Federico Di Palma

Raccolta di Temi d'esame di "Statistica" risolti e commentati.

A.A. 2012 - 2013

A Claudia

Prefazione

Il presente fascicolo racchiude i Temi d'esame del corso di base di "Statistica" proposti agli studenti negli AA. AA. 2010/2011 e 2011/2012 nella Facoltà di "Biotecnologie" dell'Università degli Studi di Verona ed è rivolto agli studenti di un corso di base di statistica descrittiva ed inferenziale. Per ogni tema d'esame viene proposta una soluzione commentata con richiami teorici.

Si prega di segnalare ogni refuso al docente tramite e-mail (federico.dipalma@univr.it).

Federico Di Palma

Indice generale

- Appello del 4 Febbraio 2011 - Fila A

Esercizio 1)
Nella tabella seguente viene riportata la distribuzione delle assenze relative al'intero anno scolastico 2009/2010 di una classe IV superiore.

Giorni di assenza	4	5	8	11	16	18	19	25	28
Frequenza	3	2	1	4	5	2	1	4	6

Determinare

a) La tipologia del carattere.
b) Un indice sintetico di posizione.
c) Se possibile, un indice sintetico di variabilità.
d) Una rappresentazione grafica adeguata.
e) L'eventuale presenza di outlier.

Esercizio 2)
E' data la seguente tabella di ricavata da un indagine svolta su 200 lavoratrici di un industria per conoscere le preferenze riguardo all'orario di lavoro in relazione allo stato civile.

		Y:stato civile		
		Nubili	Coniugate	Vedove
X:orario preferito	Diviso (oltre 2 ore di pausa)	12	20	18
	Continuato con breve interruzione	36	50	14
	Continuato senza interruzione	20	20	10

Il candidato

a) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di posizione
b) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di variabilità
c) Verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti.

Esercizio 3)
L'istituto descritto nell'esercizio 1 dichiarava nel A.S. 2008/2009 che il valore atteso delle assenze fosse di 10 gg per una classe IV. Considerando la classe illustrata nell'esercizo 1 come campione è possibile confermare tale asserzione?

Esercizio 4)
Si considerino i seguenti eventi legati all'estrazione di una delle lavoratrici descritte nell'Esercizio 2.

E_1 : si estragga una lavoratrice sposata
E_2: si estragga una lavoratrice che preferirebbe avere un orario continuato

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

- Appello del 4 Febbraio 2011 -
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 numero naturali e concettualmente limitate ad un numero massimo pari ai giorni di lezione presenti nell'anno scolastico)

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 (una persona che ha fatto una grave malattia o un incidente...)

Per calcolare la mediana si deve valutare la numerosità della popolazione (N=28) facilmente ottenibile comulando le frequenze assolute

Giorni di assenza	**4**	**5**	**8**	**11**	**16**	**18**	**19**	**25**	**28**
Frequenza	3	2	1	4	5	2	1	4	6
F. ass. cumulata	*3*	*5*	*6*	*10*	*15*	*17*	*18*	*22*	*28*

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 = 13,5 elementi. Poichè non esiste il l'osservazione di posto 14,5 viene preso come mediana la media fra il 14° ed 15° valore. Analizzando le frequenze cumulate si ottiene che ambo le osservazioni mostrano la modalità 16. Pertanto la mediana (q2) è 16

c) Se possibile, un indice sintetico di variabilità.

Un carattere di tipo quantitativo ammette quattro indici sintetici di varaibilità: il range (o campo di variazione) la distanza interquartile, la varianza e la devizione 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 (q1) edi l terzo (q3) quartile. Dave q1, una volta ordinate le osservazioni, lascia alla propria sinistra (N-1)/4 = 6,75 osservazioni mentre q3 lascia alla proria destra (N-1)/4 = 6,75 osservazioni. Anche in questo caso non ottenendo numeri interi dovremo mediare le posizioni intere più vicine. Si ha dunque che

q1 = media 7° e 8° valore =11 q3 = media 21° e 22° valore =25

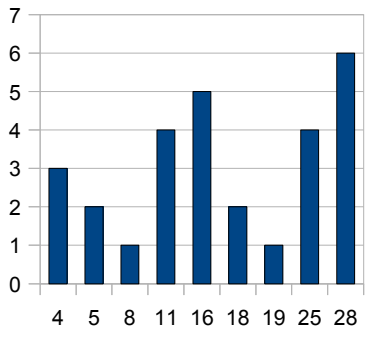
Si ha che la distanza interquartile D = q3 - q1 = 25 - 11 = 14

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 digramma 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 quatile 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

$$VAI = 11 - 14 = -3 \qquad VAS = 25 + 1*14 = 39$$

Non esistendo alcuna osservazioni esterna all'intervallo [-3 ; 39] possiamo concludere che la popolazione presumibilmente non presenta outlier.

Esercizio 2)

L'esercizio verte sull'analisi di una serie bivariata, ottenuta misurando due caratteri qualitativi non ordinabili.

a) 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 sintetici 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 50 da cui si ha le seguente moda

(Continuato con breve interruzione ; Coniugate)

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

Una serie bivariata ottenuta misurando due caratteri qualitativi non ordinabili non ammette indice sintetici di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

c) Verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti.

Per verificare se i due caratteri sono indipiendeti si può effettuare un test di ipotesi volto a verificare se le frequenze delle osservazioni rilevate nel campione sono sufficiente mente vicine (ad un determinato livello di significatività) a quelle teoriche ottenute dall'ipotesi di indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadranto avente gradi di libertà paria quelli del 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 ripotano le frequenze marginali e quelle teoriche fra parentesi

		Y:stato civile			Totali
		Nubili	Coniugate	Vedove	
X:orario preferito	Diviso (oltre 2 ore di pausa)	12 (17)	20 (22,5)	18 (10,5)	50
	Continuato con breve interruzione	36 (34)	50 (45)	14 (21)	100
	Continuato senza interruzione	20 (17)	20 (22,5)	10 (10,5)	50
Totali		68	90	42	200

A questo punto è possibile valutare la convergenza dell 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%.

$$A=\left[0;\chi^2_{1-\alpha}\left((M_x-1)(M_y-1)\right)\right]=\left[0;\chi^2_{1-0.05}\left((3-1)(3-1)\right)\right]=\left[0;\chi^2_{0.95}(4)\right]=[0;9.49]$$

Si può ora procedere al calcolo dello stimatore vero e proprio

$$\frac{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(n_{i,j}-\hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}}=$$

$$=\frac{(12-17)^2}{17}+\frac{(20-22.5)^2}{22.5}+\frac{(18-10.5)^2}{10.5}+\frac{(36-34)^2}{34}+\frac{(50-45)^2}{45}+\frac{(14-21)^2}{21}+\frac{(20-17)^2}{17}+\frac{(20-22.5)^2}{22.5}+\frac{(10-10.5)^2}{10.5}=$$

$$\frac{25}{17}+\frac{6.25}{22.5}+\frac{56.25}{10.5}+\frac{4}{34}+\frac{25}{45}+\frac{49}{21}+\frac{9}{17}+\frac{6.25}{22.5}+\frac{0.25}{10.5}=10.94$$

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 per cento.

Esercizio 3)

Nel testo viene richiesto di verificare se il valore atteso della popolazione da cui si è estratto il campione indicato nell'esercizio 1 e pari a 10.
Questo test si appoggia allo stimatore media campionaria e richiede un campione la cui dimensione sia di almeno 30 elementi. Non soddisfacendo questa ipotesi non è possibie confermare o smentire l'ipotesi.

Esercizio 4)

a) Il candidato calcoli le seguenti Probabilità: P(E_1); P(E_2); P(E_1 U E_2) P(E_1 | E_2).
Essendo gli eventi elementari equiprobabili, le probabilità degli eventi E_1 ed E_2 e dell'evento intersezione (estrarre donne coniugate che prefriscono orairio continuato) possono essere ricavate utilizzando la definizione classica; secondo la quale la probabilità è il rapporto dei casi favorevoli sui casi totali. Pertanto si ha che:

$$P(E_1)=\frac{90}{200}=0.45 \qquad P(E_2)=\frac{100+50}{200}=0.75 \qquad P(E_1\cap E_2)=\frac{50+20}{200}=0.35$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1\cup E_2)=P(E_1)+P(E_2)-P(E_1\cap E_2)=\frac{90+150-70}{200}=0.85 \qquad P(E_1|E_2)=P\frac{(E_1\cap E_2)}{P(E_2)}=\frac{70}{200}\frac{200}{170}=\frac{7}{17}$$

b) Il candidato indichi se i due eventi E_1 ed E_2 sono indipendenti.
Se due eventi sono indipendenti si ha che la probabilità condizionata è data dal prodotto delle probabilità, pertanto essendo

$$P(E_1)P(E_2)=\frac{90}{200}\frac{150}{200}=\frac{27}{80}\neq\frac{7}{17}=P(E_1|E_2)$$

Gli eventi non sono indipendenti.

- Appello del 4 Febbraio 2011 -Fila B

Esercizio 1)
Nella tabella seguente viene riportata la valutazione del livello di gradimento del corso di statistica dell A.A. 2007/2008 per una facoltà di Economia.

Gradimento	Ottimo	Buono	Discreto	Sufficiente	Insufficiente	Gravemente Insufficiente
Frequenza	3	2	1	4	5	2

Determinare

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

Esercizio 2)
Da un'indagine si sono rilevate in 6 piccole aziende italiane (indicate con lettere A-F) il profitto ed il valore delle spese sostenute per ammodernare gli impianti espressi in migliaia di euro. I dati ottenuti sono rappresentati nella forma Azienda(Profitto; Spesa).

A(50; 20) B(60; 40) C(30; 14) D(85;50) E(95;60) F(40;26)

Il candidato,

a) Indichi e fornisca una rappresentazione grafica adeguata alla serie ottenuta.
b) Se possibile, indichi e calcoli un opportuno indice di variabilità
c) Ipotizzando un legame di tipo lineare,
    1. Calcoli l'opportuna regressione
    2. Ipotizzi quale sarebbe l'investimento previsto nel caso si riscontrasse un profitto di 100 mila euro
    3. Il legame ipotizzato è attendibile? Motivare numericamente la risposta.

Esercizio 3)
Si vuole verficare la bontà di una roulette classica composta da 36 numeri (18 neri e 18 rossi), e due numeri detti "verdi" (zero e doppio zero). In particolare, si è interessati a verificare che la probabilità che vinca il banco (esca zero o doppio zero) sia equa.
Il candidato:

a) determini il numero di osservazioni necessarie affinchè si possa procedere a tale verifica
b) supposto di aver eseguito 380 prove, indicare se la roulette è equa a fronte delle seguenti frequenze assolute

Esito	Rosso	Nero	"0"	"00"
Frequenza	176	188	10	6

Esercizio 4)
Si considerino i seguenti eventi legati all'estrazione di una delle 6 aziende descritte nell'Esercizio 2.

E_1: si estragga un'azienda che spende oltre 45 mila euro
E_2: si estragga un'azienda che ricava oltre 45 mila euro

a) Il candidato calcoli le seguenti Probabilità: $P(E_1); P(E_2); P(E_1 \cup E_2)\ P(E_1 \mid E_2)$.
b) Il candidato indichi se gli eventi E_1 ed E_2 possono ritenersi statisticamente indipendenti.

- Appello del 4 Febbraio 2011 -
Svolgimento - Fila B

Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo qualitativo (in quanto non espresso da numeri) ordinabili (in quanto e possibile fissare un ordine fra le modalità)

b) Tutti gli indici sintetici di posizione possibili da calcolare.

Un carattere di tipo qualitativo ordinabile ammette due indici sintetici di posizione: la moda e la mediana.

La moda è che la modalità con la frequenza maggiore: pertanto la moda è "Insufficiente"

Per calcolare la mediana si deve valutare la numerosità della popolazione (N=17) facilmente ottenibile comulando le frequenze assolute

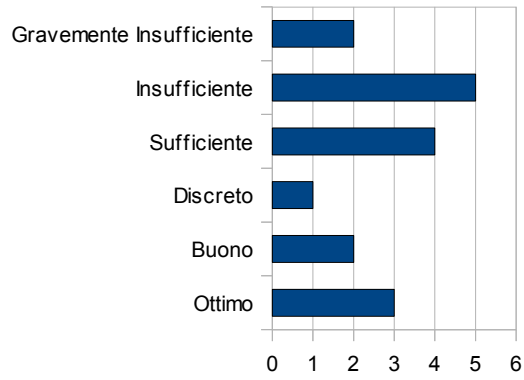
Gradimento	Ottimo	Buono	Discreto	Sufficiente	Insufficiente	Gravemente Insufficiente
Frequenza	3	2	1	4	5	2
Cumulata	3	5	6	10	15	17

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 = 8 elementi; ovvero il nono elemento. Analizzando le frequenze cumulate si ottiene che la mediana indicherà la modalità "Sufficiente" (che infatti raccoglie le osservazioni dal 7° al 10 ° posto).

c) Tutti gli indici sintetici di posizione possibili da calcolare.

Un carattere di tipo qualitativo non ammette alcun indici sintetici di variabilità.

d) Una rappresentazione grafica adeguata.



Un carattere di tipo qualitativo ordinabile 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.

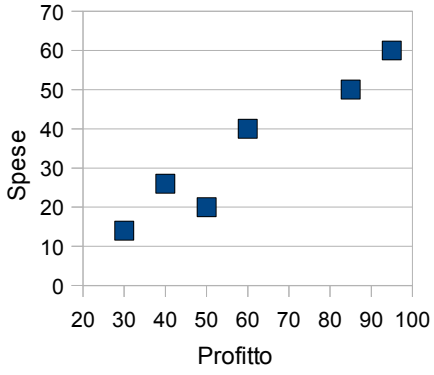
A lato si riporta il digramma a barre ricavato dalla distribuzione in esame

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: Profitto realizzato dall'aziende

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{n} x_i=\frac{50+60+30+85+95+40}{6}=60$$
$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{n} x_i^2-\bar{x}^2=\frac{2500+3600+900+7225+9025+1600}{6}-3600=\frac{24850-21600}{6}=\frac{3250}{6}$$

Y: Spesa per ammodernamento effettuata dall'aziende

$$\bar{y}=\frac{1}{N}\sum_{i=1}^{n} y_i=\frac{20+40+14+50+60+26}{6}=35$$
$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{n} y_i^2-\bar{y}^2=\frac{400+1600+196+2500+3600+676}{6}-1225=\frac{8972-7350}{6}=\frac{1622}{6}$$

La covarianza si ottiene

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

I cui conti sono ripostati in tabella

	X	**Y**	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
	50	**20**	-10	-15	150
	60	**40**	0	5	0
	30	**14**	-30	-21	630
	85	**50**	25	15	375
	95	**60**	35	25	875
	40	**26**	-20	-9	180
somma	**360**	**210**			**2210**

Per tanto la matrice varianza covarianza risulta essere

$$\Sigma=\begin{bmatrix}\frac{3250}{6} & \frac{2210}{6}\\ \frac{2210}{6} & \frac{1622}{6}\end{bmatrix}$$

c 1) Ipotizzando un legame di tipo lineare, si calcoli l'opportuna regressione
La retta di regressione ha equazione

$$\hat{y}=\frac{\sigma_{xy}}{\sigma_x^2}x+\bar{y}-\frac{\sigma_{xy}}{\sigma_x^2}\bar{x} \qquad \hat{y}=\frac{2210}{3250}x+35-\frac{2210}{3250}60 \qquad \hat{y}=0.68\,x-5.8$$

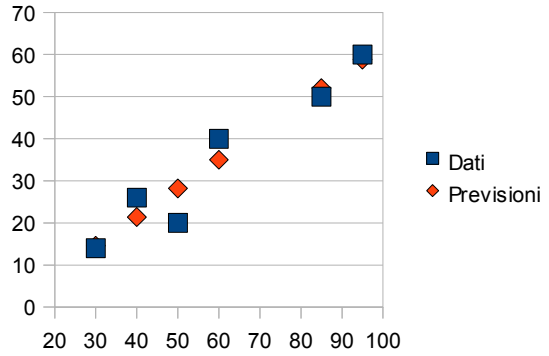
c 2) Ipotizzando un legame di tipo lineare, si ipotizzi quale sarebbe l'investimento previsto nel caso si riscontrasse un profitto di 100 mila euro
La risposta a questo quesito si ottiene applicando la retta nel punto x= 100. si ottiene quindi un investimento previsto di 62.2 mila euro.

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

$$R^2=\frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}=0.9265 \qquad R=0.9625$$

Poiche l'indice risulta superiore a 0.7 si può asserire che il legame è 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 ben descrivono l'andamento dei dati.



Esercizio 3)

L'indagine statistica mira a verificare mediante inferenza se la probabilità di ottenere uno zero od un doppio zero e onesta. L'analisi si riduce a verificare le probabilità di due eventi complementari:

A: esca un numero "verde" A: non esca un numero "verde"

Si ha inoltre che P(A) = 2/38 = 1/19 e P(A) = 1- P(A) = 18/19

Pertanto è possibile modellare la popolazione di riferimento mediante una bernoulliana *P~Ber(1/19)* dove *E*[*P*]=1/19.

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

la dimensione nel campione varia a seconda del tipo di test da effettuare: se si utilizza il test di adattamento alla distribuzione empirica o quello sul valore atteso.

Nell'ipotesi di agire usando il test sul valore atteso si ha che la dimensione del campione deve essere superiore alle 30 unità statistiche.

b) supposto di aver eseguito 380 prove, indicare se la roulette è equa a fronte delle seguenti frequenze assolute

Il test impostato è un test di ipotesi sul parametro valore atteso. Si hanno le seguenti ipotesi

$$H_0: E[P]=1/19 \qquad H_1: E[P]\neq 1/19$$

Questo test utilizza come stimatore la media campionaria e, poiche la dimensione del campione è superiore alle 30 unità, si ha la convergenza della sua distribuzione ad una normale, si ha infatti che

$$\bar{x}\sim N\left(p,\frac{p(1-p)}{n}\right)\Rightarrow \bar{x}\sim N\left(\frac{1}{19},\frac{\frac{1}{19}\cdot\frac{18}{19}}{380}\right)\Rightarrow \bar{x}\sim N\left(\frac{1}{19},\left(\frac{3}{19}\right)^2\cdot\frac{1}{19}\right)$$

Verificata la convergenza dello stimatore è possibile determinare la regione di accettazione *A*. Essendo l'ipoesi alternativa un ipotesi di disuguaglianza il test da eseguire è di tipo bilaterale. Fissato un livello di significatività del 5 % si ha che

$$A=\left[-z_{1-\frac{\alpha}{2}};z_{1-\frac{\alpha}{2}}\right]\Rightarrow A=[-1.96;1.96]$$

Si calcola ora il valore dello simatore standardizzato

$$\bar{x}=\frac{10+6}{380}\Rightarrow z_{\bar{x}}=\frac{\bar{x}-E[P]}{\sqrt{Var[P]}}=\frac{\frac{16}{380}-\frac{1}{19}}{\sqrt{\frac{3^2}{19^2}\cdot\frac{1}{19}}}=\frac{-\frac{2}{190}}{\frac{3}{19}\cdot\frac{1}{\sqrt{19}}}=-\frac{2}{190}\cdot\frac{19}{3}\cdot\sqrt{19}=-\frac{2}{10}\cdot\frac{\sqrt{19}}{3}=-0.2906$$

Poichè il valore ottenuto è interno alla regione di accettazione possiamo accettare l'ipotesi nulla.

Esercizio 4)

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

Essendo gli eventi elementari equiprobabili, le probabilità degli eventi E_1 ed E_2 e dell'evento intersezione (estrarre un'azienda che spenda e ricavi oltre 45 mila euro) possono essere ricavate utilizzando la definizione classica; secondo la quale la probabilità è il rapporto dei casi favorevoli sui casi totali. Pertanto si ha che:

$$P(E_2)=\frac{4}{6}=0.667 \qquad P(E_1)=\frac{3}{6}=0.5 \qquad P(E_1 \cap E_2)=\frac{2}{6}=0.333$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1 \cup E_2)=P(E_1)+P(E_2)-P(E_1 \cap E_2)=\frac{4+3-2}{6}=0.833 \qquad P(E_1|E_2)=P\frac{(E_1 \cap E_2)}{P(E_2)}=\frac{2/6}{3/6}=0.667$$

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

Se due eventi sono indipendenti si ha che la probabilità condizionata è data dal prodotto delle probabilità, pertanto essendo

$$P(E_1)P(E_2)=\frac{1}{2}\frac{1}{3}=0.5 \neq 0.667=P(E_1|E_2)$$

Gli eventi non sono indipendenti.

- Appello del 18 Febbraio 2011 - Fila A

Esercizio 1)
Si vuole valutare il tempo di incubazione (espresso in giorni) di un agente virale. Da un osservazione su di una popolazione di 20 elementi si sono ottenute le frequenze assolute indicate nella tabella a lato.

inf_i	sup_i	n_i
0	8	1
8	12	2
12	16	5
16	20	4
20	24	5
24	28	2
28	36	1

Il candidato

a) Determini la tipologia del carattere.
b) Se possibile, tracci l'istogramma.
c) Se possibile, calcoli la mediana.
d) Se possibile, calcoli la varianza.

N.b. L'estremo superiore delle varie classi di modalità è da ritenersi escluso

Esercizio 2)
I dati raccolti nel precedente esercizio sono stati organizzati tenendo conto del diverso genere del soggetto che ha contratto il virus, ottenendo la seguente tabella.

		Y:tempo di incubazione				
		fino a 12 gg	da 12 a 16 gg (16 escluso)	da 16 a 20 gg (20 esluso)	da 20 a 24 gg (24 escluso)	24 gg e oltre
X:Genere	Maschile	1	2	2		
	Femminile				2	1

Il candidato

a) Completi la tabella con i dati mancanti.
b) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di posizione
c) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di variabilità
d) Se possibile, verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti.

Esercizio 3)
Un laboratorio di ricerca vuole stimare la varianza di un microscopio elettronico. Per fare questo effettuate 11 misure di un campione di lunghezza nota 5 nm. Le misure (espresse in nm) ottenute sono:

5,01 5,00 4,99 5,01 5,00 5,01 5,00 5,00 5,00 4,98 5,00

Il candidato stimi puntualmente e per intervallo lo scarto quadratico medio delle misurazioni.

Esercizio 4)
Si considerino i due eventi E_1 ed E_2. Sapendo che i due eventi sono indipendenti e $P(E_1) = 1/2$; $P(E_2) = 1/3$. Il candidato calcoli le probabilità dei seguenti eventi

a) evento E_2 condizionato E_1
b) evento E_1 intersezione E_2.
c) evento E_2 unito E_1.

- Appello del 18 Febbraio 2011 -
Svolgimento Fila A

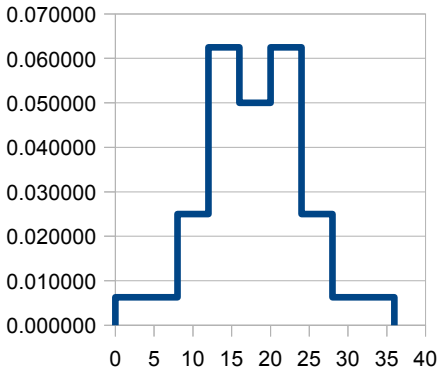
Esercizio 1)

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 concettualemente è continuo).

b) Se possibile, tracci l'istogramma..

L'istogramma è una rappresentazione comunemente utilizzata quando si tratta un dato quantitativo continuo che, per diverse esigenze, viene rappresentato in classi di modalita 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 istrogramma 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	8	4	8	1	0,050	0,050	0,006250	0,200	16	0,80
2	8	12	10	4	2	0,100	0,150	0,025000	1,000	100	10,00
3	12	16	14	4	5	0,250	0,400	0,062500	3,500	196	49,00
4	16	20	18	4	4	0,200	0,600	0,050000	3,600	324	64,80
5	20	24	22	4	5	0,250	0,850	0,062500	5,500	484	121,00
6	24	28	26	4	2	0,100	0,950	0,025000	2,600	676	67,60
7	28	36	32	8	1	0,050	1,000	0,006250	1,600	1024	51,20
				Totali	20	1			18		364,40

c) Se possibile, calcoli la mediana.

La mediana è il valore che bipartisce la popolazione, ovvero, una volta ordinate le osservazioni si ricerca quella che lascia alla sua destra 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. Dal calcolo delle frequenze cumulate (F_i) si vede come la mediana cada nella 4 classe (prima classe a superare lo 0,5). Per determinare l'esatto valore basta imporre che l'area presente nella classe 5 sia sufficiente a raggiungere il valore di 0.5. Si ottiene quindi il seguente conto:

$$(q_2-16)*0.05=(0.5-0.4)\Rightarrow q_2-16=0.1/0.05\Rightarrow q_2=2+16=18$$

d) Se possibile, si calcoli la varianza.

La varianza nel caso siano presenti osservazioni raggruppate in classi si calcola utilizzando come modalità i valori centrali delle classi (c_i). Nella tabella alla fine del punto b) è stato riportato il calcolo della varianza utilizzando la formula abbreviata.

$$\sigma^2=\left(\sum_{i=1}^{M} c_i^2 * f_i\right)-\left(\sum_{i=1}^{M} c_i * f_i\right)^2=364{,}40-18^2=40{,}4$$

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).

Esercizio 2)

a)Completi la tabella con i dati mancanti.

La tabella si completa tenendo conto che la somma delle colonne deve coincidere con i dati illustrati

nell'esercizio 1. Si noti che nella nuova formulazione alcune classi di modalità sono state aggregate.

		Y:tempo di incubazione				
		fino a 12 gg	da 12 a 16 gg (16 escluso)	da 16 a 20 gg (20 esluso)	da 20 a 24 gg (24 escluso)	24 gg e oltre
X:Genere	Maschile	1	2	2	3	2
	Femminile	2	3	2	2	1

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativo non ordinabile ammette un solo indice sintetico di posizione: la moda. La moda di una bi-variata si ottiene valutando la o le modalità della serie corrispondenti alla frequenza (assoluta o relativa) maggiore. Nel caso in esame la frequenza assoluta maggiore è 3 cui corrispondono due modalità (distribuzione bi-modale)

(Maschile; Da 20 a 24 gg) e (Femminile; Da 12 a 16 gg)

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativio non ordinabile non ammette indice sintetici di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

d) Verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti.

Per verificare se i due caratteri sono indipendenti 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 indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadrato avente gradi di libertà pari a quelli del 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 ripotano le frequenze marginali e quelle teoriche fra parentesi

		Y:tempo di incubazione					
		fino a 12 gg	da 12 a 16 gg (16 escluso)	da 16 a 20 gg (20 esluso)	da 20 a 24 gg (24 escluso)	24 gg e oltre	Totali
X:Genere	Maschile	1 (1.5)	2 (2.5)	2 (2)	3 (2.5)	1 (1.5)	10
	Femminile	2 (1.5)	3 (2.5)	2 (2)	2 (2.5)	2 (1.5)	10
Totali		3	5	4	5	3	20

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 non è verificata si può concludere che non è possibile ricevare l'informazione richiesta dalle osservazioni fornite.

Esercizio 3)

Nel testo si effettuano diverse misure di una grandezza nota. Possiamo modellare questo problema come l'estrazione di una variabile casuale X avente distribuzione ignota e valore atteso 5.

Si sono effettuate N= 11 estrazioni aventi M=4 modalità

a) stimare puntualmente la varianza.
Continuando con il modello precedentemente fatto il punto richiede di stimare lo scarto quadratico medio ovvero la radice quadrata di Var[X]. Questa stima può essere effettuata ricordando che la varianza viene stimata correttamente mediante la varianza campionaria (s^2). Il calcolo di s^2 in presenza di osservazioni ripetute, (frequenze assolute maggiori di uno) è dato dalla seguente:

$$s^2 = \frac{\sum_{i=1}^{M} n_i \left(x_i - \bar{x}\right)^2}{N-1} = \frac{0.0008}{10} = 0.00008 \Rightarrow S = 0.00894$$

Il calcolo della varianza è stato fatto utilizzando la seguente tabella

i	x_i	n_i	$x_i\, n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$
1	4,98	1	4,9800	-0,0200	0,0004	0,00040
2	4,99	1	4,9900	-0,0100	0,0001	0,00010
3	5,00	6	30,0000	0,0000	0,0000	0,00000
4	5,01	3	15,0300	0,0100	0,0001	0,00030
	Totali	11	55,0000			0,00080

b) stimare per intervallo la varianza.
La stima della varianza per intervallo si ha considerando la distribuzione di partenza gaussiana ed *n* grande. Nel caso in esame considerare la distribuzione di partenza gaussiana non introduce un errore elevato (trattasi di errori di misura quindi nello specifico simmetrici) per quanto riguarda la dimensione del campione è possibile ritenere *n = 11* una dimensione sufficiente.

Validate le ipotesi si ha che la stima per intervallo della varianza è data dalla $\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$

ponendo un livello del 5 % si ha che: $Var[X] \in \left[\frac{10 \cdot 0.00008}{20.5}; \frac{10 \cdot 0.00008}{3.25}\right] = [0.000039; 0.000246]$

Pertanto l'intevallo richiesto è: $sqm = \sqrt{Var[X]} \in [\sqrt{0.000039}; \sqrt{0.000246}] = [0.0062; 0.01569]$

Esercizio 4)

Si noti come l'esercizio fissa la probabilità degli eventi elementari e richiede il computo delle probabilità di eventi complessi, pertanto richiede l'applicazione della definizione assiomatica di probabilità.

*a) Il candidato calcoli Probabilità dell'*evento E_2 condizionato E_1

La probabilità richiestà $P(E_2 \mid E_1)$ viene calcota immediatamente ricordando che gli eventi statisticamente indipendenti sono quelli per cui il verificarsi di un evento non altera la probabilità di verificarsi dell'altro. Pertanto si ha che $P(E_2 \mid E_1) = P(E_2) = 1/3$.

b) Il candidato calcoli Probabilità dell'evento E_1 intersezione E_2.

La probabilità dell'evento intersezione di due eventi indipendenti (ovvero che i due eventi si verifichino entrambi) è data dal prodotto delle due probabilità. Si ha infatti

$$P(E_1 \cap E_2) = P(E_1)P(E_2) = \frac{1}{2}\frac{1}{3} = \frac{1}{6}$$

Si noti come lo stesso risultato poteva essere raggiunto elaborando la definzione di probabilità condizionata:

$$P(E_2|E_1) = P\frac{(E_1 \cap E_2)}{P(E_1)} \Rightarrow P(E_2|E_1)P(E_1) = P(E_1 \cap E_2) \underset{E_1, E_2 \text{ indip.}}{\Rightarrow} P(E_2)P(E_1) = P(E_1 \cap E_2)$$

c) Il candidato calcoli la Probabilità dell'evento E_1 unito E_2..

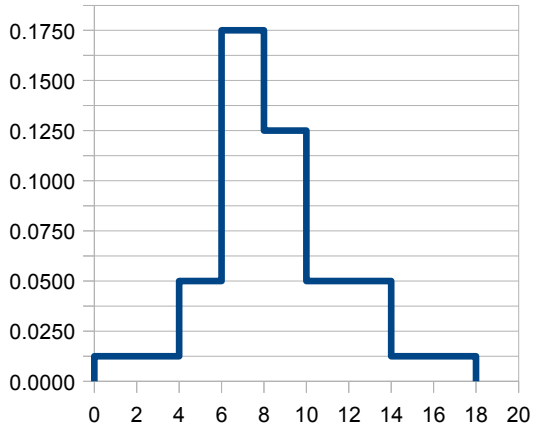
Note le probabilità degli eventi elementari e dell'evento intersezione si ha che

$$P(E_1 \cup E_u) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{3+2-1}{6} = \frac{2}{3}$$

- Appello del 18 Febbraio 2011 - Fila B

Esercizio 1)

Si vuole valutare il tempo di incubazione (espresso in giorni) di un agente virale. Da un osservazione su di una popolazione di 20 elementi si è ottenuto l'istogramma nella figura a lato.



Il candidato

a) Determini la tipologia del carattere.
b) Fornisca una rappresentazione tabellare dei dati (mettendo in risalto le frequenze assolute).
c) Se possibile, calcoli la mediana.
d) Se possibile, calcoli la varianza.

N.b. L'estremo superiore delle singole classi di modalità è da ritenersi escluso

Esercizio 2)

Per verificare la difficoltà di un corso di laurea si è voluto monitorare il numero di anni fuori corso che un laureato magistrale ha maturato durante il suo percorso di studi. I dati relativi ad un campione di 100 laureati è riassunto nella seguente tabella a doppia entrata.

		Y: anni fuori corso laurea triennale				
		0	**1**	**2**	**3**	**Totali**
X:Anni fuori corso magistrale	**0**		6	3		10
	1	6		10	10	
	2	3	10		10	40
Totali			40	30	20	100

Il candidato

a) Completi la tabella con i dati mancanti.
b) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di posizione
c) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di variabilità
d) Se possibile, verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti.

Esercizio 3)

Si supponga di modellare mediante una v.c. W il numero di anni fuori corso totali (triennio + biennio) accumulati da un laureato specialistico. Date le osservazioni dell'Esercizio 2 il candidato stimi puntualmente e per intervallo il valore atteso di W.

Esercizio 4)

Si considerino i due eventi E_1 ed E_2. Sapendo che i due eventi sono incompatibili e che $P(E_1) = 1/2$; $P(E_2) = 1/3$. Il candidato calcoli le probabilità dei seguenti eventi

a) evento E_1 intersezione E_2
b) evento E_2 condizionato E_1
c) evento E_2 unito E_1.

- Appello del 18 Febbraio 2011 -
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 concettualemente è continuo).

b) Fornisca una rappresentazione tabellare dei dati.

L'istogramma è una rappresentazione comunemente utilizzata quando si tratta un dato quantitativo continuo che viene, per diverse esigenze, raccolto in classi di modalita c_i. Il grafico riporta le modalità sull'asse delle ascisse e sulle ordinate la densità di frequenza di ogni classe. Esso si compone di un rettangolo per ogni classe. I rettangoli sono fra di loro adiacenti e dalle loro basi si ricavano gli estremi della classe corrispondente (sup_i e inf_i) mentre l'altezza coincide con la densità di frequenza (d_i). Quindi l'area di ogni rettangolo sarà uguale alla frequenza relativa della classe ($f_i = d_i$.*(sup_i e inf_i)). Pertanto, la frequenza assoluta può essere ottenuta moltiplicando l'area del rettangolo per la dimensione del campione (N=20). Applicando quanto descritto è possibile ottenere la seguente rappresentazione tabellare.

i	inf_i	sup_i	sup_i-inf_i	di	f_i	n_i	F_i	c_i	$c_i f_i$	c_i^2	$c_i^2 f_i$
1	0	4	4	0,01250	0,050	1	0,050	2	0,10	4	0,200
2	4	6	2	0,05000	0,100	2	0,150	5	0,50	25	2,500
3	6	8	2	0,17500	0,350	7	0,500	7	2,45	49	17,150
4	8	10	2	0,12500	0,250	5	0,750	9	2,25	81	20,250
5	10	14	4	0,05000	0,200	4	0,950	12	2,40	144	28,800
6	14	18	4	0,01250	0,050	1	1,000	16	0,80	256	12,800
				Totali	1,000	20			8,50		81,70

c) Se possibile, calcoli la mediana.

La mediana è il valore che bipartisce la popolazione, ovvero, una volta ordinate le osservazioni si ricerca quella che lascia alla sua destra 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. Dal calcolo delle frequenze cumulate (F_i) si vede come la mediana cada all'estremità superiore della classe 3. Pertanto si può asserrire $q_2 = 8$.

d) Se possibile, si calcoli la varianza.

La varianza nel caso siano presenti osservazioni raggruppate in classi si calcola utilizzando come modalità i valori centrali delle classi (c_i). Nella tabella alla fine del punto b) è stato riportato il calcolo della varianza utilizzando la formula abbreviata.

$$\sigma^2 = \left(\sum_{i=1}^{M} c_i^2 * f_i\right) - \left(\sum_{i=1}^{M} c_i * f_i\right)^2 = 81{,}70 - 8.5^2 = 9.45$$

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).

Esercizio 2)

a)Completi la tabella con i dati mancanti.

La tabella si completa tenendo conto che la somma delle colonne e delle righe deve coincidere con le distribuzioni marginali e con il totale delle osservazioni (*N = 100*).

		Y: anni fuori corso laurea triennale				
		0	**1**	**2**	**3**	**Totali**
X:Anni fuori corso magistrale	**0**	1 (1)	6 (4)	3 (3)	0 (2)	10
	1	6 (5)	24 (20)	10 (15)	10 (10)	50
	2	3 (4)	10 (16)	17 (16)	10 (8)	40
	Totali	10	40	30	20	100

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

La bivariata è composta da due caratteri quantitativi discreti. Pertanto è possibile calcolare la media come indice di poszione. In una bi-variata la media può essere calcolata raccogliendo in un vettore le medie dei due caratteri calcolate separatamente a partire dalle rispettive marginali.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{M_x} n_{i,+} \cdot x_i = \frac{0\cdot 10 + 1\cdot 50 + 2\cdot 40}{100} = \frac{130}{100} = 1.3$$
$$\bar{y} = \frac{1}{N}\sum_{j=1}^{M_y} n_{+,j} \cdot y_j = \frac{0\cdot 10 + 1\cdot 40 + 2\cdot 30 + 3\cdot 20}{100} = \frac{160}{100} = 1.6$$

da cui si ricava che la media è (1.3; 1.6).

c) 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: Anni fuori corso durante la laurea triennale

$$\sigma_x^2 = \frac{\sum_{i=1}^{M_x} n_{i,+} x_i^2}{N} - \bar{x}^2 = \frac{10\cdot 0^2 + 50\cdot 1^2 + 40\cdot 2^2}{100} - 1.3^2 = 2.1 - 1.69 = 0.41$$

Y: Anni fuori corso durante la laurea magistrale

$$\sigma_y^2 = \frac{\sum_{j=1}^{M_y} n_{+,j} y_j^2}{N} - \bar{y}^2 = \frac{10\cdot 0^2 + 40\cdot 1^2 + 30\cdot 2^2 + 20\cdot 3^2}{100} - 1.6^2 = 3.4 - 2.56 = 0.84$$

La covarianza si ottiene

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{M_x}\sum_{j=1}^{M_y} n_{i,j}(x_i - \bar{x})(y_j - \bar{y}) = 0.14$$

Pertanto la matrice varianza covarianza risulta essere

$$\Sigma = \begin{bmatrix} 0.41 & 0.14 \\ 0.14 & 0.84 \end{bmatrix}$$

d) Verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti.

Per verificare se i due caratteri sono indipiendeti si può effettuare un test di ipotesi volto a verificare se le frequenze delle osservazioni rilevate nel campione sono sufficiente mente vicine (ad un determinato livello di significatività) a quelle teoriche ottenute dall'ipotesi di indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadranto avente gradi di libertà pari a quelli del 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 a doppia entrata indicata al punto *a)* si ripotano le frequenze teoriche fra parentesi

A questo punto è possibile valutare la convergenza dell stimatore di Pizzetti Pearson, possibile solo se tutte le frequenze teoriche sono superiori a 5. Constatato che la condizione non è verificata si può concludere che non è possibile ricevare l'informazione richiesta dalle osservazioni fornite.

Esercizio 3)
La W è combinazione lineare delle v.c. X ed Y introdotte nell'esercizio 2, in particolare si ha che W = X + Y.
Ciononstante non è possivibile utilizzare le informazioni calcolate su X ed Y (valore atteso e varianza) per trarre conclusioni su W in quanto non siamo in grado di verificare l'indipendenza di X ed Y.
Pertanto si deve considerare la distribuzione di W e procedere al calcolo senza considerare le informazioni ottenute dall'analisi delle vv.cc. X ed Y. La distribuzione è riportata nella tabella sottostante

i	w_i	n_i	$w_i n_i$	$w_i - \overline{w}$	$(w_i - \overline{w})^2$	$n_i(w_i - \overline{w})^2$
1	0	1	0	-2,9	8,41	8,41
2	1	12	12	-1,9	3,61	43,32
3	2	30	60	-0,9	0,81	24,3
4	3	20	60	0,1	0,01	0,2
5	4	27	108	1,1	1,21	32,67
6	5	10	50	2,1	4,41	44,1
Totali		100	290			153

Da cui si ricavano facilemente sfruttando i calcoli in tabella media ($\overline{w}$ = 290/100 = 2.9), varianza (σ^2 = 153 /100 = 1.53) e varianza campionaria (s^2 = 153/99 = 1.55)

La stima puntuale del valore atteso di W è coincide con la media campionaria $E[W] = 2.9$.

La stima per intervallo del valore atteso non conoscendo il corretto valore della varianza della v.c. W è data dalla formula $E[W] \in \left[\overline{w} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \; ; \; \overline{w} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]$ dove s è la radice quadrata della varianza campionaria, α è il livello di confidenza e Z è la normale standardizzata.

Fissando un livello di confidenza al 5% si ottiene la seguente stima.

$$E[W] \in \left[\overline{w} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}} \; ; \; \overline{w} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}} \right] = \left[2.9 - 1.96 \cdot \sqrt{\frac{1.55}{100}} \; ; \; 1.3 + 1.96 \cdot \sqrt{\frac{1.55}{100}} \right] = [2.66 \; ; \; 3.14]$$

Esercizio 4)
Si noti come l'esercizio fissi la probabilità degli eventi elementari e richieda il computo delle probabilità di eventi complessi, pertanto richiede l'applicazione della definizione assiomatica di probabilità.

a) Il candidato calcoli Probabilità dell'evento E_1 intersezione E_2.

Due eventi incompatibili non possono verificarsi contemporaneamente, pertanto l'insieme intersezione è l'insieme nullo. Quindi la probabilità dell'evento intersezione (ovvero l'evento rappresentato dal verificarsi contemporaneo dei due eventi di partenza) è nulla.

$$P(E_1 \cap E_2) = 0$$

b) Il candidato calcoli la Probabilità dell'evento E_2 condizionato E_1

La probabilità richiesta viene calcolata applicando la definzione di probabilità condizionata:

$$P(E_2|E_1) = P\frac{(E_1 \cap E_2)}{P(E_1)} = \frac{0}{P(E_1)} = 0$$

c) Il candidato calcoli la Probabilità dell'evento E_1 unito E_2..

Note le probabilità degli eventi elementari e dell'evento intersezione, si ha che

$$P(E_1 \cup E_u) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{1}{2} + \frac{1}{3} - 0 = \frac{3+2}{6} = \frac{5}{6}.$$

- Appello del 24 Giugno 2011 - Fila A

Esercizio 1)
Si vuole stabilire il livello medio di glucosio che un adulto sano presenta nel sangue durante le ore di sonno. Pertanto si è condotta una sperimentazione in cui sono stati coinvolti 10 soggetti. Per ogni soggetto si sono effettuati 3 prelievi (uno ogni due ore) ottenendo le seguenti misurazioni di concetrazione di glucosio espresse in *mg/dl* ed ordinate in maniera crescente.

140	140	140	143	145	145	147	148	148	148
148	148	149	149	149	150	151	151	151	151
152	152	152	152	152	155	156	158	160	170

Il candidato

a) Determini la tipologia del carattere.
b) Se possibile, tracci il box plot.
c) Se possibile, calcoli la varianza.
d) Se possibile, calcoli un indice di asimmetria adeguato.

Esercizio 2)
I dati raccolti nel precedente esercizio sono stati organizzati tenendo conto del diverso genere del soggetto coinvolto nella sperimentazione, ottenendo la seguente tabella.

		Y: concentrazione di glucosio mg/dl						
		fino a 142	da 143 a 149 (149 escluso)	da 149 a 152 (152 esluso)	da 152 a 155 (155 escluso)	da 155 a 160 (160 escluso)	160 ed oltre	
X: Genere	M	3	6	4				
	F				1	2	2	12
		3	9		5	3	2	

Il candidato

a) completi la tabella con i dati mancanti.
b) indichi e calcoli, se possibile, un opportuno indice di posizione per la serie bivariata
c) indichi e calcoli, se possibile, un opportuno indice di variabilità per la serie bivariata
d) se possibile, verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti. Nel caso non fosse possibile indichi una possibile strategia per effettuare il calcolo.

Esercizio 3)
Il candidato stimi puntualmente e per intervallo il valore atteso della concentrazione di glucosio in un adulto basandosi sulle misurazioni di concentrazione riportate nell'Esercizio 1.

Esercizio 4)
Si considerino i due eventi relativi ai dati dell'Esercizio 2

E_1 : estraendo a caso un componente della sperimentazione, questa è una donna.
E_2 : estraendo a caso una misura di concetrazione di glucosio fra le 30 effettuate durante la sperimentazione, questa è compresa fra 152 e 155 (155 escluso).

Il candidato calcoli le probabilità dei seguenti eventi

a) E_1 ed E_2
b) evento E_1 intersezione E_2
c) evento E_2 condizionato E_1
d) evento E_2 unito E_1.

- Appello del 24 Giugno 2011 -
Svolgimento

Esercizio 1)

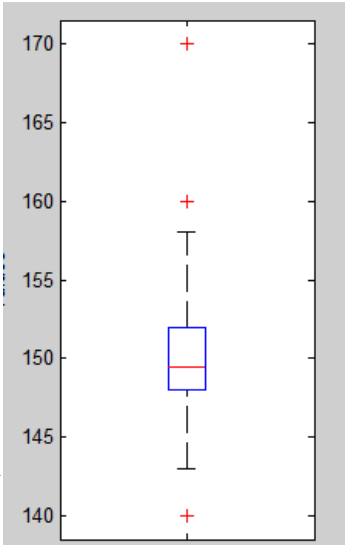
a) Determini la tipologia del carattere.

Il carattere è di tipo quantitativo (in quanto espresso da numeri) continuo (in quanto si vuole monitorare una concentrazione che concettualemente è continua).

b) Se possibile, tracci il boxplot.

Il box-plot è una rappresentazione grafica utile per rappresentare dati quantitativi siano essi continui o discreti. Deve essere infatti possibile calcolare i quartili delle osservazioni e poter svolgere semplici operazioni di conto. Come primo passo si debbono valutare i quartitili. Il primo quartile è quell'osservazione che lascia alla sua sinistra un quarto delle restanti osservazioni (ovvero $\frac{N-1}{4}=7.25$) poichè il numero non risulta tondo il primo quartile si otterra medianto l'ottava e la nona osservazione. Con una procedura analoga si ottiene che la mediana (secondo quartile) risutlata la media fra la 15ª e la 16ª osservazione mentre il terzo quartile sarà la media fra la 21ª e la 22ª osservazione. Si ha:

$$q_1=\frac{o_8+o_9}{2}=\frac{148+148}{2}=148 \qquad q_2=\frac{o_{15}+o_{16}}{2}=\frac{149+150}{2}=149.5$$
$$q_3=\frac{o_{21}+o_{22}}{2}=\frac{152+152}{2}=152$$



Per poter tracciare il box-plot si devono identificare gli estremi dei due "baffi" che completano il boxplot. Il baffo inferiore viene delimitato dal massimo fra il valore adiacente inferiore (VAI) e la minima osservazione ($o_1=140$); mentre il baffo superiore viene delimitato dal massimo fra il valore adiacente inferiore (VAI) e la massima osservazione ($o_N=170$). Posto la costante $k=1.5$ si ha che:

$$VAI=q_1-1.5*(q_3-q_1)=142 \qquad VAS=q_3+1.5*(q_3-q_1)=158$$

Da cui si ricava agevolemte diagramma a lato in cui si nota la presenza di alcuni outliers.

c*) Se possibile, calcoli la varianza.*

Lo scato quadratico medio (σ) può essere calcolato per ogni carattere quantitativo, e si ha che

$$\sigma^2=\left(\sum_{i=1}^{M} f_i*m_i^2\right)-\bar{o}^2=22536-22500=36 \quad \text{da cui} \quad \sigma=\sqrt{36}=6$$

I conti sono stati svolti nella tabella in calce.

i	m_i	n_i	f_i	$m_i f_i$	m_i^2	$m_i^2 f_i$	$m_i-\bar{x}$	$(m_i-\bar{x})^3$	$(m_i-\bar{x})^3 f_i$
1	140	3	0.100	14	19600	1960	-10	-1000	-100.000
2	143	1	0.033	4.77	20449	681.63	-7	-343	-11.433
3	145	2	0.067	9.67	21025	1401.67	-5	-125	-8.333
4	147	1	0.033	4.9	21609	720.3	-3	-27	-0.900
5	148	5	0.167	24.67	21904	3650.67	-2	-8	-1.333
6	149	3	0.100	14.9	22201	2220.1	-1	-1	-0.100
7	150	1	0.033	5	22500	750	0	0	0.000
8	151	4	0.133	20.13	22801	3040.13	1	1	0.133
9	152	5	0.167	25.33	23104	3850.67	2	8	1.333
10	155	1	0.033	5.17	24025	800.83	5	125	4.167
11	156	1	0.033	5.2	24336	811.2	6	216	7.200
12	158	1	0.033	5.27	24964	832.13	8	512	17.067
13	160	1	0.033	5.33	25600	853.33	10	1000	33.333
14	170	1	0.033	5.67	28900	963.33	20	8000	266.667
Totali		30	1	150	--	22536	--	--	207.8

d) Se possibile, calcoli un indice di asimmetria.
Un indice di asimmetria per caratteri quantiativi è il momento centrale terzo standardizzato

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\sum_{i=1}^{M} f_i (m_i - \bar{o})^3}{\sigma^3} = \frac{207.8}{6^3} = 0.926$$

I conti sono stati svolti nella tabella riportata in precedenza.

Esercizio 2)
a) Completi la tabella con i dati mancanti.

La tabella si completa tenendo conto che la somma delle colonne deve coincidere con i dati illustrati nell'esercizio 1. Si noti che nella nuova formulazione le osservazioni sono state aggregate in classi. Le frequenze assolute richieste sono riportate nella tabella seguente (numeri non tra parentesi).

		Y: concentrazione di glucosio mg/dl						
		fino a 142	da 143 a 149 (149 escluso)	da 149 a 152 (152 esluso)	da 152 a 155 (155 escluso)	da 155 a 160 (160 escluso)	160 ed oltre	
X: Genere	M	3 (1.8)	6 (5.4)	4 (4.8)	**4 (3)**	**1 (1.8)**	**0 (1.2)**	**18**
	F	**0 (1.2)**	**3 (3.6)**	**4 (3.2)**	1 (2)	2 (1.2)	2 (0.8)	**12**
		3	**9**	**8**	**5**	**3**	**2**	**30**

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativo non ordinabile ammette un solo indice sintetico di posizione: la moda. La moda di una bi-variata si ottiene valutando la o le modalità della serie corrispondenti alla frequenza (assoluta o relativa) maggiore. Nel caso in esame la frequenza assoluta maggiore è 4 cui corrisponde la modalità (Maschile; da 149 a 152)

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativio non ordinabile non ammette indice sintetici di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

d) Se possibile, verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti, nel caso non fosse possibile indichi una possibile strategia per effettuare il calcolo.

Per verificare se i due caratteri sono indipendenti 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 indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadrato avente gradi di libertà pari a quelli del 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 .

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

le frequenze marginali e quelle teoriche fra parentesi sono state inserite nella tabella riportata al punto a).

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 non è verificata si può concludere che non è possibile ricevare l'informazione richiesta dalle osservazioni fornite. Un modo per ottenere delle fequenze teoriche superiori a 5 (e quindi poter eseguire il test) è quello di accorpare più classsi nella speranza di ottenere delle frequenze attese meggiori. (si veda la soluzione della seconda fila).

Esercizio 3)

Nel testo si effettuano diverse misure di una grandezza ignota da stimare. Possiamo modellare questo problema come l'estrazione di una variabile casuale

X :*concentrazione di glucosio in un adulto*

avente distribuzione ignota. Si sono effettuate *N = 30* estrazioni in cui si son rilevate *M = 14* modalità

a) stimare puntualmente il valore atteso.

Continuando con il modello precedentemente fatto il punto richiede di stimare *E[X]*. Questa stima può essere effettuata ricordando che la varianza viene stimata correttamente mediante la media campionaria. Il calcolo è gia stato effettuato nello svolgimento del primo esercizio, ottenendo

$$E[X]=\bar{o}=150$$

b) stimare per intervallo del valore atteso.

La stima del valore atteso per intervallo ha come ipotesi che considerando la distribuzione di partenza gaussiana ed *n* grande. Nel caso in esame considerare la distribuzione di partenza gaussiana non introduce un errore elevato (trattasi di errori di misura quindi nello specifico simmetrici) per quanto riguarda la dimensione del campione è possibile ritenere *N = 30* una dimensione sufficiente.

Validate le ipotesi si ha che la stima per intervallo della valor atteso in caso che la varianza della popolazione sia ignota è data dalla

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

ponendo un livello di confidenza del 95 % e ricordando la formula del calcolo della varianza campionaria si ha che:

$$s^2=\sigma^2\frac{n}{n-1}=36\frac{30}{29}=6\sqrt{\frac{30}{29}}\qquad z_{\frac{\alpha}{2}}=1.96$$

Pertanto l'intervallo richiesto è: $E[X]\in\left[150-1.96\frac{6}{\sqrt{30}}\ ;\ 150+1.96\frac{6}{\sqrt{30}}\right]=[152.18\ ;\ 147.82]$

Esercizio 4)

a) E_1 ed E_2

Le due probabilità possono essere calcolate utilzzando la definizione frequentistica, dove gli esiti favorevoli vengono determinati dalle marginali della tebella a doppia etrata dell'esercizio 2.

$$P(E_1) = 12/30 \quad P(E_2) = 5/30$$

b) Il candidato calcoli Probabilità dell'evento E_1 intersezione E_2.

La probabilità dell'evento intersezione di due eventi (ovvero che i due eventi si verifichino entrambi) è ottenibile mediante la definizione frequentistica della probabilità. Si ha infatti ottenuti i casi favorevoli dalla tabella a doppia entrata (casella in posizione 2,4) si ha che

$$P(E_1\cap E_2)=\frac{1}{30}$$

c) Il candidato calcoli Probabilità dell'evento E_2 condizionato E_1

Applicando la definzione di probabilità condizionata si ha che:

$$P(E_2|E_1)=P\frac{(E_1\cap E_2)}{P(E_1)}=\frac{1/30}{5/30}=\frac{1}{5}$$

d) Il candidato calcoli la Probabilità dell'evento E_1 unito E_2..

Note le probabilità degli eventi elementari e dell'evento intersezione si ha che

$$P(E_1\cup E_2)=P(E_1)+P(E_2)-P(E_1\cap E_2)=\frac{12}{30}+\frac{5}{30}-\frac{1}{30}=\frac{16}{30}=\frac{8}{15}$$

- Appello del 24 Giugno 2011 - Fila B

Esercizio 1)
Si vuole stabilire il livello medio di glucosio che un adulto sano presenta nel sangue durante le ore di sonno. Pertanto si è condotta una sperimentazione in cui sono stati coinvolti 10 soggetti. Per ogni soggetto si sono effettuati 3 prelievi (uno ogni due ore) ottenendo le seguenti misurazioni di concetrazione di glucosio espresse in *mg/dl* ed ordinate in maniera crescente.

140	140	140	143	145	145	147	148	148	148
148	148	149	149	149	150	151	151	151	151
152	152	152	152	152	155	156	158	160	170

Il candidato

a) Determini la tipologia del carattere.
b) Se possibile, tracci l'istogramma.
c) Se possibile, calcoli un indice di curtosi adeguato.
d) Se possibile, calcoli la varianza.

Esercizio 2)
I dati raccolti nel precedente esercizio sono stati organizzati tenendo conto del diverso genere del soggetto coinvolto nella sperimentazione, ottenendo la seguente tabella.

		Y: concentrazione di glucosio mg/dl		Marginali
		da 149 a 150 (150 esluso)	da 150 a 170 (170 incluso)	
X: Genere	Maschile	11		
	Femminile			12
Marginali			15	

Il candidato

a) completi la tabella con i dati mancanti.
b) indichi e calcoli, se possibile, un opportuno indice di posizione per la serie bivariata
c) indichi e calcoli, se possibile, un opportuno indice di variabilità per la serie bivariata
d) se possibile, verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti. Nel caso non fosse possibile indichi una possibile strategia per effettuare il calcolo.

Esercizio 3)
Il candidato stimi puntualmente e per intervallo la varianza della concentrazione di glucosio in un adulto basandosi sulle misurazioni di concentrazione riportate nell'Esercizio 1.

Esercizio 4)
Si considerino i due eventi relativi ai dati dell'Esercizio 2

E_1 : estraendo a caso un componente della sperimentazione, questo è un uomo.
E_2 : estraendo a caso una misura di concetrazione di glucosio fra le 30 effettuate durante la sperimentazione, questa è maggiore di 139.

Il candidato valuti calcoli le probabilità dei seguenti eventi

a) le probabilità di E_1 ed E_2
b) la probabilità dell'evento E_1 intersezione E_2
c) se i due eventi sono statisticamente indipendenti

- Appello del 24 Giugno 2011 -
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 una concentrazione che concettualemente è continua).

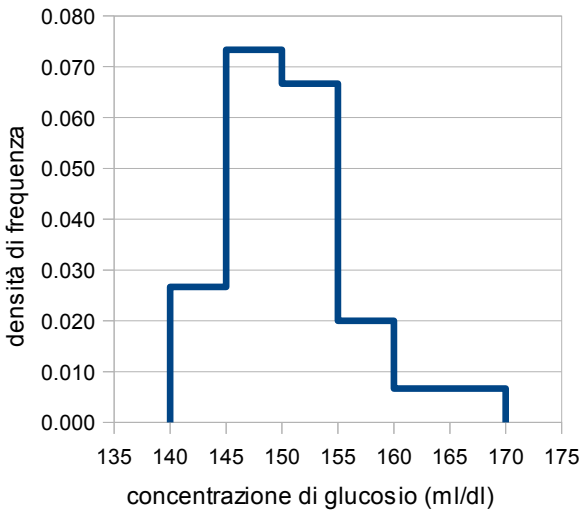
b) Se possibile, tracci si tracci l'istogramma.

L'istogramma è una rappresentazione utilizzata per rappresentare dati quantitativi continui raggruppati in classi. Pertanto per ottenere l'istogramma i dati debbono essere raccolti in classi. Per determinare il numero delle classi C si può utilizzare le seguente formula empirica.

$$C = 1 + \log_2 N = 1 + \log_2 30 = 1 + 4.907 \approx 6$$

Utilizzando classi aventi eguale ampiezza si ha che l'ampiezza di una classe è

$$sup_i - inf_i = \frac{o_N - o_1}{C} = \frac{170 - 140}{6} = 5$$



Procedendo con i conti riportati nella tabella sotto indicata si ha l'istogramma a lato.

i	inf_i	sup_i	sup_i-inf_i	n_i	f_i	f_i /(sup_i-inf_i)
1	140	145	5	4	0.13	0.027
2	145	150	5	11	0.37	0.073
3	150	155	5	10	0.33	0.067
4	155	160	5	3	0.1	0.020
5	160	165	5	1	0.03	0.007
6	165	170	5	1	0.03	0.007
Totali			---	30	1.00	---

N.b. gli estremi superiori sono da riternersi esclusi dalla classe per ogni classe tranne l'ultima per cui è incluso.

c) Se possibile, calcoli la varianza.

La varianza può essere calcolata per ogni carattere quantitativo, quindi anche nel caso in esame e si ha che $\sigma^2 = \left(\sum_{i=1}^{M} f_i * m_i^2\right) - \bar{o}^2 = 22536 - 22500 = 36$ I conti sono stati svoli nella tabella in calce.

i	m_i	n_i	f_i	$m_i f_i$	m_i^2	$m_i^2 f_i$	$m_i - \bar{x}$	$(m_i - \bar{o})^4$	$(m_i - \bar{o})^4 f_i$
1	140	3	0.100	14	19600	1960	-10	10000	1000.000
2	143	1	0.033	4.77	20449	681.63	-7	2401	80.033
3	145	2	0.067	9.67	21025	1401.67	-5	625	41.667
4	147	1	0.033	4.9	21609	720.3	-3	81	2.700
5	148	5	0.167	24.67	21904	3650.67	-2	16	2.667
6	149	3	0.100	14.9	22201	2220.1	-1	1	0.100
7	150	1	0.033	5	22500	750	0	0	0.000
8	151	4	0.133	20.13	22801	3040.13	1	1	0.133
9	152	5	0.167	25.33	23104	3850.67	2	16	2.667
10	155	1	0.033	5.17	24025	800.83	5	625	20.833
11	156	1	0.033	5.2	24336	811.2	6	1296	43.200
12	158	1	0.033	5.27	24964	832.13	8	4096	136.533
13	160	1	0.033	5.33	25600	853.33	10	10000	333.333
14	170	1	0.033	5.67	28900	963.33	20	160000	5333.333
Totali		30	1	150	--	22536	--	--	6997.2

d) Se possibile, calcoli un indice di curtosi.

Un indice di curtosi per caratteri quantiativi è il momento centrale quarto standardizzato

$$\gamma_1 = \frac{\mu_4}{\sigma^4} = \frac{\sum_{i=1}^{M} f_i (m_i - \bar{o})^4}{\left(\sigma^2\right)^2} = \frac{6997.2}{36^2} = 5.40$$

I conti sono stati svolti nella tabella riportata in precedenza.

Esercizio 2)

a) Completi la tabella con i dati mancanti.

La tabella si completa tenendo conto che la somma delle colonne deve coincidere con i dati illustrati nell'esercizio 1. Si noti che nella nuova formulazione le osservazioni sono state aggregate in classi in maniera diversa da quanto fatto nel primo esercizio.

		Y: concentrazione di glucosio mg/dl		Marginali
		da 149 a 150 (150 esluso)	da 150 a 170 (170 incluso)	
X: Genere	Maschile	11 (9)	**7 (9)**	**18**
	Femminile	**4 (6)**	**8 (6)**	12
Marginali		**15**	15	**30**

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativo non ordinabile ammette un solo indice sintetico di posizione: la moda. La moda di una bi-variata si ottiene valutando la o le modalità della serie corrispondenti alla frequenza (assoluta o relativa) maggiore. Nel caso in esame la frequenza assoluta maggiore è 12 cui corrisponde la modalità (Femminile; da 149 a 152)

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativio non ordinabile non ammette indice sintetici di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

d) Se possibile, verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti, nel caso non fosse possibile indichi una possibile strategia per effettuare il calcolo.

Per verificare se i due caratteri sono indipendenti 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 indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadrato avente gradi di libertà pari a quelli del 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 .

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

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 ha la convergenza dello stimatore

$$\hat{n}_{i,j} > 5 \Rightarrow \sum_{i=1}^{M} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} \sim \chi^2\left((M_x - 1)(M_y - 1)\right).$$

Poiché entrambi i caratteri della bi-variata hanno 2 modalità ($M_x = M_y = 2$), la regione di accettazione per un test al 5 % è la seguente.

$$A = \left[0\,;\, \chi^2_{0.95}(1)\right] = [0\,;\, 3.84]$$

Non rimane che da calcolare il valore dello stimatore e verificare se appartiene alla regione di accettazione.

Il valore dello stimatore è

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{i,j}-\hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}}=\frac{(11-9)^2}{9}+\frac{(7-9)^2}{9}+\frac{(4-6)^2}{6}+\frac{(8-6)^2}{6}=\frac{8}{9}+\frac{4}{3}=\frac{20}{9}$$

e risulta interno ad A. Quindi l'ipotesi di indipendenza viene accettata.

Esercizio 3)

Nel testo si effettuano diverse misure di una grandezza ignota da stimare. Possiamo modellare questo problema come l'estrazione di una variabile casuale

X :*concentrazione di glucosio in un adulto*

avente distribuzione ignota. Si sono effettuate $N = 30$ estrazioni in cui si son rilevate $M = 14$ modalità

a) stimare puntualmente la varianza.
Continuando con il modello precedentemente fatto il punto richiede di stimare Var[X]. Questa stima può essere effettuata ricordando che la varianza viene stimata correttamente mediante la varianza campionaria. Ricordando il legame fra la varianza di un campione e la varianza campionaria si ha che

$$s^2=\sigma^2\frac{N}{N-1}=36\frac{30}{29}=37.24$$

(si ricorda che la varianza è stata calcolata nel primo esercizio)

b) stimare per intervallo del valore atteso.
La stima del valore atteso per intervallo ha come ipotesi che considerando la distribuzione di partenza gaussiana ed n grande. Nel caso in esame considerare la distribuzione di partenza gaussiana è un ipotesi un po forte ma legittima mentre per quanto riguarda la dimensione del campione è possibile ritenere $N = 30$ una dimensione sufficiente.

Validate le ipotesi si ha che la stima per intervallo della varianza è data dalla

$$Var[X]\in\left[\frac{(N-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(N-1)};\frac{(N-1)s^2}{\chi^2_{\frac{\alpha}{2}}(N-1)}\right]$$

ponendo un livello di confidenza del *95 %* si ha che:

$$Var[X]\in\left[\frac{(29)36\frac{30}{29}}{47.5};\frac{(29)36\frac{30}{29}}{16.0}\right]=\left[\frac{1080}{47.5};\frac{1080}{16.0}\right]=[22.74;67.5]$$

Esercizio 4)

a) E_1 ed E_2

Le due probabilità possono essere calcolate utilzzando la definizione frequentistica, dove gli esiti favorevoli vengono determinati dalle marginali della tebella a doppia etrata dell'esercizio 2.

$$P(E_1) = 18/30 = 1/5 = 0.2 \quad P(E_1) = 30/30$$

Si nota come l'evento sia l'evento certo.

b) Il candidato calcoli Probabilità dell'evento E_1 intersezione E_2.

La probabilità dell'evento intersezione di un evento con l'evento certo coincide con la probabilità del'evento non certo

$$P(E_1\cap E_2)=\frac{1}{5}$$

c) Il candidato valuti se i due eventi sono indipendenti
Avere informazioni su un qualunque evento non può influenzare la probabilità dell'evento certo (esso si verifica sempre)... quindi i due eventi sono per definizioni indipendenti.

- Appello del 08 Luglio 2011 -

Esercizio 1)
Uno studio di analisi veterinarie vuole monitare il livello di soddifazione dei pazienti riguardo al trattamento ricevuto. Questo viene espresso mediante cinque possibili valutazioni Insufficiente (I), Scarso (S), Accettabile (A), Discreto (D) e Buono (B). Dopo due settimane si sono ottenute le seguenti valutazioni

A	D	I	S	B	A	I	I	S	A
S	D	B	S	A	B	D	A	S	I
B	I	D	A	A	D	I	S	D	A
D	B	S	I	A	I	D	B	A	D

Il candidato

a) determini la tipologia del carattere.
b) illustri la serie utilizzando una rappresentazione grafica opportuna
c) descriva cosa indica l'indice di posizione e, se possibile, ne calcoli uno adeguato alla serie di dati in esame.
d) descriva cosa indicano gli indici di curtosi e, se possibile, ne calcoli uno adeguato alla serie di dati in esame.

Esercizio 2)
Un ricercatore sospetta che la produzione di glucagone da parte del pancreas possa essere stimolata dall'assunzione della vitamina C. Per verificare la sua teoria fa assumere ad un soggetto un preciso quanitativo di vitamina C due ore dopo il pranzo ed osserva la produzione di glucagone dopo 15 minuti dall'assunzione. L'esperimento viene ripetuto per 5 gg ottenendo i seguenti risultati

Giorno	1	2	3	4	5
Vitamina C [mg]	0	10	50	60	80
Glucagone [mg/dl]	1	3	5	6	10

Il candidato

a) descriva il tipo di serie ottenuta e ne fornisca una opportuna rappresentazione grafica.
b) indichi e calcoli, se possibile, un opportuno indice di posizione
c) indichi e calcoli, se possibile, un opportuno indice di variabilità
d) Ipotizzando un legame di tipo lineare,
    1. calcoli l'opportuna regressione
    2. ipotizzi quale sarebbe il livello di glucagone corrispondente ad una ingestione di 92 mg di vitamina C.
    3. indichi, motivando numericamente la risposta, se il legame ipotizzato è attendibile.

Esercizio 3)
Il candidato stimi puntualmente e per intervallo la varianza della concetrazione di glucagone in adulto basandosi sulle misurazioni di concentrazione riportate nell'Esercizio 2.

Esercizio 4)
Si considerino i due eventi incompatibili E_1 ed E_2. Sapendo che $P(E_1) = 0.4$ e $P(E_2) = 50\%$ il candidato calcoli le probabilità dei seguenti eventi:

a) evento E_1 intersezione E_2
b) evento E_2 condizionato E_1
c) evento E_2 unito E_1.

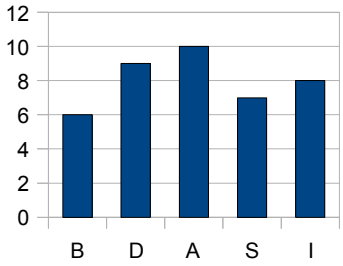
- Appello dell' 8 Luglio 2011 -
Svolgimento

Esercizio 1)

a) Determini la tipologia del carattere.

Il carattere è di tipo qualitativo (in quanto espresso da giudizi e non numeri) ordinabile (in quanto i giudizi possono essere ordinati).

b) illustri la serie utilizzando una rappresentazione grafica opportuna.



Per caratteri qualitativi ordinabili sono possibili due tipologie di rappresentazioni grafiche: il diagramma a barre ed il diagramma a torta.
In questa soluzione si è scelto di rappresentare il primo. Un diagramma a barre è costituito da una serie di barre (orizzontali o verticali) poste in un diagramma cartesiano in cui su di un asse son riportate le modalità del carattere e sull'altro si riportano le frequenze assolute.

c*) descriva cosa indica l'indice di posizione e, se possibile, ne calcoli uno adeguato alla serie di dati in esame..*

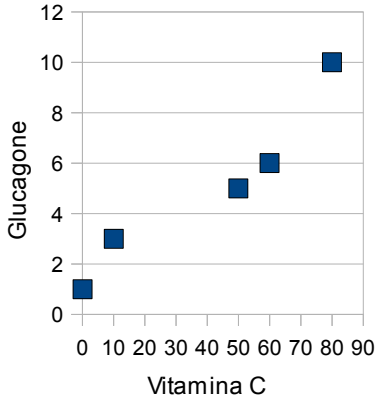
L'indice di posizione di una serie di osservazioni indica il valore centrale che viene assunto dalla serie. Gli indici di posizione visti a lezione sono tre: moda, mediana e media. Per i caratteri in esame sono calcolabili sono i primi due. La moda (ovvero la modalità cui corrisponde la modalità maggiore) della serie è A.

d) descriva cosa indica l'indice di curtosi e, se possibile, ne calcoli uno adeguato alla serie di dati in esame.

Gli indici di curtosi indicano quanto la serie di osservazioni si discosta da una distribuzione normale avente stessa stessa media e varianza della serie di osservazioni. Gli indici di curtosi sono calcolabili solo per caratteri quantitativi.

Esercizio 2)

a) descriva il tipo di serie ottenuta e se ne fornisca una opportuna rappresentazione grafica.



La serie descritta è una serie bivariata composta dai seguenti caratteri
X: vitamina C fornita dopo il pasto.
Y: glucagone misurato nell'individuo dopo 15 min dal pasto.
Una opportuna rappresentazione per una bivariata in cui tutti i caratteri sono di tipo quantitativo è il dia gramma a dispersione riportato a lato

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

Una serie bivariata ottenuta misurando due caratteri quantitativi ammette due indici di posizione la media e la moda. Poichè la seconda non fornisce informazioni (non vi sono frequenze assolute maggiori di uno) si calcola la media che viene calcolata come il vettore delle medie dei due caratteri. Si ha quindi che

$$\bar{x}=\frac{1}{N}\sum_{1=1}^{N} x_i=\frac{200}{5}=40 \qquad \bar{y}=\frac{1}{N}\sum_{1=1}^{N} y_i=\frac{25}{5}=5$$

Da cui la media richiestà è il punto (40 ; 5).

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

Una serie bivariata composta da due indici carattere quantiativi ammette un indice sintetico di variabilità: la matrice varianza/covarianza. Questa matrice si compone della varianze dei due caratteri e della covarianza della serie. Impostati i conti nella forma tabellare seguente

i	x_i	y_i	x_i^2	y_i^2	$x_i-\bar{x}$	$y_i-\bar{y}$	$(y_i-\bar{y})(x_i-\bar{x})$
1	0	1	0	1	-40.000	-4.000	160
2	10	3	100	9	-30.000	-2.000	60
3	50	5	2500	25	10.000	0.000	0
4	60	6	3600	36	20.000	1.000	20
5	80	10	6400	100	40.000	5.000	200
	200	25	12600	171			440.000

si possono ricavare i tre indici richiesti

$$\sigma_x^2=\left(\frac{1}{N}\sum_{1=1}^{N}x_i^2\right)-\overline{x}^2=\frac{12600}{5}-40^2=920 \qquad \sigma_y^2=\left(\frac{1}{N}\sum_{1=1}^{N}y_i^2\right)-\overline{y}^2=\frac{171}{5}-5^2=9.2$$

$$\sigma_{xy}=\frac{1}{N}\sum_{1=1}^{N}(x_i-\overline{x})(y_i-\overline{y})=\frac{440}{5}=88$$

Da cui si ricava la seguente matrice varianza/covaranza

$$\Sigma=\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}=\begin{bmatrix}920 & 88\\ 88 & 9.2\end{bmatrix}$$

d.1) Ipotizzando un legame di tipo lineare, calcoli l'opportuna regressione
Ipotizzando una modello di tipo lineare y = ax +b si dimostra che i parametri hanno i seguenti valori

$$a=\frac{\sigma_{xy}}{\sigma_x^2}=\frac{88}{920} \qquad b=\overline{y}-a\,\overline{x}=5-\frac{88}{920}40=5-\frac{88}{23}=\frac{(115-88)}{23}=\frac{27}{23} \quad \text{da cui} \quad \hat{y}=\frac{88}{920}x+\frac{27}{23}$$

d.2) Ipotizzando un legame di tipo lineare, ipotizzi quale sarebbe il livello di glucagone corrispondente ad una ingestione di 92 mg di vitamina C.
In questo caso e sufficiente applicare la retta di regressione nel punto in cui si vuole ottenere la stima. Si ha quindi

$$\hat{y}=\frac{88}{920}92+\frac{27}{23}=\frac{88}{10}+\frac{27}{23}=9.973$$

d.3) Ipotizzando un legame di tipo lineare, indichi, motivando numericamente la risposta, se il legame ipotizzato è attendibile.
Una buona stima della bontà dell'approssimazioneè è data dal coefficiente di correlazione lineare di Pearson

$$R^2=\frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}=\frac{88^2}{920\cdot 9.2}=0.915$$

Un valore prossimo ad 1 del coefficiente di correlazione linare del Pearson indica una buona approssimazione del modello adottato.

Esercizio 3)

Nel testo si effettuano diverse misure di una grandezza ignota da stimare. Possiamo modellare questo problema come l'estrazione di una variabile casuale

Y :*concentrazione di glucagone in un adulto*

avente distribuzione ignota. Si sono effettuate *N = 5* realizzazioni della v.c.

a) stimare puntualmente la varianza.
Continuando con il modello precedentemente fatto il testo richiede di stimare Var[Y] puntualmente. Questa stima può essere effettuata ricordando che la varianza viene stimata correttamente mediante la varianza campionaria s^2. Sfruttando i conti svolti nel precedente esercizio, e ricordando la formula della varianza campionaria otteniamo che

$$Var[Y]=s^2=\frac{\sigma_y^2}{N-1}N=\frac{9.2}{4}5=11.5$$

b) stimare per intervallo la varianza.

Nei termini del modello precedentemente illustrato il testo richiede di stimare Var[Y] per intervallo. Questa stima può essere effettuata ricordando lo stimatore varianza campionaria al crescere della dimensione del campione si tende a distribuirsi come un $\frac{Var[Y]}{N-1}\chi^2(N-1)$. Pertanto per si ha che la stima voluta ad un livello di confidenza α è data dal seguente intervallo

$$Var[Y]\in\left[\frac{(N-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(N-1)};\frac{(N-1)s^2}{\chi^2_{\frac{\alpha}{2}}(N-1)}\right]$$

Considerare *N = 5* un valore sufficiente per la convergenza dello stimatore è un ipotesi molto forte, quindi si considererebbe valida l'asserzione che mancano dati per fare la stima richiesta. In ogni caso, se si ritenesse la dimensione del campione adeguata si avrebbe la seguente stima al 95%

$$Var[Y]\in\left[\frac{4\cdot 11.5}{\chi^2_{0.975}(4)};\frac{4\cdot 11.5}{\chi^2_{0.025}(4)}\right]=\left[\frac{47}{11.1};\frac{47}{0.484}\right]=[4.234;97.107]$$

Esercizio 4)
I due eventi considerati sono incompatibili il che vuol dire che non possono verificarsi contemporaneamente. In termini probabilistici questo implica che:

$$P(E_1\cap E_2)=0$$

a) Il candidato calcoli Probabilità dell'evento E_1 intersezione E_2.
Vedi definizione iniziale

b) Il candidato calcoli Probabilità dell'evento E_2 condizionato E_1
Poiché i due eventi sono incompatibili non è possibile che si verifichi un evanto sapendo se che i è verificato l'altro. In ogni caso lo stesso risultati poteva essere ottenuto utilizzando la definzione di probabilità condizionata:

$$P(E_2|E_1)=\frac{P(E_1\cap E_2)}{P(E_1)}=\frac{0}{0.4}=0$$

c) Il candidato calcoli la Probabilità dell'evento E_1 unito E_2..
La probabilità dell'evento unione di due eventi indipendenti è pari alla somma delle probabilità (90 %). Lo stesso risultato può essere ottenuto applicando la definzione assiomatica di probabilità

$$P(E_1\cup E_2)=P(E_1)+P(E_2)-P(E_1\cap E_2)=0.4+0.5-0=0.9$$

- Appello del 09 Settembre 2011 -

Esercizio 1)
Un laboratorio farmaceuto vuole verificare l'affidabilità dei suoi fornitori di acqua borica. A questo scopo ha effettuato diverse misure di concentrazione del boro presente nelle diverse boccette, ottenendo i seguenti 11 dati espressi in punti percentuali

2.9500 3.0300 3.0000 3.0900 3.0400 2.9200 2.9600 3.0300 3.0600 3.0200 2.9000

Il candidato

a) Determini la tipologia del carattere.
b) Scelga una rappresentazione grafica idonea.
c) Definisca gli indici di posizione e, se possibile, ne calcoli uno adeguato alla serie.
d) Definisca gli indici di variabilità e, se possibile, ne calcoli uno adeguato alla serie.

Esercizio 2)
Si vuole verificare l'efficacia di un nuovo antibiotico chemiterapico utilizzare contro l'infezione dovuta ad un particolare batterio. A tale scopo si sono prese 100 cavie da laboratorio, le si sono infettate con il suddetto batterio ma solo metà di esse è stata sottoposta a trattamento con il nuovo farmaco. Dopo due settimane si è osservato l'estensione dell'infezione ottenendo le seguente tabella a doppia entrata:

		Y: estensione dell'infezione			
		Contenuta	Media	Estesa	
X: Trattamento	Non applicato		10		
	Applicato	10			
		15		60	

Il candidato

a) completi la tabella con i dati mancanti.
b) indichi e calcoli, se possibile, un opportuno indice di posizione per la serie bivariata
c) indichi e calcoli, se possibile, un opportuno indice di variabilità per la serie bivariata
d) se possibile, verifichi, ad un opportuno livello di significatività, l'efficacia del nuovo farmaco. Nel caso non fosse possibile indichi una possibile strategia per effettuare il calcolo.

Esercizio 3)
Si consideri la statitistca relativa all'acqua borica analizzata nell'Esercizio 1. Il candidato

a) verifichi, se possibile, che ad una significatività del 90% l'acqua borica abbia concentrazione pari al 3%;
b) stimi puntualmente la varianza della concentrazione dell'acqua borica fornita.

Esercizio 4)
Si considerino i due eventi relativi ai dati dell'Esercizio 2

E_1 : estraendo a caso una cavia, questa sia stata trattata con il nuovo farmaco.
E_2 : estraendo a caso una cavia della sperimentazione, questa abbia un'infezione contenuta.

Il candidato calcoli le probabilità dei seguenti eventi

a) E_1 ed E_2
b) evento E_1 intersezione E_2
c) evento E_2 condizionato E_1
d) evento E_2 unito E_1.

- Appello del 9 Settembre 2011 -
Svolgimento

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 una concentrazione che concettualemente è continua).

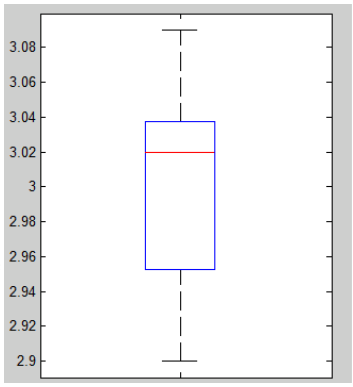
b) Se possibile, tracci una rappresentazione adeguata.

Poichè non vi sono molti dati ripetuti, ed essi sono in numero superiore alla decina (*N=11*) una buona rappresentazione è data dal box-plot. Il box-plot è una rappresentazione grafica utile per rappresentare dati quantitativi siano essi continui o discreti. Deve essere infatti possibile calcolare i quartili delle osservazioni e poter svolgere semplici operazioni di conto. Come primo passo si debbono ordinare le osservazioni

2.90 2.92 2.95 2.96 3.00 3.02 3.03 3.03 3.04 3.06 3.09

e valutare i quartitili. Il primo quartile è quell'osservazione che lascia alla sua sinistra un quarto delle restanti osservazioni (ovvero $\frac{N-1}{4}=2.5$) poichè il numero non risulta tondo il primo quartile si otterra medianto la terza e la quarta osservazione ordinate; con una procedura analoga si ottiene che il terzo quartile sarà la media fra la 8ª e la 9ª osservazione. Mentra la mediana risulta essere il valore che ha alla sua sinistra $\frac{N-1}{2}=5$ osservazioni. Pertanto si ha:

$$q_1=\frac{o_3+o_4}{2}=\frac{2.95+2.96}{2}=2.955 \qquad q_2=o_6=3.02$$
$$q_3=\frac{o_8+o_9}{2}=\frac{3.03+3.04}{2}=3.035$$



Per poter tracciare il box-plot si devono identificare gli estremi dei due "baffi" che completano il boxplot. Il baffo inferiore viene delimitato dal massimo fra il valore adiacente inferiore (VAI) e la minima osservazione ($o_1=2.90$); mentre il baffo superiore viene delimitato dal massimo fra il valore adiacente inferiore (VAI) e la massima osservazione ($o_N=3.09$). Posto la costante $k=1.5$ si ha che:

$$VAI=q_1-1.5*(q_3-q_1)=2.835 \qquad VAS=q_3+1.5*(q_3-q_1)=3.155$$

Da cui si ricava agevolemte diagramma a lato in cui si nota l'assenza di outliers.

c*) Definisca gli indici di posizione e, se possibile, ne calcoli uno adeguato alla serie..*

Gli indici di posizione calcolabili sono tre: la media (ricavata dalla sequente formula $\bar{o}=\frac{1}{N}\sum_{i=1}^{N} o_i=3$), la mediana introdotta al punto precedente e la moda (ovvero l'osservazione con frequenza più elevata).
Per questa serie l'unico indice che non pare adeguato è la moda che , pur essendo unica e calcolabile riferisce l'unica osservazione ripetuta della serie di dati (3.03)

d) Definisca gli indici di variabilità e, se possibile, ne calcoli uno adeguato alla serie.

Gli indici di variabilità indicano quanto le osservazioni si discostano dal valor centrale. Essi sono tutti validi per la serie in esame e sono i seguenti.

- Il campo di variazione: l'osservazione massima meno l'osservazione minima e vale $o_{11}-o_1=0.19$
- La distanza interquartile: la differenze fra il terzo ed il primo quartile $q_3-q_1=0.08$
- Lo scato quadratico medio (σ): radice quadrata della media degli scarti dalla media

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}\left(o_i-\bar{o}\right)^2=\frac{0.036}{11} \quad \text{da cui} \quad \sigma=\sqrt{\frac{0.036}{11}}=0.0572$$

I conti sono stati svolti nella tabella in calce.

o_i	2.9000	2.9200	2.9500	2.9600	3.0000	3.0200	3.0300	3.0300	3.0400	3.0600	3.0900
$o_i - \bar{o}$	-0.1000	-0.0800	-0.0500	-0.0400	0.0000	0.0200	0.0300	0.0300	0.0400	0.0600	0.0900
$(o_i - \bar{o})^2$	0.0100	0.0064	0.0025	0.0016	0.0000	0.0004	0.0009	0.0009	0.0016	0.0036	0.0081

Esercizio 2)

a) Completi la tabella con i dati mancanti.

La tabella si completa tenendo conto che la somma delle righe deve essere pari a 50 (metà delle cavie). Le frequenze assolute teoriche sono riportate nella tabella seguente (numeri non tra parentesi).

		Y: estensione dell'infezione			
		Contenuta	Media	Estesa	
X: Trattamento	Non Applicato	**5** *(7.5)*	10 *(12.5)*	**35** *(30)*	**50**
	Applicato	10 *(7.5)*	**15** *(12.5)*	**25** *(30)*	**50**
		15	**25**	60	**100**

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativo non ordinabile ammette un solo indice sintetico di posizione: la moda. La moda di una bi-variata si ottiene valutando la o le modalità della serie corrispondenti alla frequenza (assoluta o relativa) maggiore. Nel caso in esame la frequenza assoluta maggiore è 35 cui corrisponde la modalità (Non applicato; Estesa)

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativio non ordinabile non ammette indice sintetici di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

d) Se possibile, verifichi, ad un opportuno livello di significatività, l'efficacia del nuovo farmaco. Nel caso non fosse possibile indichi una possibile strategia per effettuare il calcolo.

Un requisito di base richiesto dal nuovo farmaco è che la sua applicazione influisca sul carattere relativo alla diffusione del contagio ovvero che i due caratteri in esame siano dipendenti. Per verificare se i due caratteri sono indipendenti 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 indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadrato avente gradi di libertà pari a quelli del 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 .

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

le frequenze marginali e quelle teoriche fra parentesi sono state inserite nella tabella riportata al punto a).
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 ha che.

$$\sum_{i=1}^{M} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} \sim \chi^2\left((M_x - 1)(M_y - 1)\right)$$

Poichè entrambi i caratteri della bivarianta hanno $M_x = 2$ e $M_y = 3$, la regione di accettazione per un test al 5 % è la seguente.

$$A = \left[0\,;\, \chi^2_{0.95}(2)\right] = [0\,;5.99]$$

Non rimane che da calcolare il valore dello stimatore e verificare se appartiene alla regione di accettazione.
Il valore dello stimatore è

$$\sum_{i=1}^{6} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} = \frac{(5-7.5)^2}{7.5} + \frac{(10-12.5)^2}{12.5} + \frac{(35-30)^2}{30} + \frac{(10-7.5)^2}{7.5} + \frac{(15-12.5)^2}{12.5} + \frac{(15-12.5)^2}{12.5} + \frac{(25-30)^2}{30} = \frac{13}{3}$$

e risulta interno ad *A*. Quindi l'ipotesi di indipendenza viene accettata. Pertanto il farmaco è inefficace.

Esercizio 3)

Nel risolvere questo esercizio si ipotizza che la concnetrazione di boro, espressa in punti percentuali, presente nell'acqua fornita sia modellabile mediante una V.C. *X* da cui sono state effettuate diverse estrazioni i.i.d. .

a) verifichi, se possibile, che ad una significatività del 90% l'acqua borica abbia concentrazione pari al 3%

Continuando con il modello precedentemente definito il punto richiede di verificare l'ipotesi che che E[X]=3. La verifica di quest'ipotesi può essere realizzata, utilizzando gli strumenti forniti nel corso, solo se la dimensione del campione è pari a 30. Pertanto non è possibile verificare quest'ipotesi avendo a disposizione solo 11 valori.

b) stimi puntualmente la varianza della concentrazione dell'acqua borica fornita

Una stima puntale corretta della varianza di una V.C. mediante osservazioni i.i.d. può essere ottenuta mediante la varianza campionaria s2 ottenuta mediante la seguente formula.

$$s^2 = \sigma^2 \frac{n}{n-1}$$

quindi ricordando quando ricavato nell'esercizio 1 si ha che

$$s^2 = \sigma^2 \frac{n}{n-1} = \frac{0.036}{11}\frac{11}{10} = 0.0036$$

Pertanto la stima richiest è *Var*[*X*]*= 0.0036*

Esercizio 4)

a) E_1 ed E_2

Le due probabilità possono essere calcolate utilzzando la definizione frequentistica, dove gli esiti favorevoli vengono determinati dalle marginali della tebella a doppia etrata dell'Esercizio 2.

$$P(E_1) = 50/100 = 0.5 \quad P(E_2) = 15/100 = 0.15$$

b) Il candidato calcoli Probabilità dell'evento E_1 intersezione E_2.

La probabilità dell'evento intersezione di due eventi (ovvero che i due eventi si verifichino entrambi) è ottenibile mediante la definizione frequentistica della probabilità. Si ha infatti ottenuti i casi favorevoli dalla tabella a doppia entrata (casella in posizione 2,1) si ha che

$$P(E_1 \cap E_2) = \frac{10}{100} = 0.1$$

c) Il candidato calcoli Probabilità dell'evento E_2 condizionato E_1

Applicando la definzione di probabilità condizionata si ha che:

$$P(E_2|E_1) = P\frac{(E_1 \cap E_2)}{P(E_1)} = \frac{1/10}{1/2} = \frac{5}{100} = 0.05$$

d) Il candidato calcoli la Probabilità dell'evento E_1 unito E_2..

Note le probabilità degli eventi elementari e dell'evento intersezione si ha che

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{50}{100} + \frac{15}{100} - \frac{10}{100} = \frac{55}{100} = 0.55$$

- Appello del 23 Settembre 2011 -

Esercizio 1)
Si sono rilevati i giudizi relativi alla qualità delle lezioni di un corso di statistica seguito da 30 studenti, ottenendo le seguenti rilevazioni

Modalità (giudizio)	Ottimo	Buono	Discerto	Sufficiente	Insufficiente	Gravemente insufficiente
Frequenza assoluta	3	5	9	6	4	3

Il candidato

a) Determini la tipologia del carattere.
b) Scelga una rappresentazione grafica idonea.
c) Definisca gli indici di posizione e, se possibile, ne calcoli uno adeguato alla serie.
d) Definisca gli indici di variabilità e, se possibile, ne calcoli uno adeguato alla serie.

Esercizio 2)
I dati descritti nell'Esercizio 1 sono stati caratterizzati considerando il genere dei votanti

		X: giudizi al corso						Tot
		Ottimo	Buono	Discerto	Sufficiente	Insufficiente	Gravemente ins	
Y: genere	Maschi		2	3	3	3		16
	Femmine	1				1		
Tot								

Il candidato

a) completi la tabella con i dati mancanti.
b) indichi e calcoli, se possibile, un opportuno indice di posizione per la serie bivariata
c) indichi e calcoli, se possibile, un opportuno indice di variabilità per la serie bivariata
d) se possibile, verifichi, ad un opportuno livello di significatività, se il genere influenza il voto senza alterare la statistica.

Esercizio 3)
Si consideri la statitistca ricavata da quella descritta nell'Esercizio 1 dove per ogni giudizio sono stati asseganti i seguenti punteggi.

Voto	Ottimo	Buono	Discerto	Sufficiente	Insufficiente	Gravemente insufficiente
Punteggio (p_i)	9	8	7	6	5	4

Il candidato consideri la v.c. *P* le cui osservazioni sono avvenute secondo le frequenze descritte nell'Esercizio 1 e

a) verifichi, se possibile, mediante un test di ipotesi ad una significatività del 90%, se il valor atteso di *P* sia è superiore al 7;
b) stimi puntualmente la varianza di *P*.

Esercizio 4)
Date le sue variabili casuali $X \sim Ber(0.4)$ e $Y \sim Chi(10)$, si considerino i due eventi considerati indipendenti

$$E_1 : X = 0 \qquad E_2 : Y < 3.94$$

Il candidato calcoli le probabilità dei seguenti eventi

a) E_1 ed E_2
b) evento E_1 intersezione E_2
c) evento E_2 condizionato E_1
d) evento E_2 unito E_1.

- Appello del 9 Settembre 2011 -
Svolgimento

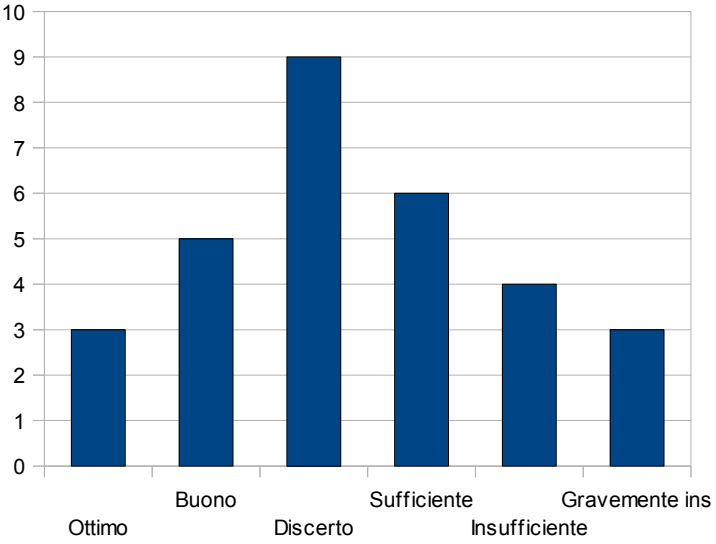
Esercizio 1)

a) Determini la tipologia del carattere.

Il carattere è di tipo qualitativo (in quanto espresso da giudizio) ordinabile (in quanto è possibile dare un ordine alle modalità).

b) Se possibile, tracci una rappresentazione adeguata.

Essendo il dato qualitativo solo alcune rappresentazioni grafiche sono possibili. Una rappresentazione efficace per dati ordinabili è il diagramma a barre. Questo diagramma si rappresenta inserendo dei rettangoli (barre) nel primo quadrante di un piano catesiano in cui nell'asse delle ascisse sono riportate le modalità mentre in quello dello ordinate le frequenze (relative o assolute). I diversi rettangoli sono posti in corrispondenza delle varie modalità ottenendo un diagramma come quello riportato a lato.



c*) Definisca gli indici di posizione e, se possibile, ne calcoli uno adeguato alla serie..*

Gli indici di posizione sono tre: la media (ricavata dalla sequente formula $\bar{o}=\frac{1}{N}\sum_{i=1}^{N} o_i=3$), la mediana (ovvero l'osservazione che bipartisce le osservazioni ordinate) e la moda (ovvero l'osservazione con frequenza più elevata).

Per questa serie l'unico indice non calcolabile è la media. Gli altri indici sono

- Moda = Disceto (la frequenza assoluta è la maggiore 9)
- Mediana = Discerto (Poiche vi sono 30 ossevazioni la mediana è la media fra la *15ª* e la *16ª* osservazione. Poichè esse coincidono con la modalità "Discreto" esse coincidono con la mediana.)

d) Definisca gli indici di variabilità e, se possibile, ne calcoli uno adeguato alla serie.

Gli indici di variabilità indicano quanto le osservazioni si discostino dal valor centrale.

Nel corso sono stati illustrati i seguenti indici.

- Il campo di variazione: l'osservazione massima meno l'osservazione minima
- La distanza interquartile: la differenze fra il terzo ed il primo quartile
- Lo scato quadratico medio (σ): radice quadrata della media degli scarti dalla media

In questo caso nessun indice è calcolabile.

Esercizio 2)

a) Completi la tabella con i dati mancanti.

La tabella si completa tenendo conto dei totali di colonna sono dati all punto 1e che il numero totale della osservazioni deve essere 30. Le frequenze assolute teoriche sono riportate nella tabella seguente (numeri non tra parentesi).

		X: giudizi al corso						Tot
		Ottimo	Buono	Discerto	Sufficiente	Insufficiente	Gravemente ins	
Y: genere	Maschile	2 (8/5)	2 (8/3)	3 (24/5)	3 (16/5)	3 (36/15)	3 (8/5)	16
	Femminile	1 (7/5)	3 (7/3)	6 (21/5)	3 (14/5)	1 (28/15)	0 (8/5)	14
Tot		3	5	9	6	4	3	30

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativo non ordinabile ammette un solo indice sintetico di posizione: la moda. La moda di una bi-variata si ottiene valutando la o le modalità della serie corrispondenti alla frequenza (assoluta o relativa) maggiore. Nel caso in esame la frequenza assoluta maggiore è 6 cui corrisponde la modalità (Discreto; Femminile)

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativio non ordinabile non ammette indice sintetici di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

d) Se possibile, verifichi, ad un opportuno livello di significatività, se il genere influenza il voto senza alterare la statistica.

Nel caso il genere non influenzi il voto, vorrebbe dire che le due variabili sono scorrelate, pertanto un buom modo per vedere se vi sia un legame fra le variabili e verificarne la dipendenza mediante un test di indipendenza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadrato avente gradi di libertà pari a quelli del 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 .

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

le frequenze teoriche sono state inserite fra parentesi nella tabella riportata al punto a).

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 questa non è verificata si può conludere che non è possibile verificare l'indipendenza senza alterare la statistica.

Esercizio 3)

Nel risolvere questo esercizio si ipotizza che il voto espresso dagli studenti sia modellabile mediante una V.C. X da cui sono state effettuate diverse estrazioni i.i.d. .

a) verifichi, se possibile, mediante un test di ipotesi ad una significatività del 90%, se il valor atteso di P sia è superiore al 7;

Il punto richiede di verificare l'ipotesi alternativa H_1: E[P]>7 contro l'ipotesi H_0: E[P]=7. La verifica di quest'ipotesi può essere realizzata, utilizzando gli strumenti forniti nel corso, solo se la dimensione del campione è pari o superiore a 30. Verificata questa ipotesi si procede al calcolo della media e della varianza campionaria (stimatori puntuali rispettiavemente di valore atteso e varianza). I conti sono riporati nella tabella seguente

$$\bar{p} = \sum_{i=1}^{6} f_i\, p_i = 6.6 \qquad s^2 = \sigma^2 \frac{n}{n-1} = \frac{n}{n-1}\left(\sum_{i=1}^{6} f_i\, p_i^2 - \bar{p}^2\right) = \frac{30}{29}\left(45.6 - 6.6^2\right) = 2.11$$

p_i	9	8	7	6	5	4	Tot
n_i	3	5	9	6	4	3	30
f_i	1/10	1/6	3/10	1/5	2/15	1/10	
$v_i f_i$	3/10	4/3	21/10	6/5	2/3	2/5	6.6
p^2_i	81	64	49	36	25	16	
$p^2_i f_i$	81/10	20/3	147/10	36/5	8/3	6/5	45.6

Noti questi valori è possibile standardizzare il valore di riferimento (7) ottenendo

$$z_7 = \frac{7-\bar{p}}{\sqrt{\frac{s^2}{n}}} = \frac{7-6.6}{\sqrt{\frac{2.11}{30}}} = 1.5$$

Questo valore va confrontato con la ragione di accettazione del test. Trattandosi di un test ad una coda con un intervallo di signinficatività pari al 90% essa è pari a $A = [-\infty; 1.28]$. Poiche il valore è esterno alla regione di accettazione possiamo considerare buona l'ipotesi alternativa.

b) stimi puntualmente la varianza dei P
Una stima puntale corretta della varianza di una V.C. mediante osservazioni i.i.d. può essere ottenuta mediante la varianza campionaria s^2 ottenuta al punto precedente. Pertanto la stima richiest è $Var[P]= 2.11$

Esercizio 4)

a) E_1 ed E_2

L'evento E_1 si verifica quando una v.c. Bernoulliana avente p=0.4 vale 0. Poichè il parametro p indica la probabilità che la v.c. ha valore 1 si ha che

$$P(E_1) = 1 - P(\overline{E}_1) = 1- P(X=1) = 1- p = 0.6$$

La probabilità dell'evento E2 è ottenibile direttamente dalle tavole dei Chi quadrato. Infatti condsiderando un Chi quadrato a 10 gradi di libertà si ha che in corrispondenza del valore 3.94 l'area sottesa dalla curva della d.d.p. vale 0.05. Pertanto si ha che

$$P(E_2)=P(Y<3.94)=\int_0^{3.94} f(y)dy=0.05$$

b) Il candidato calcoli Probabilità dell'evento E_1 intersezione E_2.

La probabilità dell'evento intersezione di due eventi (ovvero che i due eventi si verifichino entrambi) indipendenti è pari al prodotto delle probabilità.

$$P(E_1\cap E_2)=P(E_1)P(E_2)=0.6*0.05=0.03$$

c) Il candidato calcoli Probabilità dell'evento E_2 condizionato E_1

Applicando la definzione di probabilità condizionata si ha che:

$$P(E_2|E_1)=P\frac{(E_1\cap E_2)}{P(E_1)}=\frac{0.03}{0.6}=0.05$$

d) Il candidato calcoli la Probabilità dell'evento E_1 unito E_2..
Note le probabilità degli eventi elementari e dell'evento intersezione si ha che

$$P(E_1\cup E_2)=P(E_1)+P(E_2)-P(E_1\cap E_2)=0.6+0.05-0.03=0.62$$

- Appello del 8 Febbraio 2012 -

Esercizio 1)
Una grossa azienda vuole monitorare il numero di permessi richiesti in un anno da un singolo dipendete. La seguente tabella riporta le frequenze assolute delle richieste annue dei permessi nell'intero anno 2011 da parte di tutto il personale.

Permessi richiesti	1	2	3	4	5	6	7	8	9
Frequenza	80	83	87	77	90	183	180	90	30

Determinare

a) La tipologia del carattere.
b) Un indice sintetico di posizione.
c) Se possibile, un indice sintetico di variabilità.
d) Una rappresentazione grafica adeguata.
e) L'eventuale presenza di outlier.

Esercizio 2)
Le osservazioni descritte nell'esercizio precedente sono state catalogate in base allo stato civile e raccolte in classi.

		Y:stato civile			
		Nubile/Celibe	Coniugato/a	Vedovo/a	
X: Giorni di assenza	meno di 4		105	65	
	Fra 4 ed 6	100			
	Piu di 6		160		
		270	450		

Il candidato

a) Completi la tabella con i dati mancanti.
b) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di posizione
c) Se possibile, indichi e calcoli per la serie ottenuta un opportuno indice di variabilità
d) Verifichi, ad un opportuno livello di significatività, se i due caratteri si possano dire indipendenti.

Esercizio 3)
Utilizzando i dati raccolti nel primo esercizio come campione, il candidato stimi puntualmente e per intervallo la varianza dei giorni di permesso richiesti in un anno, indicando le opportune ipotesi. Il candidato proceda al calcolo anche se queste risultino non verificate.

Esercizio 4)
Si considerino i seguenti eventi legati all'estrazione di un lavoratore fra tutti quelli descritti nell'Esercizio 2.

E_1 : Si estragga un lavoratore sposato
E_2: Si estragga un lavoratore che ha fatto al massimo 6 giorni di assenza

a) Il candidato calcoli, se possibile, le seguenti Probabilità $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$ $P(E_1 \mid E_2)$.
b) Il candidato indichi se i due eventi E_1 ed E_2 sono statisticamente indipendenti.

- Appello del 8 Febbraio 2012 -
Svolgimento

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 espresse da numeri naturali)

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 (una persona che ha fatto una grave malattia o un incidente...)

Per calcolare la mediana si deve valutare la numerosità della popolazione (N=900) facilmente ottenibile comulando le frequenze assolute

Giorni di assenza	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Tot.**
Frequenza	80	83	87	77	90	183	180	90	30	900
F. ass. cumulata	*80*	*163*	*250*	*327*	*417*	*600*	*780*	*870*	*900*	--

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 = 449,5 elementi. Poichè non esiste l'osservazione di posto 450,5 viene presa come mediana la media fra il 450° ed 451° valore. Analizzando le frequenze cumulate si ottiene che ambo le osservazioni mostrano la modalità 6. Pertanto la mediana (q2) è 6

c) Se possibile, un indice sintetico di variabilità.

Un carattere di tipo quantitativo ammette quattro indici sintetici di varaibilità: il range (o campo di variazione) la distanza interquartile, la varianza e la devizione 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 (q1) edi l terzo (q3) quartile. Dave q1, una volta ordinate le osservazioni, lascia alla propria sinistra (N-1)/4 = 6,75 osservazioni mentre q3 lascia alla proria destra (N-1)/4 = 224,75 osservazioni. Anche in questo caso non ottenendo numeri interi dovremo mediare le posizioni intere più vicine. Si ha dunque che

q1 = media 225° e 226° valore = 3 q3 = media 675° e 676° valore =7

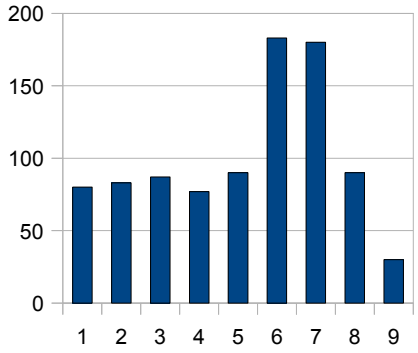
Si ha che la distanza interquartile D = q3 - q1 = 7 - 3 = 4

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 digramma 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 quatile 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

$$VAI = 3\text{-}1^* 4 = -1 \qquad VAS = 7 + 1^*4 = 11$$

Non esistendo alcuna osservazioni esterna all'intervallo [-1 ; 11] possiamo concludere che la popolazione presumibilmente non presenta outlier.

Esercizio 2)

L'esercizio verte sull'analisi di una serie bivariata, ottenuta misurando due caratteri qualitativi non ordinabili.

a) completi la tabella con i dati mancanti

Le marginali del carattere X si ottengono dalla monovarita descrtta nell'esercizio 1; mentre gli altri numeri si ottengono ricordanto che si debbono rispettare i totali di riga e colonna (ovvero le marginali).

		Y:stato civile			Marginali
		Nubile/Celibe	Coniugato/a	Vedovo/a	
X: Giorni di assenza	meno di 4	**80 (75)**	105 (125)	65 (50)	**250**
	Fra 4 ed 6	100 (105)	**185 (175)**	**65 (70)**	**350**
	Piu di 6	**90 (90)**	160 (150)	**50 (60)**	**300**
Marginali		270	450	**180**	**900**

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

Una serie bivariata ottenuta misurando un carattere qualitativo non ordinabili (lo stato civile) ed uno quantitativo discreto ammette un solo indice sintetici 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 185 da cui si ha le seguente moda

(Fra 4 e 6 ; Coniugato/a)

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

Una serie bivariata ottenuta misurando almeno un carattere qualitativo non ammette indice sintetici di variabilità in quanto non è possibile ottenere il concetto di distanza in maniera oggettiva.

c) Verifichi, ad un opportuno livello di significatività, se i due caratteri si possono dire indipendenti.

Per verificare se i due caratteri sono indipiendeti si può effettuare un test di ipotesi volto a verificare se le frequenze delle osservazioni rilevate nel campione sono sufficiente mente vicine (ad un determinato livello di significatività) a quelle teoriche ottenute dall'ipotesi di indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadranto avente gradi di libertà paria quelli del 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 .

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

le frequenze teoriche sono state riportate nella tabella al punto precedente fra parentesi. A questo punto è possibile valutare la convergenza dell 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%.

$$A = \left[0\,; \chi^2_{1-\alpha}\left((M_x - 1)(M_y - 1)\right)\right] = \left[0\,; \chi^2_{1-0.05}\left((3-1)(3-1)\right)\right] = \left[0\,; \chi^2_{0.95}(4)\right] = [0\,; 9.49]$$

Si può ora procedere al calcolo dello stimatore vero e proprio

$$\frac{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(n_{i,j}-\hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} =$$

$$= \frac{(80-75)^2}{75}+\frac{(105-125)^2}{125}+\frac{(65-50)^2}{50}+\frac{(100-105)^2}{105}+\frac{(185-175)^2}{175}+\frac{(65-70)^2}{70}+\frac{(90-90)^2}{90}+\frac{(160-150)^2}{150}+\frac{(50-60)^2}{60} =$$

$$\frac{25}{75}+\frac{400}{125}+\frac{225}{50}+\frac{25}{105}+\frac{225}{175}+\frac{25}{70}+0+\frac{100}{150}+\frac{100}{60}=11.53$$

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%.

Esercizio 3)

Nel testo viene richiesto di stimare la varianza puntualmente e per intervallo dei giorni di permesso richiesti in un anno. Le ipotesi sotto cui questa stima è possibile, con gli strumenti visti nel corso sono

- Si considerano i dati come estrazioni dalla stessa variabile casuale *P.*
- Le estrazioni sono tutte i.i.d.

La seconda ipotesi è un po forzata infatti non si sono "estratti" da tutti i possibili lavoratori italiani 900 possibili studenti ma è stata scelta a caso un azienda e si sono considerati tutti i lavoratori di quell'azienda. In ogni caso il teato richiede di calcolare la stima usando quel particolare campione, senza discuterne la validità.
Stima puntuale. Per stimare puntualmente la varianza si ricorre al suo stimatore la varianza campionaria s^2. Essa è espressa della seguente formula

$$s^2=\sigma^2\frac{n}{n-1}=\left(\frac{\sum_{i=1}^{M}x_i^2 n_i}{n}-\bar{x^2}\right)\frac{n}{n-1}=\left(\frac{\sum_{i=1}^{M}x_i^2 n_i}{n}-\left(\frac{\sum_{i=1}^{M}x_i n_i}{n}\right)^2\right)\frac{n}{n-1}=\left(\frac{28275}{900}-\left(\frac{4613}{900}\right)^2\right)\frac{900}{899}=5.15107$$

Dove il calcolo della varianza e della media campionaria sono state calcolate secondo la seguente tabella:

Giorni richiesti (x_i)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Tot.**
Frequenza (n_i)	80	83	87	77	90	183	180	90	30	900
$x_i\, n_i$	*80*	*166*	*261*	*308*	*450*	*1098*	*1260*	*720*	*270*	*4613*
x_i^2	*1*	*4*	*9*	*16*	*25*	*36*	*49*	*64*	*81*	*--*
$x_i^2\, n_i$	*80*	*332*	*783*	*1232*	*2250*	*6588*	*8820*	*5760*	*2430*	*28275*

Per procedere alla stima per intervallo e necessario fissare un livello di confidenza (1-α); nel nostro caso si è deciso di fissare un livello di confidenza del 95% (corrispondente ad α=5%).
La stima richiesta prevedel il calcolo di un chi quadrato ad n-1 (899) gradi di libertà.
La maggioranza delle tabelle statistche non riporta tale distribuzione, pertanto si utilizza la distribuzione limite del chi quadrato. Infatti, il numero di gradi di libertà è così elevato da poter approssimare la distribuzione in esame con una gaussiana avente pari valor atteso (899) e pari varianza (899*2).

Peranto i due valori critici del chi quadrato che lascisano alla propria sinistra / destra 2.5% possono essere ben approssimati dai rispettivi valori della normale *N(899, 1798)* x_1 *e* x_2. I valori x_1 *e* x_2 si ottengono destandardizzando i rispettivi valori della normale standardizzata secondo la seguente formula $z_1 = -1.96$ e z_2=1.96.

$$x=z*\sqrt{Var[X]}+E[X]=z*\sqrt{1798}+899 \quad , \quad x_1=816 \quad , \quad x_1=982$$

Ora è possibile applicare la formula della stima per intervallo della varianza:

$$Var[P]\in\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)};\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]=\left[\frac{(899)5.15107}{982};\frac{(899)5.15107}{816}\right]=[4{,}716\,;5{,}675]$$

Esercizio 4)

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

Essendo gli eventi elementari equiprobabili, le probabilità degli eventi E_1 ed E_2 e dell'evento intersezione (estrarre un lavoratore sposato che abbia fatto al massimo sei giorni di assenza) possono essere ricavate utilizzando la definizione classica; secondo la quale la probabilità è il rapporto dei casi favorevoli sui casi totali. Pertanto si ha che:

$$P(E_1)=\frac{450}{900}=\frac{1}{2} \qquad P(E_2)=\frac{250+350}{900}=\frac{2}{3} \qquad P(E_1 \cap E_2)=\frac{105+185}{900}=\frac{29}{90}$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1 \cup E_2)=P(E_1)+P(E_2)-P(E_1 \cap E_2)=\frac{450+600-290}{900}=\frac{76}{90} \qquad P(E_1|E_2)=\frac{P(E_1 \cap E_2)}{P(E_2)}=\frac{29}{90}\frac{90}{60}=\frac{29}{60}$$

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

Se due eventi sono indipendenti si ha che la probabilità condizionata è data dal prodotto delle probabilità, pertanto essendo

$$P(E_1)P(E_2)=\frac{1}{2}\frac{2}{3}=\frac{1}{3}\neq\frac{29}{60}=P(E_1 \cap E_2)$$

Gli eventi non sono indipendenti.

- Appello del 22 Febbraio 2012 -

Esercizio 1)

Un laboratorio di analisi vuole verificare la precisione di una macchina di test per misurare la glicemia. Pertanto ha effettuato diverese misure relative a campioni avente concentrazione nota ed ottenendo i seguenti errori.

0.015 0.005 0.030 0.010 0.035 0.020 0.020 0.025 0.010 0.020 0.030

Determinare

a) La tipologia del carattere.
b) Tutti gli indici sintetici di posizione possibili da calcolare indicando se siano opportuni.
c) Se possibile, un indice sintetico di variabilità.
d) Una rappresentazione grafica adeguata.

Esercizio 2)

Un ricercatore vuole verificare se esiste un legame fra la frequenza cardiaca (HR) ed il quoziente ventilatorio (rapporto fra il volume ossigeno ed il volume co_2 nell'aria espulsa ad ogni espirazione). Per tanto, ha sottoposto un paziente ad una prova sotto sforzo (30 minuti di ciclette a velocità costante) ed ogni 5 mininuti ha misurato entrambe le grandezze ottendo le seguenti osservazioni

Minuto	5	10	15	20	25	30
HR [batt/min]	70	76	77	77	80	82
Vo2/Vco2	0.84	0.87	0.89	0.9	0.8	1.1

Il candidato,

a) Indichi e fornisca una rappresentazione grafica adeguata alla serie ottenuta.
b) Se possibile, indichi e calcoli un opportuno indice di variabilità
c) Ipotizzando un legame di tipo lineare,
    1. Calcoli l'opportuna regressione
    2. Il legame ipotizzato è attendibile? Motivare numericamente la risposta.
    3. Ipotizzi quale sarebbe il quoziente respiratorio previsto nel caso si riscontrasse una frequenza cardiaca pari a 72 battiti al minuto

Esercizio 3)

Il candidato, utilizzando i dati dell'Esercizio 1, stimi puntualmente e per intervallo la percentuale d'errore atteso della macchina in esame evidenziando le ipotesi necessarie. Il candidato proceda al calcolo anche se queste risultassero non verificate.

Esercizio 4)

Si considerino i seguenti eventi legati all'estrazione di una delle misure descritte nell'Esercizio 2.

E_1: si estragga una misura in cui la frequenza cardiaca sia supeiore a 77
E_2: si estragga una misura in cui il quoziente respiratorio sia inferiore 0.9

a) Il candidato calcoli le seguenti Probabilità: $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$; $P(E_1 \mid E_2)$.
b) Il candidato indichi se gli eventi E_1 ed E_2 possono ritenersi incompatibili.

- Appello del 22 Febbraio 2012 -
Svolgimento

Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo quantitativo (in quanto espresso da numeri) continuo (in quanto concettualmente un errore relativo può assumere qualsiasi valore positivo)

b) Tutti gli indici sintetici di posizione possibili da calcolare indicando se siano opportuni.

Un carattere di tipo quantitativo ammette tre indici sintetici di posizione: la moda, la mediana e la media.

La moda è che la modalità con la frequenza maggiore: in questo caso essendo le osservazioni continue la moda è l'unico indice di posizione che generalmente non è affidabile. In ogni caso la moda risulta essere 0.020 che ha frequenza assoluta 3 (vedi tabella 1).

Per calcolare la mediana si devono ordinare le osservazioni

0.005 0.010 0.010 0.015 0.020 0.020 0.020 0.025 0.030 0.030 0.035

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 = 5 elementi; Ovvero il sesto elemento, quindi q2 = 0.2.

La media, raggruppati i dati in modalità risulta essere :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{M} x_i * n_i = \frac{0.22}{11} = 0.2$$

Dove i conti sono stati svolti nella seguente tabella in calce all'esercizio

c) Se possibile, un indice sintetico di variabilità..

Il carattere in esame (quantitativo continuo) ammette tutti gli indici di variabilità visti nel corso (range, varianza e distanza interquartile e sqm). Fra questi è stato deciso di calcolare la varianza utizzando la seguente formula

$$\sigma_x^2 = \frac{1}{N}\left(\sum_{i=1}^{M} n_i * x_i^2\right) - \bar{x}^2 = \frac{0.0053}{11} - 0.02^2 = 0{,}00048182 - 0.0004 = 0.00008182$$

Errore (x_i)	0.005	0.010	0.015	0.020	0.025	0.030	0.035	
Frequenza (n_i)	1	2	1	3	1	2	1	
$x_i * n_i$	0.01	0.02	0.02	0.06	0.03	0.06	0.04	0.22
x_i^2	0.000025	0.000100	0.000225	0.000400	0.000625	0.000900	0.001225	
$x_i^2 * n_i$	0.000025	0.000200	0.000225	0.001200	0.000625	0.001800	0.001225	0.005300

Tabella 1) analisi dati Esercizio 1

d) Una rappresentazione grafica adeguata.

Una rappresentazione adeguata per dati qualitativi di una buona numerosità è il box-plot. Per descrivere un boxplot si debbono calcolare i quartili. I quartili sono 5 numeri dove il primo (quartile zero) e l'ultimo (quartile 4) sono gli estremi delle osservazioni, mentre il terzo è la mediana. Il primo ed il terzo quartile sono quelle osservazioni che lascisano rispettivamente alla destra ed alla sinistra un quarto delle osservazioni ordinate ovvero *(N-1)/2 = 2.5*. Poichè il numero non è tondo si mediano le due osservazioni successive; pertanto si ha che il primo quartile è la media fra la terza e quarta osservazione mentre il secondo sarà la media fra l'ottava ed il nona osservazione

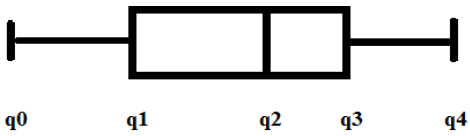
$$q_0 = 0.005 \qquad q_1 = (0.010 + 0.015)/2 = 0.0125 \qquad q_2 = 0.020 \qquad q_3 = (0.025 + 0.030)/2 = 0.0275 \qquad q_4 = 0.035$$

Per completare la raccolta di infomrazioni richiesta per il box-plot si debbono calcolare il 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 quatile *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 = 0.0125 - 1 (0.0275- 0.0125) = -0.0025* *VAI = 0.0275 + 1* (0.0275- 0.0125) = 0.045*

Essendo tutte le osservazioni interne all'intervallo [VAI ; VAS] possiamo concludere che non vi siano outlier. A lato si riporta il box-plot ottenuto

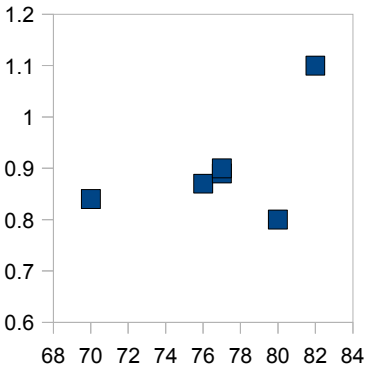


Esercizio 2)

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. Questi diagrammi sono diagrammi cartesiani i cui le modalità dei caratteri vengono posti sui due essa 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: Frequenza cardiaca

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{n} x_i=\frac{70+76+77+77+80+82}{6}=77$$

$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{7^2+1^2+0^2+(-3)^2+(5)^2}{6}=\frac{49+1+9+25}{6}=14$$

Y: Coefficiente Ventilatorio

$$\bar{y}=\frac{1}{N}\sum_{i=1}^{n} y_i=\frac{0.84+0.87+0.89+0.90+0.80+1.1}{6}=0.90$$

$$\sigma_y^2=\frac{1}{N}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{(-0.06)^2+(-0.03)^2+(-0.01)^2+(0)^2+(0.1)^2+(0.2)^2}{6}=\frac{5.4}{6}=0.9$$

La covarianza si ottiene

$$\sigma_{xy}=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{0.42+0.03+0+0-0.3+1}{6}=\frac{1.15}{6}$$

I cui conti sono ripostati in Tabella 2.Pertanto la matrice varianza covarianza risulta essere

$$\Sigma=\begin{bmatrix} 14 & \frac{1.15}{6} \\ \frac{1.15}{6} & 0.9 \end{bmatrix}$$

x_i	70	76	77	77	80	82
y_i	0.84	0.87	0.89	0.9	0.8	1.1
$x_i - \bar{x}$	-7	-1	0	0	3	5
$y_i - \bar{y}$	-0.06	-0.03	-0.01	0	-0.1	0.2
$(y_i - \bar{y})(x_i - \bar{x})$	0.42	0.03	0	0	-0.3	1

Tabella 2) Dati relativi Esercizio 2 e 4

c 1) Ipotizzando un legame di tipo lineare, si calcoli l'opportuna regressione
La retta di regressione ha equazione

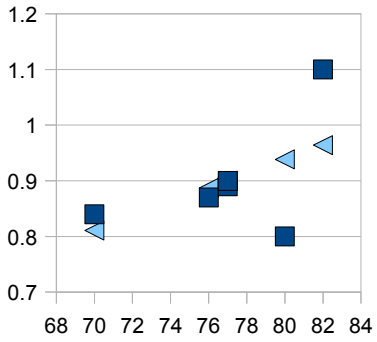
$$\hat{y}=\frac{\sigma_{xy}}{\sigma_x^2}x+\bar{y}-\frac{\sigma_{xy}}{\sigma_x^2}\bar{x} \qquad \hat{y}=\frac{1.15}{6*14}x+0.9-\frac{1.15}{6*14}77 \qquad \hat{y}=0.0137x-0.1542$$

c 2) 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

$$R^2=\frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}=0.0029 \qquad R=0.053$$

Poiche l'indice risulta inferiore a 0.3 si può asserire che il legame è difficilemente possibile. Ovviamente il dato deve essere confermato dalla visualizzazione del modello. Infatti il coefficientei di Pearson può anche dare risultati molto errati. A lato si riportano le presevisioni effettuate dal modello lineare che non sempre descrivono l'andamento dei dati



c 3) Ipotizzi quale sarebbe il quoziente respiratorio previsto nel caso si riscontrasse una frequenza cardiaca pari a 72 battiti al minuto
La risposta a questo quesito si ottiene applicando la retta nel punto *x = 72;* si ottiene quindi un quoziente ventilatorio previsto di 0.8322

Esercizio 3)
Le tecniche di stima viste nel corso prevedono che:

- la popolazione sia describile mediante una variabile casuale,
- che il campione abbia una numerosità tale da far convergere lo stimatore e
- che le prove siano indipendenti ed identicamente distribuite (i.i.d.).

Nel caso in esame

- descrivere l'esperimento mediante la seguente variabile casuale *X: errore relativo di una misura*.
- la grandezza da stimare risulta *E[X]* il cui stimatore è la media campionaria che converge in legge per campioni avente numerosità superiore a 30 (ipotesi non confermata).
- L'ipotesi di prove i.i.d. viene confermata in quanto si suppone che una sequenza di 6 prove (ovvero l'effettuare una misura) non alteri la distribuzione di probabilità di *X* (rovini lo strumento di misura).

La stima puntuale si ottiene semplicemente dall'applicazione dello stimatore, pertanto

$$\hat{\theta}=E[\hat{X}]=\bar{x}=0.2$$

Per effettuare una stima per intervallo si deve come prima cosa fissare un livello di confidenza, nel nostro caso 95% (α=0.05). Definita la stima (stima per intervallo al 95%), si ha che la stima i intervallo è data dalla seguente

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

Dove la varianza viene stimata mediante la varianza campionaria ed il valore della normale standardizzata si ricava dalle tavole:

$$\bar{x}=0.2 \qquad \hat{Var}[X]=s^2=\sigma^2\frac{n}{n-1}=0.00008122\frac{11}{10}=0{,}000090 \qquad z_{\alpha/2}=1.96$$

Infine si otiene quindi:

$$E[X]\in[0.02-1.96*0.002864 \;;\; 0.02+1.96*0.002864]\Rightarrow E[P]\in[0.0143936 \;;\; 0.025606]$$

Esercizio 4)

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

Essendo gli eventi elementari equiprobabili (ogni coppia di misure ha la stessa probabilità di essere estratta), le probabilità degli eventi E_1 ed E_2 e dell'evento intersezione (estrarre una misura in cui la frequenza cardiaca sia supeiore a 77 ed il quoziente respiratorio sia inferiore a 0.9) possono essere ricavate utilizzando la definizione classica; secondo la quale la probabilità è il rapporto dei casi favorevoli sui casi totali. Pertanto si ha che:

$$P(E_1)=\frac{2}{6}=0.333 \qquad P(E_2)=\frac{4}{6}=0.667 \qquad P(E_1 \cap E_2)=\frac{1}{6}=0.167$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1 \cup E_2)=P(E_1)+P(E_2)-P(E_1 \cap E_2)=\frac{2+4-1}{6}=0.833 \qquad P(E_1|E_2)=P\frac{(E_1 \cap E_2)}{P(E_2)}=\frac{1/6}{4/6}=0.250$$

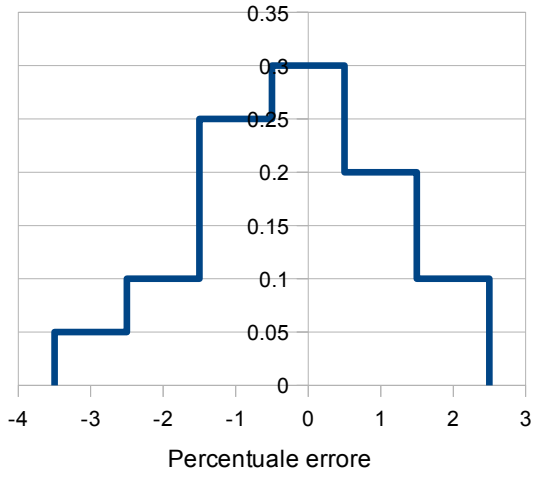
b) Il candidato indichi se i due eventi E_1 ed E_2 sono incompatibili.

Se due eventi sono incompatibili se non possono verificarsi contemporaneamente pertanto la probabilità dell'evento intersezione è nulla. Essendo questa probabilità pari ad un sesto possiamo affermare che gli eventi non sono incompatibili (quindi sono compatibili).

- Appello del 27 Giugno 2012 -

Esercizio 1)

Un laboratorio di analisi vuole comprare una macchina per la rilevazione dell'emoglobina. Il costruttore fornisce a testimonianza della bontà della propria macchina l'istogramma dell'errore di misura riportato a lato relativo alla misurazione di 1000 campioni



Il candidato

a) Determini la tipologia del carattere.
b) Fornisca una rappresentazione tabellare dei dati (mettendo in risalto le frequenze assolute).
c) Se possibile, calcoli la mediana.
d) Se possibile, calcoli la varianza.

Esercizio 2)

Un ricercatore vuole verificare se esista un legame fra le ore di sonno ed il livello di glicemia al risveglio in un soggetto diabetico. Per far ciò ha sottoposto lo stesso soggetto ad un protocollo sperimentale che prevede il monitoraggio di 6 notti di sonno ottenendo i seguenti dati

Notte	I	II	III	IV	V	VI
Ore di Sonno	5	6	6	7	7	8
Glicemia alle 7:30 [mg/dl]	71	75	70	75	82	80

Il candidato,

a) Indichi e fornisca una rappresentazione grafica adeguata alla serie ottenuta.
b) Se possibile, indichi e calcoli un opportuno indice di variabilità
c) Ipotizzando un legame di tipo lineare,
   1. Calcoli l'opportuna regressione
   2. Il legame ipotizzato è attendibile? Motivare numericamente la risposta.
   3. Ipotizzi quale sarebbe il valore di glicemia se il soggetto dormisse 24 ore.

Esercizio 3)

Il candidato, utilizzando i dati dell'Esercizio 1, stimi puntualmente e per intervallo la varianza della percentuale d'errore attesa della macchina in esame evidenziando le ipotesi necessarie. Il candidato proceda al calcolo anche se queste risultassero non verificate.

Esercizio 4)

Si considerino i seguenti eventi considerati indipendenti:

E_1: si abbia $z < 0$ dove z è una normale standardizzata
E_2: si abbia $x = 0$ dove $x \sim Ber(0.3)$

a) Il candidato calcoli le seguenti Probabilità: $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$; $P(E_1 \mid E_2)$.
b) Il candidato indichi se gli eventi E_1 ed E_2 possono ritenersi incompatibili.

- Appello del 27 Giugno 2012 -
Svolgimento

Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo quantitativo (in quanto espresso da numeri) continuo (in quanto concettualmente un errore percentuale può assumere qualsiasi valore)

b) Fornisca una rappresentazione tabellare dei dati (mettendo in risalto le frequenze assolute).

In un istogramma viene riportata sulle ordinate la densità di frequenza (d_i) data dal rapporto fra le frequenze relative (f_i) e l'ampiezza della classe cui sono riferite (sup_i - inf_i). Ricordando che le frequenze relative son il rapporto fra le frequenze assolute (n_i) ed il totale delle osservazioni (N), si ha che

$$d_i = \frac{n_i}{N(sup_i - inf_i)} \quad \text{da cui si ottiene che} \quad n_i = N\, d_i (sup_i - inf_i)$$

Poiché dal grafico è possibile leggere direttamente gli estremi delle classi ed i valori della densità di frequenza è possibile applicando banalmente la formula testè ricavata ottenere i dati riportati in Tabella 1.

c) Se possibile, calcoli la mediana.

La mediana è il valore che bipartisce le osservazioni ordinate, ovvero, quel valore che bipartisce l'area sottesa dell'istogramma. Considerando le frequenze cumulate in Tabella 1, si osserva come la mediana cada nella quarta classe ($i^* = 4$) contente i valori fra 40% e 70% delle misurazioni ordinate. Per calcolare la mediana si deve trovare la parte del rettangolo relativo alla quarta classe che sottenda solo il 10% ($50\ \% - F_{i^*}$) delle misurazioni. Poichè l'atezza del rettangolo è nota ($d_{i^*} = 0.3$) possiamo facilmente ricavarne la base (0.1 / 0.3 = 1/3). Quindi la mediana si avrà sommando questo valore all'estremo inferiore della classe ($inf_{i^*} = -0.5$) ricavando il valore di -1/6.
Lo stesso risultato poteva essere ottenuto applicando la seguente formula che riassume il procedimento appena descritto

$$Me = inf_{i^*} + (0.5 - F_{i^*}) / d_{i^*}$$

d) Se possibile, si calcoli la varianza.

Il carattere in esame (quantitativo continuo) ammette tutti gli indici di variabilità visti nel corso (range, varianza e distanza interquartile e sqm) anche se ottenuto con sola rappresentazioni per classi di osservazioni. In questo caso gli indici sono ricavabili abbinando ad ogni classe il valore centrale della classe (c_i). La varianza delle osservazioni è stata calcolata utilizzando i dati ricavati in Tabella 1 nella seguente formula

$$\sigma_x^2 = \left(\sum_{i=1}^{M} f_i * c_i^2\right) - \bar{x}^2 = \left(\sum_{i=1}^{M} f_i * c_i^2\right) - \left(\sum_{i=1}^{M} f_i * c_i\right)^2 = 1.7 - (-0.2)^2 = 1.7 - 0.04 = 1.66$$

i	inf_i	sup_i	c_i	n_i	f_i	F_i	$c_i * f_i$	c_i^2	$c_i^2 * f_i$
1	-3.5	-2.5	-3	50	0.050	0.050	-0.1500	9	0.45
2	-2.5	-1.5	-2	100	0.100	0.150	-0.2000	4	0.4
3	-1.5	-0.5	-1	250	0.250	0.400	-0.2500	1	0.25
4	-0.5	0.5	0	300	0.300	0.700	0.0000	0	0
5	0.5	1.5	1	200	0.200	0.900	0.2000	1	0.2
6	1.5	2.5	2	100	0.100	1.000	0.2000	4	0.4
				Totali	1		-0.2000		1.7000

Tabella 1) analisi dati 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. Questi diagrammi sono diagrammi cartesiani i cui le modalità dei caratteri vengono poste sui due assi ed ogni osservazione viene rappresentata da un punto. Il grafico ottenuto dai dati nella consegna viene riportato in Figura 1 (serie "Dati Reali").

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 varianze dei distinti caratteri e la covarianza della serie bivariata. Si seguito riportiamo i calcoli relativi alle varianze dei i singoli caratteri:

X: Ore di sonno

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{n}x_i=\frac{5+6+6+7+7+8}{6}=6.5$$

$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{(5-6.5)^2+2*(6-6.5)^2+2*(7-6.5)^2+(8-6.5)^2}{6}=\frac{2.25+0.5+0.5+2.25}{6}=\frac{5.5}{6}$$

Y: Glicemia al mattino

$$\bar{y}=\frac{1}{N}\sum_{i=1}^{n}y_i=\frac{71+75+70+75+82+80}{6}=75.5$$

$$\sigma_y^2=\frac{1}{N}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{(-4.5)^2+(-0.5)^2+(-5.5)^2+(-0.5)^2+(6.5)^2+(6.75)^2}{6}=\frac{113.5}{6}$$

Sfruttando i conti ripostati in Tabella 2 si ottiene la seguente covarianza:

$$\sigma_{xy}=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{6.75+0.25+2.75-0.25+3.25+6.75}{6}=\frac{19.5}{6}$$

Pertanto la matrice varianza/covarianza risulta essere

$$\Sigma=\begin{bmatrix}\frac{5.5}{6} & \frac{19.5}{6}\\ \frac{19.5}{6} & \frac{113.5}{6}\end{bmatrix}$$

	Osservazioni						**Totali**
$\mathbf{x_i}$	5	6	6	7	7	8	6.5000
$\mathbf{y_i}$	71	75	70	75	82	80	75.5000
$\mathbf{x_i-\bar{x}}$	-1.5	-0.5	-0.5	0.5	0.5	1.5	
$\mathbf{y_i-\bar{y}}$	-4.5	-0.5	-5.5	-0.5	6.5	4.5	
$\mathbf{(y_i-\bar{y})(x_i-\bar{x})}$	6.75	0.25	2.75	-0.25	3.25	6.75	19.5000

Tabella 2) Dati relativi Esercizio 2

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

La retta di regressione ha equazione

$$\hat{y}=\frac{\sigma_{xy}}{\sigma_x^2}x+\bar{y}-\frac{\sigma_{xy}}{\sigma_x^2}\bar{x}\qquad \hat{y}=\frac{19.5}{5.5}x+75.5-\frac{19.5}{5.5}6.5\qquad \hat{y}=3.54\,x-52.45$$

c 2) 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

$$R^2=\frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}=0.61\qquad R=0.78$$

Poiche l'indice risulta superiore a 0.7 si può asserire che il legame è possibile. Ovviamente il dato deve essere confermato dalla visualizzazione del modello. Infatti il coefficiente di Pearson può anche dare risultati fuorvianti. A lato si riportano le presevisioni effettuate dal modello lineare che descrivono l'andamento dei dati con buona precisione.

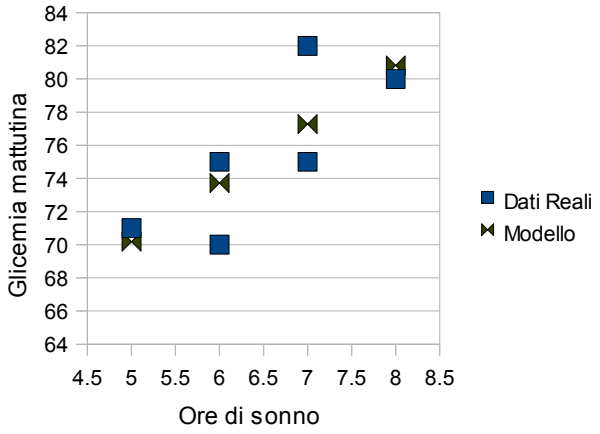


Figura 1) Rappresentazione dei dati dell Es 2

c 3) Ipotizzi quale sarebbe il valore di glicemia se il soggetto dormisse 24 ore.
La risposta a questo quesito si ottiene applicando la retta nel punto *x = 24;* si ottiene quindi un quoziente ventilatorio previsto di 137.55.
Si ricorda che il valore risulta poco attendibile poiché il modello viene applicato in ascisse (24) molto lontane da quelle usate per stimarlo (5-8).

Esercizio 3)
Le tecniche di stima viste nel corso prevedono che:

- la popolazione sia descrivibile mediante una variabile casuale,
- che il campione abbia una numerosità tale da far convergere lo stimatore e
- che le prove siano indipendenti ed identicamente distribuite (i.i.d.).

Nel caso in esame

- descrivere l'esperimento mediante la seguente variabile casuale *X: errore percentuale di una misura*.
- la grandezza da stimare risulta *E[X]* il cui stimatore è la media campionaria la quale converge in legge per campioni avente numerosità superiore a 30 (ipotesi confermata).
- L'ipotesi di prove i.i.d. viene confermata in quanto si suppone che una sequenza di 1000 prove (ovvero l'effettuare una misura) non alteri la distribuzione di probabilità di *X* (rovini lo strumento di misura) di una macchina da laboratorio progettata per svolgere molte più analisi.

La stima puntuale si ottiene semplicemente dall'applicazione dello stimatore, pertanto

$$\hat{\theta} = \hat{Var}[X] = s_x^2 = \sigma^2 \frac{n}{n-1} = 1.7 * 1000/999 = 1.7017$$

Si noti come con grandi numeri la varianza e la varianza campionaria tendano a conicidere

Per effettuare una stima per intervallo si deve come prima cosa fissare un livello di confidenza, nel nostro caso 95% (α=0.05). Definita la tipologia di stima (stima per intervallo al 95%), si ha che essa è data dalla seguente

$$Var[P] \in \left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} ; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right]$$

Dove i valori del della $\chi^2(n-1)$ si dovrebbero ricavare dalle tavole, ma esse normalmente non riportano valori per un numero così elevato di gradi di libertà. In questi casi è possibile utilizzare la convegenza in legge della varibile chi qudrato. Infatti essa per un alto numero di gradi di libertà converge ad una normale con stesso valor medio e varianza. Si ha infatti che

$$\lim_{n \to +\infty} \chi^2(n) = N(n, 2n) \quad \text{nel nostro caso} \quad \chi^2(999) \sim N(999, 1998)$$

Per ottenere i valori critici richiesti è possibile reperire i valori su di una standardizzata $z_{\alpha/2} = 1.96$ e poi destandardizzarli secondo la seguente formula

$$x = \sigma * z + \mu$$

Si ottengono quindi i due valori critici richiesti

$$\chi^2_{\frac{\alpha}{2}}(n-1) = \sqrt{Var[\chi^2(n-1)]}\, z_{\frac{\alpha}{2}} + E[\chi^2(n-1)] = \sqrt{1998}(-1.96) + 999 = 991$$

$$\chi^2_{1-\frac{\alpha}{2}}(n-1) = \sqrt{Var[\chi^2(n-1)]}\, z_{\frac{\alpha}{2}} + E[\chi^2(n-1)] = \sqrt{1998}\, 1.96 + 999 = 1086$$

Infine si ottiene la stima richiesta:

$$Var[P] \in \left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} ; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right] = \left[\frac{999 * 1.7017}{1086} ; \frac{999 * 1.7017}{991} \right] = [1.565 ; 1.7154]$$

Esercizio 4)

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

L'evento E_1 è dato dalla probabilità di estrarre un numero negativo da una normale standardizzata. Poiché essa ha mediana pari a 0, suddetta probabilità è pari al 50%. (Si ricorda che la mediana bipartisce l'area sottesa da una densità di probabilità in parti uguali)

L'evento E_2 è dato dalla probabilità di avere un esito negativo in una prova di Bernoulli con probabilità di esito positivo pari al 30%. Poiché in una prova di Bernoulli vi sono solo due esiti fra loro complementari abbiamo che

$$P(E_2) = 1 - P(\overline{E}_2) = 1 - P(X = 1) = 1 - 0.3 = 70\%$$

Essendo gli eventi indipendenti la probabilità dell'evento intersezione è data dal prodotto delle probabilità

$$P(E_1 \cap E_2) = P(E_1)\,P(E_2) = 0.5 * 0.7 = 0.35$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.5 + 0.7 - 0.35 = 0.85 \qquad P(E_1 | E_2) = P\frac{(E_1 \cap E_2)}{P(E_2)} = \frac{0.35}{0.7} = 0.5 = P(E_1)$$

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

Due eventi sono incompatibili se non possono verificarsi contemporaneamente ne consegue che la probabilità dell'evento intersezione è nulla. Nel caso in esame questa probabilità è non nulla, quindi è possibile affermare che gli eventi non sono incompatibili.

- Appello del 11 Luglio 2012 -

Esercizio 1)
Si vuole monitorare la presenza di effetti collaterali nella somministrazione di un farmaco contro una particolare emicrania dovuta a cause vascolari. A tale scopo si sono scelti 20 soggetti malati e per ognuno di questi si è si sono rilevati il numero di effetti collaterali

Numero di Effetti Collaterali	0	1	2	3
Frequenza assoluta	10	6	3	1

a) Determini la tipologia del carattere.
b) Fornisca una rappresentazione grafica dei dati.
c) Se possibile, calcoli un indice di variabilità.
d) La distribuzione dei dati è approssimabile con una normale? Motivare numericamente la risposta.

Esercizio 2)

Si vuole verificare se l'uso del farmaco come placebo (ovvero in soggetti che non necessitano di trattamento), si è voluto verificare se l'incidenza degli effetti colaterali sia influenzato dalla presenza della malattia. Pertanto si sono presi dei soggetti sani e si è ripetuto il protocollo descritto all'Esercizio 1 ottenendo i seguenti dati:

Numero di Effetti Collaterali	0	1	2	3
Frequenza assoluta	12	10	6	2

Il candidato,

a) realizzi una tabella a doppia entrata che raccolga i dati ricavati nel primo e nel secondo Esercizio (effetti collaterali in pazienti sani e non);
b) se possibile, indichi e calcoli un opportuno indice di posizione per la serie bivariata;
c) se possibile, indichi e calcoli un opportuno indice di variabilità per la serie bivariata;
d) verifichi se l'uso dell farmaco in soggetti sani (ovvero la cui emicrania è di origine psicologica) aumenta l'incidenza degli effetti collateriali. Il candidato indichi le necessarie ipotesi e proceda al calcolo anche se queste non fossero soddisfatte.

Esercizio 3)
Il candidato, utilizzando i dati riportati negli Esercizi 1 e 2, stimi puntualmente e per intervallo il valore atteso del numero di effetti collaterali evidenziando le ipotesi necessarie. Il candidato proceda al calcolo anche se queste risultassero non verificate.

Esercizio 4)

Si considerino i seguenti eventi considerati incompatibili:

E_1: si abbia $x < 5$ dove x e estratto da una normale con valore atteso 5 e varianza 1.
E_2: estraendo un soggetto a caso dalla sperimentazione descritta all'Esercizio 1, questi abbia mostrato più di un effetto collaterale

a) Il candidato calcoli le seguenti Probabilità: $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$; $P(E_2 \mid E_1)$.
b) Il candidato indichi se gli eventi E_1 ed E_2 possono ritenersi dipendenti.

- Appello del 11 Luglio 2012 -
Svolgimento

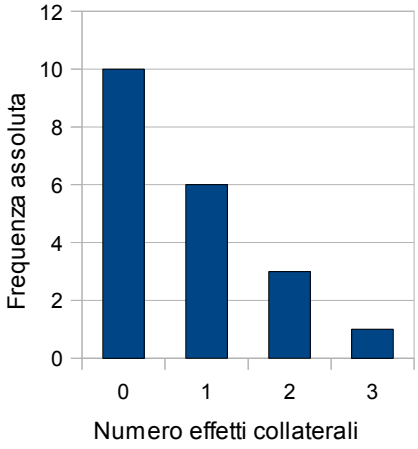
Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo quantitativo (in quanto espresso da numeri) discreto (in quanto non è possibile mostrare 1 effetto collaterale e mezzo).

b) Fornisca una rappresentazione grafica adeguata dei dati .

Per un carattere quantitavo discreto una rappresentazione dei dati idonea può essere il diagramma a barre. Questo diagramma è ottenuto ponendo sulle ascisse di un piano cartesiano le modalità delle osservazioni e disegnando per ogni modalità un rettangolo la cui altezza è pari alla relativa frequenza assoluta. A lato si mostra il diagramma ottenuto dai dati in oggetto.



c) Se possibile, calcoli un indice di variabilità.

Il carattere in esame (quantitativo discreto) ammette tutti gli indici di variabilità visti nel corso (range, varianza e distanza interquartile e sqm). La varianza delle osservazioni è stata calcolata utilizzando i dati ricavati in Tabella 1 nella seguente formula

$$\sigma_x^2=\left(\sum_{i=1}^{M} f_i * c_i^2\right)-\bar{m}^2=\left(\sum_{i=1}^{M} f_i * c_i^2\right)-\left(\sum_{i=1}^{M} f_i * c_i\right)^2=1.35-0.75^2=1.35-0.7875=1.7875$$

d) La distribuzione dei dati è approssimabile con una normale? Motivare numericamente la risposta.

Una distribuzione viene approssimata ad una normale se è simmetrica (la normale è una d.d.p. simmetrica) e se il sua d.d.p. presenta un andamento simile a quello di una normale con lo stesso valor atteso e la stessa varianza. Spesso non si testa la simmetria, in ogni caso ognuna delle due carattesistiche può essere testata con diversi indici

- *Simmetria*: momento centrale terzo; indice di simmetria (momento centrale terzo standardizzato); indice di asimmetria del Pearson
- *Vicinanza alla gaussiana*: centrale quarto oppure indice di curtosi (momento centrale quarto standardizzato) oppure eccesso curtosi.

Nel caso in esame abbiamo sfruttato i valori in Tabella 1 per il calcolo degli gli indici di simmetria e di eccesso curtosi: se entrambi i valori sono prossimi allo zero l'approssimazione può definirsi accettabile.

- simmetria $\gamma_1=\frac{\mu_3}{\sigma^3}=\frac{\sum_{i=1}^{M} f_i\left(m_i-\bar{m}\right)^3}{\left(\sqrt{\sigma^2}\right)^3}=\frac{0.6563}{\left(\sqrt{0.7875}\right)^3}=0.939$
- vicinanza alla gaussiana $\gamma_2=\frac{\mu_4}{\sigma^4}-3=\frac{\sum_{i=1}^{M} f_i\left(m_i-\bar{m}\right)^4}{\left(\sigma^2\right)^2}=\frac{1.81}{\left(0.7875\right)^2}-3=-0.09$

Sebbene l'eccesso curtosi sia prossimo allo zero la distribuzione è fortemente asimmetrica (si ricorda che il valore del coefficiente di curtosi ha come massimo l'unita). Pertanto l'approssimazione alla gaussiana è da sconsigliarsi.

Effetti Collaterali modalità m_i	frequenze assolute n_i	frequenze relative f_i	$m_i * f_i$	m_i^2	$m_i^2 * f_i$	scarto $m_i-\bar{m}$	$(m_i-\bar{m})^3$	$(m_i-\bar{m})^3 f_i$	$(m_i-\bar{m})^4$	$(m_i-\bar{m})^4 f_i$
0	10	0.50	0	0	0	-0.75	-0.42	-0.21	0.32	0.16
1	6	0.30	0.300	1	0.3	0.25	0.02	0	0.00	0
2	3	0.15	0.300	4	0.6	1.25	1.95	0.29	2.44	0.37
3	1	0.05	0.150	9	0.45	2.25	11.39	0.57	25.63	1.28
Totali	20	1	0.750		1.35			0.6563		1.81

Tabella 1) analisi dati Esercizio 1

Esercizio 2)

a) Realizzi una tabella a doppia entrata che raccolga i dati ricavati nel primo e nel secondo Esercizio (effetti collaterali in pazienti sani e non).

Nei due esercizi si descrive una bivariata che raccoglie i seguenti caratteri

- *X*: stato del soggetto (carattere qualitativo non ordinabile)
- *Y*: numero di effetti collaterali riscontrati (carattere quantitativo discreto)

Le cui osservazioni possono essere raccolte nella seguente tabella a doppia entrata.

		Y: numero di effetti collaterali riscontrati				Totali
		0	1	2	3	
X: stato	Sano	12	10	6	2	30
	Malato	10	6	3	1	20
	Totali	22	16	9	3	50

b) Se possibile, indichi e calcoli un opportuno indice di posizione per la serie bivariata

Per serie bivariate con almeno un carattere quantitativo, l'unico indice di posizione possibile è la moda ovvero la modalita corrispondente alla frequenze maggiore. Nel nostro caso la frequenza magiore è 12, pertanto la moda è:

(Sano; 0)

c) Se possibile, indichi e calcoli un opportuno indice di variabilità per la serie bivariata

Non è possibile, con le tecniche viste nel corso, calcolare indici di variabilità per serie bivariate con almeno un carattere quantitativo.

d) verifichi se l'uso dell farmaco in soggetti sani (ovvero la cui emicrania è di origine psicologica) aumenta l'incidenza degli effetti collateriali. Il candidato indichi le necessarie ipotesi e proceda al calcolo anche se queste non fossero soddisfatte.

Se vi fosse un legame fra i due caratteri questi dovrebbero essere dipendenti; pertanto come prima cose di valuta la loro indipendenza. Per verificare se i due caratteri sono indipendenti si può effettuare un test di ipotesi volto a verificare se le frequenze delle osservazioni rilevate nel campione sono sufficiente mente vicine (ad un determinato livello di significatività) a quelle teoriche ottenute dall'ipotesi di indipendeza. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadranto avente gradi di libertà paria quelli del 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 .

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

le frequenze teoriche sono state riportate fra parentesi nella seguente tabella

		Y: numero di effetti collaterali riscontrati				Totali
		0	1	2	3	
X: stato	Sano	12 (13.2)	10 (9.6)	6 (5.4)	2 (1.8)	30
	Malato	10 (8.8)	6 (6.4)	3 (3.6)	1 (1.2)	20
	Totali	22	16	9	3	50

A questo punto è possibile valutare la convergenza dell stimatore di Pizzetti Pearson mediante la verifica dell'unica ipotesi vista a tal riguardo ovvero che tutte le frequenze teoriche siano superiori a 5.
Constatato che la condizione non è verificata ci si dovrebbe fermare nella valutazione. Nonostante ciò il testo richiede di procedere nella procedura.

Si calcola la regione di accettazione dopo aver fissato il livello di significatività (posto nel nostro caso al 5%).

$$A=\left[0\,;\chi^2_{1-\alpha}\left((M_x-1)(M_y-1)\right)\right]=\left[0\,;\chi^2_{1-0.05}\left((2-1)(4-1)\right)\right]=\left[0\,;\chi^2_{0.95}(3)\right]=[0\,;7.82]$$

Si può ora procedere al calcolo dello stimatore vero e proprio

$$\frac{\sum_{i=1}^{2}\sum_{j=1}^{4}\left(n_{i,j}-\hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}}=$$

$$=\frac{(12-13.2)^2}{13.2}+\frac{(10-9.6)^2}{9.6}+\frac{(6-5.4)^2}{5.4}+\frac{(2-1.8)^2}{1.8}+\frac{(10-8.8)^2}{8.8}+\frac{(10-8.8)^2}{8.8}+\frac{(6-6.4)^2}{6.4}+\frac{(3-3.4)^2}{3.4}+\frac{(1-1.2)^2}{1.2}=$$

$$0.11+0.02+0.07+0.02+0.16+0.03+0.10+0.03=0.54$$

Poichè il valore dello stimatore è interno all'intervallo di accettazione posso dire che i due caratteri sono indipendenti ad un livello di significatività del 5%. Quindi si può scartare l'ipotesi che esista un influenza fra il numero di effetti collaterali riscontrati e lo stato del soggetto.

Esercizio 3)

Le tecniche di stima viste nel corso prevedono che:

- la popolazione sia descrivibile mediante una variabile casuale,
- che il campione abbia una numerosità tale da far convergere lo stimatore e
- che le prove siano indipendenti ed identicamente distribuite (i.i.d.).

Nel caso in esame

- descrivere l'esperimento mediante la seguente variabile casuale *X: numero di effetti collaterali riscontrati*.
- la grandezza da stimare risulta *E[X]* il cui stimatore è la media campionaria la quale converge in legge per campioni avente numerosità superiore a 30 (ipotesi confermata, si hanno infatti 50 prove:30 provenienti dal primo esercizio e 20 dal secondo).
- L'ipotesi di prove i.i.d. è un po debole in quanto si suppone che la probabilità di avere un effetto collateriale sia distribuita in maniera identica per ogni soggetto.

La stima puntuale si ottiene semplicemente dall'applicazione dello stimatore, pertanto

$$\hat{\theta}=E[\hat{X}]=\bar{x}=\sum_1^M f_i m_i=0.86$$

Per effettuare una stima per intervallo si deve, come prima cosa, fissare un livello di confidenza: nel nostro caso 95% (α=0.05).Avendo così definito la tipologia di stima (stima per intervallo al 95%), si ha che essa è data dalla seguente formula

$$E[\hat{P}]\in\left[\bar{x}-z_{1-\frac{\alpha}{2}}\frac{\sqrt{Var[X]}}{N}\,;\bar{x}+z_{1-\frac{\alpha}{2}}\frac{\sqrt{Var[X]}}{N}\right]$$

Dove il valore della normale standardizzata si ottiene dalle tavole mentre la varianzia della popolazione si stima puntualmente come illustrato nel seguito.

- $z_{1-\alpha/2}$: con questa notazione si intende il valore di *z* che lasci alla sua sinistra una probabilità (data dall'area sottesa dalla d.d.p) pari ad *1-α/2*.

  $$\int_{-\infty}^{z_{1-\alpha/2}} f(x)\,dx=1-\alpha/2$$

  Ricodando che nelle tavole sono graficati le aree sottese dalla normale standardizzata fra 0 ed un valore positivo di *Z*, dobbiamo trovare un modo per rincodurci all'uso di questa tipologia di integrali. Questo può essere fatto spezzando l'integrale in due: fra meno infinito e zero e fra zero e $z_{1-\alpha/2}$. In simboli:

  $$\int_{-\infty}^{z_{1-\alpha/2}} f(x)\,dx=\int_{-\infty}^{0} f(x)\,dx+\int_{0}^{z_{1-\alpha/2}} f(x)\,dx=1-\alpha/2$$

  Elaborando gli ultimi due membri l'equazione si ottiene il seguente risultato

  $$\int_{0}^{z_{1-\alpha/2}} f(x)\,dx=-\int_{-\infty}^{0} f(x)\,dx+1-\alpha/2=0.5-\alpha/2=0.5-0.05/2=0.475$$

  Pertanto il valore $z_{1-\alpha/2}$ è quello a cui sulle tavole corrisponde l'area di 0.475; ottenendo

  $$z_{\alpha/2}=1.96$$

- *Stima della varianza*. La varianza viene stimata utilizzando il suo stimatore corretto: la varianza campionaria s^2. Essa viene ricavata utilizzando le sguente formula in cui i dati sono estratti dalla Tabella 2

$$s^2=\sigma^2\frac{N}{N-1}=(\sum_{i=1}^{M}f_i m_i^2-\bar{m})\frac{N}{N-1}=\left(\sum_{i=1}^{M}f_i m_i^2-\left(\sum_{i=1}^{M}f_i m_i\right)^2\right)\frac{N}{N-1}=(1.58-0.86^2)\frac{50}{49}=0.88$$

Infine si ottiene la stima richiesta:

$$E[\hat{P}]\in\left[\bar{x}-\sqrt{\frac{Var[X]}{N}};\bar{x}+z_{\frac{\alpha}{2}}\sqrt{\frac{Var[X]}{N}}\right]=\left[0.86-1.96\sqrt{\frac{0.88}{50}};0.86+1.96\sqrt{\frac{0.88}{50}}\right]=[0.60;1.12]$$

Effetti Collaterali modalità m_i	frequenze assolute n_i	frequenze relative f_i	$m_i * f_i$	m_i^2	$m_i^2 * f_i$
0	22	0.4400	0	0	0.0000
1	16	0.3200	0.3200	1	0.3200
2	9	0.1800	0.3600	4	0.7200
3	3	0.0600	0.1800	9	0.5400
Totali	50		0.8600		1.5800

Tabella 2) Dati Esercizio 3 ottenuti raggruppando i dati degli Esercizi 1 e 2.

Esercizio 4)

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

L'evento E_1 è dato dalla probabilità di estrarre un numero minore di X<5 dove $X \sim N(5,1)$. Per ottenere questa probabilità si deve standardizzare la v.c. X, ottenendo il corrispondente valore standardizzato

$$Z=\frac{X-E[X]}{\sqrt{Var[X]}}\rightarrow z=\frac{5-5}{\sqrt{1}}=0$$

Pertanto si ha che

$$P(E_1)=P(X<5) = P(Z<0) = 0.5$$

L'evento E_2 può essere ricavato utilizzando la definizione classica di probabilità (casi favorevoli su casi totali). I casi totali sono date dalle 20 osservazioni, mentre i casi favorevoli sono quelli in cui si osservano più di un effetto collaterale (quindi i casi in cui si osservano 2 o 3 effetti collaterali). Pertanto si ha che

$$P(E_2)=\frac{3+1}{20}=\frac{4}{20}=0.2$$

Essendo gli eventi incompatibili la probabilità dell'evento intersezione (ovvero che si verifichino entrambi gli eventi) è nulla.

$$P(E_1\cap E_2)=0$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1\cup E_2)=P(E_1)+P(E_2)-P(E_1\cap E_2)=0.5+0.2-0=0.7 \qquad P(E_2|E_1)=P\frac{(E_1\cap E_2)}{P(E_1)}=\frac{0}{0.5}=0$$

b) Il candidato indichi se i due eventi E_1 ed E_2 possono ritenersi dipendenti.

Due eventi sono dipendenti se il verificarsi di un evento modifica la probabilità che si verifichi l'altro. Poiché gli eventi sono incompatibili sono ovviamente dipendenti. Infatti, il verificarsi di un evento azzera la probabilità del verificarsi dell'altro.

- Appello del 05 Settembre 2012 -

Esercizio 1)
Si vuole monitorare lo sforzo percepito da un atleta durante sequenza di 11 esercizi.Tali esercizi richiedono uno sforzo crescente da parte dell'atleta. La misurazione avviene nel momento in cui sono stati svolti 3/4 dell'esercizio e si chiede all'atleta di esprimere una della seguenti valutazioni dello sforzo:(N) nullo, (L) leggero, (M) moderato, (I) intenso, (MI) molto inteso, (IN) insostenibile. Nella prima seduta si sono ottenute le seguenti misurazioni:

N L L L M I I M I MI MI

a) Determini la tipologia del carattere.
b) Fornisca una rappresentazione grafica dei dati.
c) Si indichino e si calcolino tutti gli indici di posizioni adeguati ai dati.
d) Si indichino gli indici di variabilità adeguati ai dati e, se possibile, se ne calcoli uno.

Esercizio 2)
Il soggetto descritto nell'Esercizio 1 ha ripetuto il test dopo un mese di allenamento ottenendo le seguenti misurazioni:

N L L L L M M I I MI MI

Il candidato,
a) realizzi una tabella a doppia entrata che raccolga i dati ricavati nel primo e nel secondo Esercizio (prima e seconda sessione di allenamento);
b) se possibile, indichi e calcoli un opportuno indice di posizione per la serie bivariata;
c) se possibile, indichi e calcoli un opportuno indice di variabilità per la serie bivariata;
d) verifichi se l'allenamento ha inciso in maniera significativa sulla fatica percepita dall'atleta. Il candidato indichi le necessarie ipotesi e proceda al calcolo anche se queste non fossero soddisfatte.

Esercizio 3)
Nello scenario tratteggiato nell'Esercizio 1, si vuole fornire un indicatore numerico della fatica percepita pertanto si sostiuiscono alle modalità il rispettivo numero d'ordine. (N => 1, L => 2, ... , IN =>6). Il candidato, stimi per intervallo il valore atteso della fatica percepita dall'atleta durante la prima sessione di allenamento (descritta nell'Eserczio 1). Il candidato evidenzi le ipotesi necessarie e proceda al calcolo anche se queste risultassero non verificate.

Esercizio 4)
Si considerino i seguenti eventi:

E_1: si abbia $y = 1$ estraendo da una $Ber(0.4)$
E_2: si abbia $z < 0$ dove z è estratto da una normale standardizzata.

Il candidato, sapendo che la probabilità che gli eventi si verifichino contemporaneamente è del 20%,

a) calcoli le seguenti Probabilità: $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$; $P(E_1 \mid E_2)$; $P(E_2 \mid E_1)$;
b) indichi se gli eventi E_1 ed E_2 possono ritenersi dipendenti.

- Appello del 5 Settembre 2012 -
Svolgimento

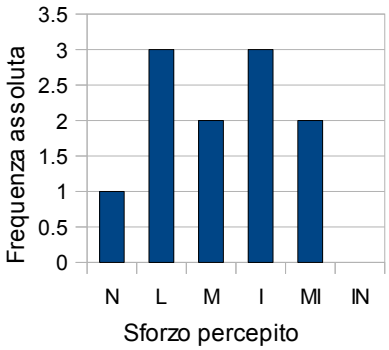
Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo qualitativo (in quanto non espresso da numeri ma da etichette) ordinabile (in quanto è possibile ordinare le modalità in maniera oggettiva N < L < M <I < MI <IN).

b) Fornisca una rappresentazione grafica adeguata dei dati .

Per un carattere qualitavo ordinabile una rappresentazione dei dati idonea può essere il diagramma a barre. Questo diagramma è ottenuto ponendo sulle ascisse di un piano cartesiano le modalità delle osservazioni e disegnando per ogni modalità un rettangolo la cui altezza è pari alla relativa frequenza assoluta. A lato si mostra il diagramma ottenuto dai dati in oggetto.



c) Si indichino e si calcolino tutti gli indici di posizioni adeguati ai dati.

Gli indici di posizione visti nel corso sono tre: media, moda e mediana. Nel caso in esame è possibile calcolare solo gli ultimi due. La moda è data dalla modalità avente la più alta frequenza assoluta. Nel caso in esame la moda non è unica: si hanno infatti due modalità a frequenza maggiore: L ed I (si parla di distribuzione *bi-modale*). La mediana è invece l'osservazione che bipartisce i dati ordinati. Avendo 11 osservazioni la mediana sarà la sesta (essa è preceduta da 5 osservazioni e seguita da 5 osservazioni) pertanto ordinano le osservazioni

N L L L M M I I I MI MI

Si evince che la mediana è M.

d) Si indichino gli indici di variabilità adeguati ai dati e, se possibile, se ne calcoli uno.

Nel caso di caratteri qualitativi non è possibile introdurre il concetto di varibilità.

Esercizio 2)

a) Realizzi una tabella a doppia entrata che raccolga i dati ricavati nel primo e nel secondo Esercizio (effetti collaterali in pazienti sani e non).

Nei due esercizi si descrive una bivariata che raccoglie i seguenti caratteri

- *X*: sessione di allenamento (carattere qualitativo ordinabile)
- *Y*: sensazione di fatica (carattere qualitativo ordinabile)

Le cui osservazioni possono essere raccolte nella seguente tabella a doppia entrata.

		Y: sensazione di fatica					Totali
		N	L	M	I	MI	
X: sessione di allenamento	Prima	1	3	2	3	2	11
	Seconda	1	4	2	2	2	11
	Totali	2	7	4	5	4	22

b) Se possibile, indichi e calcoli un opportuno indice di posizione per la serie bivariata

Per serie bivariate con almeno un carattere quantitativo, l'unico indice di posizione possibile è la moda ovvero la modalita corrispondente alla frequenze maggiore. Nel nostro caso la frequenza maggiore è 12, pertanto la moda è:

(Seconda; L)

c) Se possibile, indichi e calcoli un opportuno indice di variabilità per la serie bivariata

Non è possibile, con le tecniche viste nel corso, calcolare indici di variabilità per serie bivariate con almeno un carattere quantitativo.

d) verifichi se l'allenamento ha inciso in maniera significativa sulla fatica percepita dall'atleta. Il candidato indichi le necessarie ipotesi e proceda al calcolo anche se queste non fossero soddisfatte.

Se vi fosse un legame fra i due caratteri questi dovrebbero essere dipendenti; pertanto come primo passo si valuta la loro indipendenza. Per verificare se i due caratteri sono indipendenti 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. Il test viene fatto

sfruttando la distribuzione limite dello stimatore di Pizzetti Pearson che viene ad essere un chi quadranto avente gradi di libertà pari a quelli del 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 .

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

le frequenze teoriche sono state riportate fra parentesi nella seguente tabella

		Y: sensazione di fatica					Totali
		N	L	M	I	MI	
X: sessione di allenamento	Prima	1 (1)	3 (3.5)	2 (2)	3 (2.5)	2 (2)	11
	Seconda	1 (1)	4 (3.5)	2 (2)	2 (2.5)	2 (2)	11
	Totali	2	7	4	5	4	22

A questo punto è possibile valutare la convergenza dell stimatore di Pizzetti Pearson mediante la verifica dell'unica ipotesi vista a tal riguardo ovvero che tutte le frequenze teoriche siano superiori a 5.
Constatato che la condizione non è verificata ci si dovrebbe fermare nella valutazione. Nonostante ciò il testo richiede di procedere nella procedura. Pertanto, si procede al calcolo della regione di accettazione dopo aver fissato il livello di significatività (posto nel nostro caso al 1%).

$$A = \left[0\,;\chi^2_{1-\alpha}\left((M_x-1)(M_y-1)\right)\right] = \left[0\,;\chi^2_{1-0.01}\left((2-1)(5-1)\right)\right] = \left[0\,;\chi^2_{0.99}(4)\right] = [0;\,13.28]$$

Si può ora procedere al calcolo dello stimatore vero e proprio

$$\frac{\sum_{i=1}^{2}\sum_{j=1}^{4}\left(n_{i,j}-\hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} =$$

$$= \frac{(1-1)^2}{1} + \frac{(3-3.5)^2}{3.5} + \frac{(2-2)^2}{2} + \frac{(3-2.5)^2}{2.5} + \frac{(2-2)^2}{2} + \frac{(1-1)^2}{1} + \frac{(4-3.5)^2}{3.5} + \frac{(2-2)^2}{2} + \frac{(2-2.5)^2}{2.5} + \frac{(2-2)^2}{2} =$$

$$= 0+0.07+0+0.1+0+0+0.07+0+0.1+0=0.34$$

Poichè il valore dello stimatore è interno all'intervallo di accettazione posso dire che i due caratteri sono indipendenti ad un livello di significatività del 1%. Quindi si può scartare l'ipotesi che esista un influenza fra l'allenamento e la fatica percepita dal soggetto.

Esercizio 3)

Le tecniche di stima viste nel corso prevedono che:

- la popolazione sia descrivibile mediante una variabile casuale,
- che il campione abbia una numerosità tale da far convergere lo stimatore e
- che le prove siano indipendenti ed identicamente distribuite (i.i.d.).

Nel caso in esame

- descrivere l'esperimento mediante la seguente variabile casuale *X: fatica perceita durante lo svolgimento di un esercizio*.
- la grandezza da stimare risulta *E[X]* il cui stimatore è la media campionaria la quale converge in legge per campioni avente numerosità superiore a 30 (ipotesi non confermata, si hanno infatti 11 estrazioni in ambo i casi).
- L'ipotesi di prove i.i.d. è un molto debole in quanto si suppone che la distribuzione della v.c. di avere un non cambi attraverso le prove (sappiamo che gli esercizi sono fra di loro a difficoltà crescente).

Per effettuare una stima per intervallo si deve, come prima cosa, fissare un livello di confidenza: nel nostro caso 95% (α=0.05).Avendo così definito la tipologia di stima (stima per intervallo al 95%), si ha che essa è data dalla seguente formula

$$E[\hat{P}] \in \left[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sqrt{Var[X]}}{N}\,;\bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sqrt{Var[X]}}{N}\right]$$

Dove il valore della normale standardizzata si ottiene dalle tavole mentre la varianzia della popolazione si stima puntualmente come illustrato nel seguito.

- $z_{1-\alpha/2}$: con questa notazione si intende il valore di *z* che lasci alla sua sinistra una probabilità (data dall'area sottesa dalla d.d.p) pari ad *1-α/2*.

$$\int_{-\infty}^{z_{1-\alpha/2}} f(x)\,dx = 1-\alpha/2$$

Ricodando che nelle tavole sono graficati le aree sottese dalla normale standardizzata fra 0 ed un valore positivo di *Z*, dobbiamo trovare un modo per rincodurci all'uso di questa tipologia di integrali. Questo può essere fatto spezzando l'integrale in due: fra meno infinito e zero e fra zero e $z_{1-\alpha/2}$. In simboli:

$$\int_{-\infty}^{z_{1-\alpha/2}} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_{0}^{z_{1-\alpha/2}} f(x)dx = 1-\alpha/2$$

Elaborando gli ultimi due membri l'equazione si ottiene il seguente risultato

$$\int_{0}^{z_{1-\alpha/2}} f(x)dx = -\int_{-\infty}^{0} f(x)dx + 1-\alpha/2 = 0.5-\alpha/2 = 0.5-0.05/2 = 0.475$$

Pertanto il valore $z_{1-\alpha/2}$ è quello a cui sulle tavole corrisponde l'area di 0.475; ottenendo

$$z_{\alpha/2} = 1.96$$

- *Stima della varianza*. La varianza viene stimata utilizzando il suo stimatore corretto: la varianza campionaria s^2. Essa viene ricavata utilizzando le sguente formula in cui i dati sono estratti dalla Tabella 1

$$s^2 = \sigma^2 \frac{N}{N-1} = (\sum_{i=1}^{M} f_i m_i^2 - \bar{m}) \frac{N}{N-1} = \left(\sum_{i=1}^{M} f_i m_i^2 - \left(\sum_{i=1}^{M} f_i m_i \right)^2 \right) \frac{N}{N-1} = (11.73 - 3.18^2) \frac{11}{10} = 1.76$$

Infine si ottiene la stima richiesta:

$$E\hat{[}P] \in \left[\bar{x} - \sqrt{\frac{Var[X]}{N}} ; \bar{x} + z_{\frac{\alpha}{2}} \sqrt{\frac{Var[X]}{N}} \right] = \left[3.18 - 1.96\sqrt{\frac{1.76}{11}} ; 3.18 + 1.96\sqrt{\frac{1.76}{11}} \right] = [2.40; 3.96]$$

Fatica percepita modalità m_i	frequenze assolute n_i	frequenze relative f_i	$m_i * f_i$	m_i^2	$m_i^2 * f_i$
1	1	0.0909	0.09	1	0.0909
2	3	0.2727	0.5455	4	1.0909
3	2	0.1818	0.5455	9	1.6364
4	3	0.2727	1.0909	16	4.3636
5	2	0.1818	0.9091	25	4.5455
6	0	0.0000	0.0000	36	0.0000
Totali	11		3.1818		11.7273

Tabella 1) Dati Esercizio 3 ottenuti convertendo i dati del'Esercizio 1 in modalita numeriche.

Esercizio 4)

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

La probabilità dell'evento E_1 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 distribuzioni di Bernoulli di parametro *p* (ovvero *Ber(p)*) sono possibili solo due esiti (0 ed 1) e che il parametro indica il la probabilità dell'esito positivo si ottiene immediatamente che:

$$P(E_1) = P(Y=1) = p = 0.4$$

La probabilità dell'evento E_2 è pari alla probabilità di avere un esito maggiore di zero da una normale standardizzata. Questo probabilità è pari all'area sottesa della d.d.p. della normale standardizzata fra 0 e + infinito. Questa area può essere facilmente ricavata ricordando che la d.d.p sottende area unitaria (come tutte le d.d.p.) ed simmetrica e centrata nell'origine. Pertanto si puo dire che il valore atteso (ovvero l'origine) bipartisca l'area sottesa. Pertanto:

$$P(E_2) = P(Z>0) = 1/2 = 0.5$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.5 + 0.4 - 0.2 = 0.7 \qquad P(E_2|E_1) = P\frac{(E_1 \cap E_2)}{P(E_1)} = \frac{0.2}{0.5} = 0.4$$

b) Il candidato indichi se i due eventi E1 ed E2 possono ritenersi dipendenti.

Due eventi sono dipendenti se il verificarsi di un evento modifica la probabilità che si verifichi l'altro. Nel caso in esame questo non succede, pertanto possiamo ritenere i due eventi statisticamente indipendenti.

- Appello del 19 Settembre 2012 -

Esercizio 1)
Un mese prima di una elezione è stato chiesto ad un campione di 35 maggiorenni di esprimere una intenzione di voto in cui le risposte possibili sono (D = destra, C = centro, S= sinistra, R = non andrò a votare). Le osservazioni ottenenute sono le seguenti:

D	C	R	R	R	D	D	D	R	D
C	S	D	S	S	S	C	S	D	S
D	S	D	C	S	S	C	C	S	D
R	D	S	S	R					

Il candidato

a) determini la tipologia del carattere;
b) fornisca una rappresentazione grafica dei dati;
c) indichi e calcoli tutti gli indici di posizioni adeguati ai dati;
d) indichi gli indici di variabilità adeguati ai dati e, se possibile, se ne calcoli uno.

Esercizio 2)
Due giorni prima delle elezioni si è ripetuto il sondaggio ottenendo le seguenti misurazioni:

	Destra	Sinistra	Centro	Non Voto	Totale
Fequenza	13	14	6	2	35

Il candidato,

a) realizzi una tabella a doppia entrata che raccolga i dati ricavati nei sondaggi descritti negli Esercizi1 e 2;
b) se possibile, indichi e calcoli un opportuno indice di posizione per la serie bivariata e per le distribuzioni marginali;
c) se possibile, indichi e calcoli un opportuno indice di variabilità per la serie bivariata;
d) stabilista ad un livello di significatività del 1% se i risultati dei due sondaggi presentano differenze statisticamente significative. Il candidato indichi le necessarie ipotesi e proceda al calcolo anche se queste non fossero soddisfatte.

Esercizio 3)
Nel sondaggio descritto negli Es. 1 e 2 si è rilevata anche l'eta degli intervistati ottenendo la distribuzione a classi:

	da 19 a 25	da 25 a 35	da 35 a 55	da 55 a 85	Totale
Fequenza	10	25	15	20	70

Il candidato, stimi per intervallo il valore atteso dell'età dei votanti. Il candidato evidenzi le ipotesi necessarie e proceda al calcolo anche se queste risultassero non verificate.

Esercizio 4)
Si considerino i seguenti eventi:

E_1: si abbia $Y = 1$ estraendo da una v.c. $Y \sim Ber(0.4)$
E_2: estraendo un votante dal campione descritto nell'Esercizio 2 questo voti a sinistra.

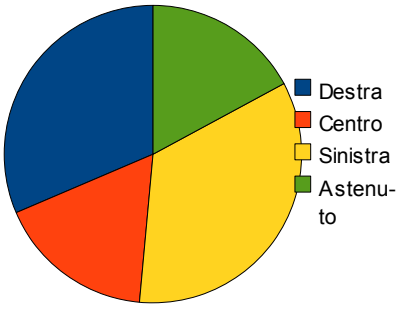
Il candidato, sapendo che gli eventi sono indipendenti, calcoli le seguenti probabilità:
$P(E_1); P(E_2); P(E_1 \cup E_2); P(E_1 \mid E_2); P(E_2 \mid E_1)$.

- Appello del 19 Settembre 2012 -
Svolgimento

Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo qualitativo (in quanto non espresso da numeri ma da etichette) non ordinabile (in quanto è non possibile ordinare le modalità in maniera oggettiva).



b) Fornisca una rappresentazione grafica adeguata dei dati .

Per un carattere qualitavo non ordinabile una rappresentazione dei dati idonea può essere il diagramma a barre o il diagramma a torta. Quest'ultimo è ottenuto dividendo una sulle ascisse di un piano cartesiano le modalità delle osservazioni e disegnando per ogni modalità un rettangolo la cui altezza è pari alla relativa frequenza assoluta. A lato si mostra il diagramma ottenuto dai dati in oggetto.

c) Si indichino e si calcolino tutti gli indici di posizioni adeguati ai dati.

Gli indici di posizione visti nel corso sono tre: media, moda e mediana. Nel caso in esame è possibile calcolare solo la moda. La moda è data dalla modalità avente la più alta frequenza assoluta. Nel caso in esame la moda è unica ed è rappresentata dalla modalità "Sinistra" avente 12 osservazioni.

d) Si indichino gli indici di variabilità adeguati ai dati e, se possibile, se ne calcoli uno.

Nel caso di caratteri qualitativi non è possibile introdurre il concetto di varibilità.

Esercizio 2)

a) realizzi una tabella a doppia entrata che raccolga i dati ricavati nei sondaggi descritti negli Esercizi1 e 2

Nei due esercizi si descrive una bivariata che raccoglie i seguenti caratteri

- *X*: sondaggio (carattere qualitativo ordinabile)
- *Y*: intenzione di voto (carattere qualitativo non ordinabile)

Le cui osservazioni possono essere raccolte nella seguente tabella a doppia entrata.

		Y: intenzione di voto				
		Destra	**Centro**	**Sinistra**	**Astenuto**	**Totali**
X: sondaggio	**primo**	11	6	12	6	35
	secondo	13	6	14	2	35
	Totali	24	12	26	8	70

b) Se possibile, indichi e calcoli un opportuno indice di posizione per la serie bivariata

Per serie bivariate con almeno un carattere quantitativo, l'unico indice di posizione possibile è la moda ovvero la modalita corrispondente alla frequenze maggiore. Nel nostro caso la frequenza maggiore è 12, pertanto la moda è:

(Seconda; Sinistra)

c) Se possibile, indichi e calcoli un opportuno indice di variabilità per la serie bivariata

Non è possibile, con le tecniche viste nel corso, calcolare indici di variabilità per serie bivariate con almeno un carattere quantitativo.

d) stabilista ad un livello di significatività del 1% se i risultati dei due sondaggi presentano differenze statisticamente significative. Il candidato indichi le necessarie ipotesi e proceda al calcolo anche se queste non fossero soddisfatte.

Se vi fossero differenze statistiche significative vorrebbe dire che un legame sarebbe ipotizzabile fra l'appartenenza ad un sondanggio (valore della v.c. X) e l'esito del sondaggio (v.c. Y). Pertanto è lecito affermare che se vi è statistica differenza fra il i due sondaggi allora è le due variabili casuali non sono indipendenti (ossia un legame sussiste fra i due caratteri). Pertanto una possibile risposta alla domanda viene data valutando l'indipendenza fra i caratteri della bivariata. Per verificare se i due caratteri sono indipendenti 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 che risulta essere l'ipotesi nulla del test. Il test viene fatto sfruttando la distribuzione limite dello stimatore di Pizzetti

Pearson che viene ad essere un chi quadranto avente gradi di libertà pari a quelli del 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 .

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

le frequenze teoriche sono state riportate fra parentesi nella seguente tabella

		Y: intenzione di voto				
		Destra	**Centro**	**Sinistra**	**Astenuto**	**Totali**
X: sondaggio	**primo**	11 (12)	6 (6)	12 (13)	6 (4)	35
	secondo	13 (12)	6 (6)	14 (13)	2 (4)	35
	Totali	24	12	26	8	70

A questo punto è possibile valutare la convergenza dell stimatore di Pizzetti Pearson mediante la verifica dell'unica ipotesi vista a tal riguardo ovvero che tutte le frequenze teoriche siano superiori a 5.
Constatato che la condizione non è verificata ci si dovrebbe fermare nella valutazione. Nonostante ciò il testo richiede di procedere nella procedura. Pertanto, si procede al calcolo della regione di accettazione dopo aver fissato il livello di significatività (posto nel nostro caso al 1%).

$$A = \left[0\,;\chi^2_{1-\alpha}\left((M_x-1)(M_y-1)\right)\right] = \left[0\,;\chi^2_{1-0.01}\left((2-1)(4-1)\right)\right] = \left[0\,;\chi^2_{0.99}(3)\right] = [0\,;11.341]$$

Si può ora procedere al calcolo dello stimatore vero e proprio

$$\frac{\sum_{i=1}^{2}\sum_{j=1}^{4}\left(n_{i,j}-\hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} =$$

$$= \frac{(11-12)^2}{12} + \frac{(6-6)^2}{6} + \frac{(12-13)^2}{13} + \frac{(6-4)^2}{4} + \frac{(13-12)^2}{12} + \frac{(6-6)^2}{6} + \frac{(14-13)^2}{13} + \frac{(2-4)^2}{4} =$$

$$= \frac{1}{12} + 0 + \frac{1}{13} + 1 + \frac{1}{12} + 0 + \frac{1}{13} + 1 = \frac{1}{6} + \frac{2}{13} + 2 = 2.32$$

Poichè il valore dello stimatore è interno all'intervallo di accettazione posso dire che i due caratteri sono indipendenti ad un livello di significatività del 1%. Quindi si può asserire che non vi sia una differenza statisticamente significativa fra i due campioni.

Esercizio 3)

Le tecniche di stima viste nel corso prevedono che:

- la popolazione sia descrivibile mediante una variabile casuale,
- che il campione abbia una numerosità tale da far convergere lo stimatore e
- che le prove siano indipendenti ed identicamente distribuite (i.i.d.).

Nel caso in esame

- descrivere l'esperimento mediante la seguente variabile casuale *X: età degli intervistati durante un sondaggio composto da due sessioni*.
- la grandezza da stimare risulta *E[X]* il cui stimatore è la media campionaria la quale converge in legge per campioni avente numerosità superiore a 30 (ipotesi confermata, si hanno infatti 70 soggetti selezionati).
- L'ipotesi di prove i.i.d. è solida in quanto si suppone che la distribuzione della v.c. di avere un non cambi attraverso le prove (si suppone che la probabilità che un soggetto entri nel campione sia fondamentalmente inalterata nelle varie estrazioni).

Per effettuare una stima per intervallo si deve, come prima cosa, fissare un livello di confidenza: nel nostro caso 95% (α=0.05).Avendo così definito la tipologia di stima (stima per intervallo al 95%), si ha che essa è data dalla seguente formula

$$E[\hat{P}] \in \left[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sqrt{Var[X]}}{N}\,;\bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sqrt{Var[X]}}{N}\right]$$

Dove il valore della normale standardizzata si ottiene dalle tavole mentre la varianzia della popolazione si stima puntualmente come illustrato nel seguito.

- $z_{1-\alpha/2}$: con questa notazione si intende il valore di *z* che lasci alla sua sinistra una probabilità (data dall'area sottesa dalla d.d.p) pari ad *1-α/2*.

  $$\int_{-\infty}^{z_{1-\alpha/2}} f(x)\,dx = 1-\alpha/2$$

  Ricodando che nelle tavole sono graficati le aree sottese dalla normale standardizzata fra 0 ed un valore

positivo di *Z*, dobbiamo trovare un modo per rincodurci all'uso di questa tipologia di integrali. Questo può essere fatto spezzando l'integrale in due: fra meno infinito e zero e fra zero e $z_{1-\alpha/2}$. In simboli:

$$\int_{-\infty}^{z_{1-\alpha/2}} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{z_{1-\alpha/2}} f(x)\,dx = 1-\alpha/2$$

Elaborando gli ultimi due membri l'equazione si ottiene il seguente risultato

$$\int_{0}^{z_{1-\alpha/2}} f(x)\,dx = -\int_{-\infty}^{0} f(x)\,dx + 1-\alpha/2 = 0.5-\alpha/2 = 0.5-0.05/2 = 0.475$$

Pertanto il valore $z_{1-\alpha/2}$ è quello a cui sulle tavole corrisponde l'area di 0.475; ottenendo

$$z_{\alpha/2} = 1.96$$

- *Calcolo della media*. Quando si ha adisposizione una statistica organizzata per classi per procedere al calcolo della media si può associare ad ogni classe il proprio valore centrale. Questo diviene in sostanza la modalità cui si riferscono le frequenze (sia relative che assolute). A questo punto si procede normalmente come mostrato in Tabella 1 ottenendo una media di 43.5.
- *Stima della varianza*. La varianza viene stimata utilizzando il suo stimatore corretto: la varianza campionaria s^2. Essa viene ricavata utilizzando le sguente formula in cui i dati sono estratti dalla Tabella 1

$$s^2 = \sigma^2 \frac{N}{N-1} = (\sum_{i=1}^{M} f_i m_i^2 - \bar{m}) \frac{N}{N-1} = \left(\sum_{i=1}^{M} f_i m_i^2 - \left(\sum_{i=1}^{M} f_i m_i\right)^2\right) \frac{N}{N-1} = (2224.5 - 43.5^2)\frac{70}{69} = 337.07$$

Infine si ottiene la stima richiesta:

$$E[\hat{P}] \in \left[\bar{x} - \sqrt{\frac{Var[X]}{N}}\,;\, \bar{x} + z_{\frac{\alpha}{2}}\sqrt{\frac{Var[X]}{N}}\right] = \left[43.5 - 1.96\sqrt{\frac{337.07}{70}}\,;\, 43.5 + 1.96\sqrt{\frac{337.07}{70}}\right] = [41.31\,;\, 45.69]$$

Classe	Valore Centrale c_i	frequenze assolute n_i	frequenze relative f_i	$m_i * f_i$	m_i^2	$m_i^2 * f_i$
19 – 25	22	10	0.1429	3.14	484	69.1429
25 – 35	30	25	0.3571	10.7143	900	321.4286
35 – 55	45	15	0.2143	9.6429	2025	433.9286
55 – 85	70	20	0.2857	20.0000	4900	1400.0000
Totali		70		43.5000		2224.5000

Tabella 1) Dati Esercizio 3.

Esercizio 4)

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

La probabilità dell'evento E_1 è pari alla probabilità di avere un esito positivo (ovvero pari ad uno) da distribuzione di Bernoulli. Ricordando che una generica distribuzioni di bernoulli di parametro *p* (ovvero *Ber(p)*) sono possibili solo due esiti (0 ed 1) e che il parametro indica il la probabilità dell'esito positivo si ottiene immediatamente che:

$$P(E_1) = P(Y=1) = p = 0.4$$

La probabilità dell'evento E_2 è pari alla probabilità di estrarre dai 35 soggetti del campione dell'esercizio 2 uno dei 14 votanti di sinistra. Ricordando che la probabilità in caso di eventi equiprobabili è data dal rapporto casi favorevi su casi totatli si ha che:

$$P(E_2) = 14/35 = 2/5 = 0.4$$

Anche se non richiesto dal testo, è utile calcolare la probabilità dell'evento intersezione che per eventi indipendenti è data dalla seguente

$$P(E_1 \cap E_2) = P(E_1)P(E_2) = 0.4^2 = 0.16$$

Noto ciò, le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

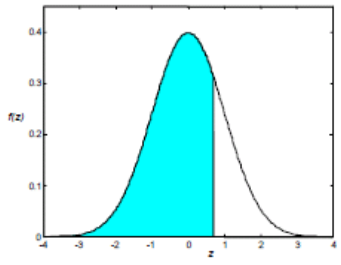
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.4 + 0.4 - 0.16 = 0.64$$

$$P(E_2|E_1) = P\frac{(E_1 \cap E_2)}{P(E_1)} = \frac{0.16}{0.4} = 0.4$$

$$P(E_1|E_2) = P\frac{(E_1 \cap E_2)}{P(E_2)} = \frac{0.16}{0.4} = 0.4$$

Tavola I - Distribuzione normale standardizzata

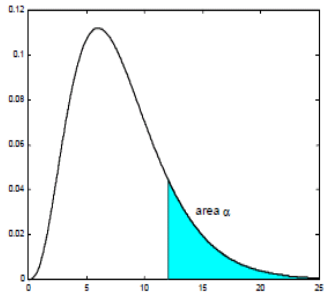
La tavola fornisce i valori sottesi dalla dstribuzione di probabilita della normale standardizzata *f(z)* da -∞ a *z*.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	0.99995	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tavola II - Distribuzione χ^2

La tabella fornisce i valori di $\chi^2_\alpha(\nu)$ per i quali $P(\chi^2 > \chi^2_\alpha(\nu)) = \alpha$ per alcuni valori notevoli della probabilità α e dei gradi di libertà ν .



ν	α = 0.995	α = 0.99	α = 0.975	α = 0.95	α = 0.05	α = 0.025	α = 0.01	α = 0.005	ν
1	0.0000393	0.000157	0.000982	0.00393	3.841	5.024	6.635	7.879	1
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597	2
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750	5
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558	24
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30
40	20.706	22.164	24.433	26.509	55.758	59.342	63.691	66.766	40
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490	50
60	35.535	37.485	40.482	43.188	79.082	83.298	88.379	91.952	60
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215	70
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321	80
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299	90
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169	100