possibile, un indice sintetico di variabilità.

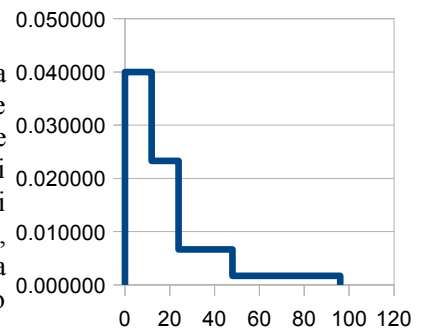
Un indice possibile è la varianza. Nel caso siano presenti osservazioni raggruppate in classi essa si calcola utilizzando come modalità i valori centrali delle classi (c_i). Nella Tabella 1 è stato riportato il calcolo della varianza utilizzando la formula abbreviata.

$$\sigma^2 = \left(\sum_{i=1}^M c_i^2 * f_i \right) - \left(\sum_{i=1}^M c_i * f_i \right)^2 = 730.08 - 19.44^2 = 352.17$$

Il risultato è stato ottenuto calcolando la media (somma colonna $c_i f_i$) e della media dei quadrati dei valori centrali (ultime due colonne della tabella).

d) Determinare una rappresentazione grafica adeguata.

L'istogramma è una rappresentazione comunemente utilizzata quando si tratta un dato quantitativo continuo che, per diverse esigenze, viene raccolto in classi di modalità c_i . Il grafico riporta le modalità sull'asse delle ascisse e sulle ordinate la densità di frequenza di ogni classe. Il grafico si compone di rettangoli fra di loro adiacenti la cui base si ricava dagli estremi della classe mentre l'altezza, l'altezza coincide con la densità di frequenza, pertanto l'area di ogni rettangolo sarà uguale alla frequenza relativa della classe. A lato si riporta l'istogramma richiesto. I conti per ricavare il suddetto istogramma sono riportati nella tabella in calce. (colonne f_i e $f_i / (sup_i - inf_i)$)



i	inf _i	sup _i	c _i	sup _i -inf _i	n _i	f _i	F _i	f _i /(sup _i -inf _i)	c _i f _i	c _i ²	c _i ² f _i
1	0	12	6	12	240	0.480	0.480	0.040000	2.880	36	17.28
2	12	24	18	12	140	0.280	0.760	0.023333	5.040	324	90.72
3	24	48	36	24	80	0.160	0.920	0.006667	5.760	1296	207.36
4	48	96	72	48	40	0.080	1.000	0.001667	5.760	5184	414.72

Totali 500 1 19.44 730.08

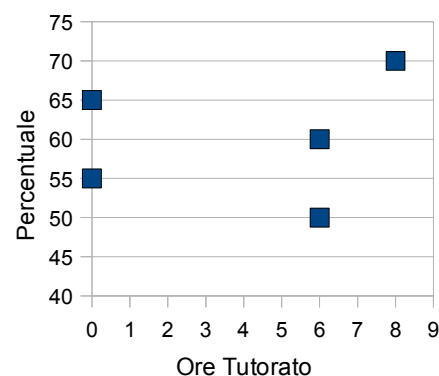
Tabella 1- Statistica Esercizio 1

Esercizio 2)

a) *Indicare e fornire una rappresentazione grafica adeguata.*

Per serie bivariate continue o discrete cui le frequenze non siano particolarmente alte si usa rappresentare la serie mediante diagrammi a dispersione. Un diagramma a dispersione è rappresentato in un piano cartesiano dove le modalità dei due caratteri vengono posti sui due assi ed ogni osservazione viene rappresentata da un punto.

A lato si mostra il diagramma a dispersione ottenuto dai dati forniti.



b) *Se possibile, indichi e calcoli un opportuno indice di variabilità*

Per serie bivariate continue o discrete l'indice di variabilità migliore è dato dalla matrice varianza/covarianza. Questa matrice si compone di 3 distinti valori: le due varianze dei distinti caratteri e la covarianza della serie bivariata.

Si seguito riportiamo i calcoli per le due varianze per i singoli caratteri:

X: Ore di tutorato

$$\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i = \frac{0+0+6+6+8}{5} = 4$$

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{0^2+0^2+6^2+6^2+8^2}{5} - 4^2 = \frac{136}{5} - 16 = 11.2$$

Y: Percentuale di superamento esame

$$\bar{y} = \frac{1}{N} \sum_{i=1}^n y_i = \frac{55+65+50+60+70}{5} = 60$$

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^n y_i^2 - \bar{y}^2 = \frac{55^2+65^2+50^2+60^2+70^2}{5} - 60^2 = \frac{18250}{5} - 3600 = 3650 - 3600 = 50$$

La covarianza si ottiene

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 4$$

I cui conti sono ripostati in tabella

	X	Y	x - $\bar{x}$	y - $\bar{y}$	(x - $\bar{x}$) (y - $\bar{y}$)
	0	65	-4	5	-20
	0	55	-4	-5	20
	6	60	2	0	0
	6	50	2	-10	-20
	8	70	4	10	40
somma	20	300			20
Media	4	60			4

Per tanto la matrice varianza covarianza risulta essere

$$\Sigma = \begin{bmatrix} 11.2 & 4 \\ 4 & 50 \end{bmatrix}$$

c 1) *Ipotizzando un legame di tipo lineare, si calcoli l'opportuna regressione*

La retta di regressione ha equazione

$$\hat{y} = \frac{\sigma_{xy}}{\sigma_x^2} x + \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2} \bar{x} \quad \hat{y} = \frac{4}{11.2} x + 60 - \frac{4}{11.2} 4 \quad \hat{y} = 0.357 x + 58.571$$

c 2) *Ipotizzando un legame di tipo lineare, si ipotizzi quale sarebbe la percentuale prevista nel caso si facessero 20 ore di tutorato.*

La risposta a questo quesito si ottiene applicando la retta nel punto $x = 20$. si ottiene quindi una percentuale di

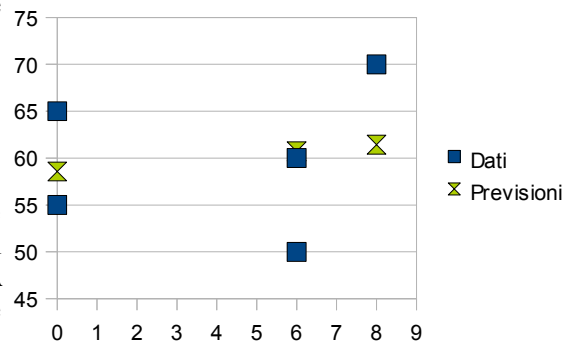
successo stimata pari al 65.71 %.

c 3) Ipotizzando un legame di tipo lineare, si verifichi il legame ipotizzato è attendibile? Motivare numericamente la risposta

Un buon indicatore della bontà del modello di regressione è dato dall'indice di correlazione di Pearson

$$R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{4^2}{11.2 \cdot 50} = 0.028 \quad R = 0.169$$

Poiché l'indice risulta inferiore a 0.3 si può asserire che il legame difficilmente sarà buono. Ovviamente il dato deve essere confermato dalla visualizzazione del modello. Infatti il coefficiente di Pearson può anche dare risultati molto errati. A lato si riportano le previsioni effettuate dal modello lineare che mal descrivono l'andamento dei dati



Esercizio 3)

L'indagine statistica mira a verificare mediante inferenza se un dado è onesto, ovvero se ogni esito è equiprobabile. Pertanto l'analisi si riduce a verificare che le probabilità degli esiti elementari (ovvero esca 1, esca 2, ..., esca 6) siano tutte pari ad 1/6. Tale verifica è possibile introducendo la mediante un test denominato di "Test sulla distribuzione (empirica¹)" in cui l'ipotesi nulla è quella che la frequenza relativa degli esiti sia pari a 1/6 per ogni esito.

a) determinare il numero di osservazioni necessarie affinché si possa procedere a tale verifica

Per il test in esame si deve ottenere che ogni frequenza assoluta teorica (ovvero quella data dall'ipotesi nulla) sia sempre non inferiore a 5. Ricordando che la frequenza assoluta teorica ($\hat{n}$) è data dalla probabilità teorica ($\hat{p}$) per il numero di prove (N). Si ricava il dato mediante i seguenti semplici passaggi

$$\hat{n} = N * \hat{p} \quad \text{imponendo} \quad \hat{n} \geq 5 \quad \text{si ricava} \quad 5 \geq \frac{N * 1}{6} \quad \text{da cui} \quad N \leq 30$$

Quindi il numero di prove necessarie deve essere almeno di 30.

b) supposto di aver eseguito 120 lanci, indicare se il dado è onesto

Il test impostato è un test di aderenza alla distribuzione empirica.

Prima di verificare il test è necessario controllare le ipotesi di validità del test

- Prove i.i.d.: Verificata in quanto nel lancio di un dado (sia esso onesto o meno) le prove sono i.i.d.
- Numero sufficienti di prove: nel punto precedente si è dimostrato come siano necessarie 30 prove, avendone 120 esse sono sufficienti.

Si hanno le seguenti ipotesi alternative

$$H_0: p_i = 1/6 \quad \forall i \quad H_1: \exists \bar{i}: p_{\bar{i}} \neq 1/6$$

Questo test utilizza come stimatore lo stimatore di Pizzetti-Pearson e con il numero di prove fornite si ha la convergenza della sua distribuzione ad un chi-quadrato, in particolare si ha che

$$\sum_{i=1}^M \frac{(n - \hat{n})^2}{\hat{n}} \sim \chi^2(M-1) \Rightarrow \bar{x} \sum_{i=1}^6 \frac{(n - \hat{n})^2}{\hat{n}} \sim \chi^2(5)$$

Verificata la convergenza dello stimatore è possibile determinare la regione di accettazione A . In questo caso il test da eseguire è di tipo bilaterale. Fissato un livello di significatività del 5 % si ha che

$$A = [0; \chi^2(5)_{0.95}] \Rightarrow A = [0; 11.07]$$

¹ Alcuni autori usano la dizione de "test di adattamento alla distribuzione empirica" volta a sottolineare che si confrontano due distribuzioni: quella teorica con quella empirica.

Per completare il test di deve calcolare lo stimatore di Pizzetti-Pearson, che, utilizzando i calcoli riportati in Tabella 2 risulta pari a 8.8.

i	n	$\check{n}$	$n-\check{n}$	$(n-\check{n})^2$	$(n-\check{n})^2/\check{n}$
1	22	20	2	4	0.200
2	12	20	-8	64	3.200
3	20	20	0	0	0.000
4	30	20	10	100	5.000
5	18	20	-2	4	0.200
6	18	20	-2	4	0.200
Tot	120				8.8

Tabella 2 - dati Esercizio 3: calcolo stimatore Pizzetti-Pearson

Poichè il valore dello stimatore è interno alla regione di accettazione si considera la distribuzione empirica aderente a quella teorica e pertanto il dado è da ritenersi onesto (ad un livello di significatività del 5%).

a) Il candidato calcoli le seguenti Probabilità: $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$ $P(E_1 | E_2)$.

Per il calcolo di $P(E_1)$, utilizzando come insieme universo le facoltà

$$U = \{A, B, C, D, E\}$$

si riconosce che gli eventi elementari sono equiprobabili, pertanto le probabilità dell'evento $E_1 = \{A, B\}$ può essere calcolata utilizzando la definizione classica; secondo la quale la probabilità è il rapporto dei casi favorevoli sui casi totali. Pertanto si ha che: $P(E_1) = 2/5 = 40\%$

La probabilità dell'evento E_2 è pari alla probabilità di avere un esito positivo (ovvero pari ad uno) estraendo un valore da una distribuzione di Bernoulli. Ricordando che una generica distribuzione di Bernoulli di parametro p (ovvero $Ber(p)$) sono possibili solo due esiti (0 ed 1) e che il parametro indica la probabilità dell'esito positivo si ottiene immediatamente che:

$$P(E_2) = P(X=2) = p = 0.2$$

Propedeutico al calcolo delle altre due probabilità e il calcolo della probabilità dell'evento intersezione (ovvero che i due eventi si verifichino contemporaneamente). Per eventi indipendenti (e solo per loro) questa è data dal prodotto delle probabilità dei singoli eventi. Pertanto si ha che

$$P(E_1 \cap E_2) = P(E_1)P(E_2) = 0.4 \cdot 0.2 = 0.08$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.4 + 0.2 - 0.08 = 0.52 \quad P(E_1 | E_2) = P \frac{(E_1 \cap E_2)}{P(E_2)} = \frac{0.08}{0.2} = 0.4$$

b) Il candidato indichi se i due eventi E_1 ed E_2 sono incompatibili.

Se due eventi sono incompatibili si ha che la probabilità dell'evento intersezione è nulla. Nel punto precedente questa probabilità risulta positiva, quindi i due eventi non sono incompatibili.