Matematica & Statistica:Modulo di Statistica

-Appello del 9 Settembre 2015 -

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)
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)
Nello scenario descritto nell'Esercizio 1, si vuole fornire un indicatore numerico della fatica percepita; pertanto si sostituiscono alle modalità il rispettivo numero d'ordine.

(N → 1, L → 2, ... , IN → 6).

Il candidato,
a) Indichi cosa cambia a livello di indici sintetici di posizione e di variabilità.
b) stimi per intervallo con una confidenza del 90% il valore atteso della fatica percepita dall'atleta durante la prima sessione di allenamento (descritta nell'Esercizio 1). Il candidato evidenzi le ipotesi necessarie e proceda al calcolo anche se queste risultassero non verificate.

Esercizio 4)
Si considerino i seguenti eventi dichiarati indipendenti.

A: si ottenga $x > -10$ dove x è estratto da una v. c. distribuita come un $\chi^2(2)$.

B: si ottenga y>2 dove y è estratto da una v. c. *Unif(-5;5).*

a) Il candidato calcoli le seguenti Probabilità: $P(E_1)$; $P(\overline{B})$; $P(B)$; $P(A \cup B)$.
b) Il candidato fornisca la definizione dei seguenti eventi notevoli: eventi statisticamente dipendenti, evento certo ed evento impossibile, evento complementare.

- Appello del 9 Settembre 2015 -

Svolgimento

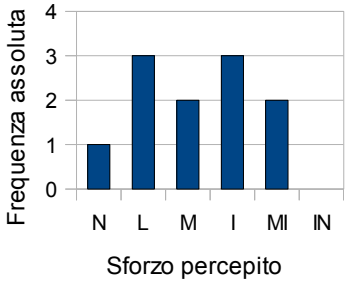
Esercizio 1)

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 qualitativo 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 variabilità.

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 riportati 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
x_i	5	6	6	7	7	8	6.5000
y_i	71	75	70	75	82	80	75.5000
$x_i - \bar{x}$	-1.5	-0.5	-0.5	0.5	0.5	1.5	
$y_i - \bar{y}$	-4.5	-0.5	-5.5	-0.5	6.5	4.5	
$(y_i - \bar{y})(x_i - \bar{x})$	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

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

Poiché 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 previsioni effettuate dal modello lineare che descrivono l'andamento dei dati con buona precisione.

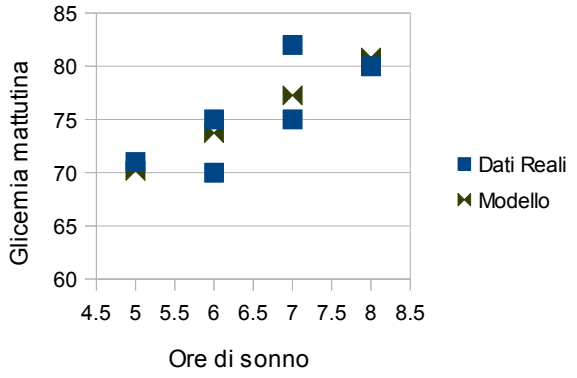


Figura 1) Rappresentazione dei dati Esercizio 2

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 una glicemia prevista 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) *Punto a)*
L'utilizzo di valori numerici consente il calcolo della media come indice di posizione.

$$\bar{m} = \sum_{i=1}^{6} m_i * f_i = 11.7273$$

Inoltre è possibile calcolare anche indici di variabilità come la varianza che risulta essere:

$$\sigma^2 = \left(\sum_{i=1}^{M} f_i m_i^2 - \bar{m}\right) = \left(\sum_{i=1}^{6} f_i m_i^2 - \left(\sum_{i=1}^{6} f_i m_i\right)^2\right) = \left(11.73 - 3.18^2\right) = 1.6176$$

Nel calcolo degli indici si sono utilizzati i conti riportati in tabella.

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 dell'Esercizio 1 in modalità numeriche.

Punto b)
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).

Il testo richiede di effettuare una stima al 92% ($\alpha = 0.08$). 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 varianza 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*.

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

  Ricordando che in alcune tavole sono riportati i valori delle le aree sottese dalla normale standardizzata fra 0 ed un valore positivo di *Z*, dobbiamo trovare un modo per ricondursi 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

  $$\int_{0}^{z_{1-\alpha/2}} f(x)\,dx = -\int_{-\infty}^{0} f(x)\,dx + 1 - \alpha/2 = 0.5 - \alpha/2 = 0.5 - 0.1/2 = 0.46$$

  Pertanto il valore $z_{1-\alpha/2}$ è quello a cui sulle tavole corrisponde l'area di 0.46; ottenendo

  $$z_{\alpha/2} = 1.75$$

- *Stima della varianza*. La varianza viene stimata utilizzando il suo stimatore corretto: la varianza campionaria s^2. Questa si ricavata applicando la seguente formula:

  $$s^2 = \sigma^2\frac{N}{N-1} = 1.6176\frac{11}{10} = 1.76$$

Infine si ottiene la stima richiesta:

$$E\hat{[}P]\in\left[\bar{x}-\sqrt{\frac{Var[X]}{N}};\bar{x}+z_{\frac{\alpha}{2}}\sqrt{\frac{Var[X]}{N}}\right]=\left[3.18-1.75\sqrt{\frac{1.76}{11}};3.18+1.75\sqrt{\frac{1.76}{11}}\right]=[2.48;3.88]$$

Esercizio 4)
a) Il candidato calcoli le seguenti Probabilità: P(A); P(E₂); P(E₁ U E₂) P(E₁ | E₂).
Applicando la definizione di densità di probabilità per ottenere di *P(A)*, si deve calcolare l'area sottesa dalla curva della densità di probabilità di un chi quadro a *2* gradi di libertà da -10 a più infinito.
Lo stesso risultato si ottiene considerando che la v.c. chi quadro per ogni possibile grado di libertà assume solo valori maggiori di zero pertanto essi saranno sempre maggiori di -10. Quindi la probabilità richiesta è pari a quella dell'evento certo.

La probabilità dell'evento *B* è pari alla probabilità di estrarre un numero *y* maggiore di due da una v. c. uniforme fra *-5* ed *5*. In questo caso è possibile ricavare la d.d.p. della v.c. *Y*. Basta imporre che l'area sottesa dalla d.d.p. sia unitaria. Pertanto si ha che

$$\int_{-\infty}^{+\infty} f_Y(\tau)\,d\tau = 1$$

poiché la d.d.p. è nulla per valori esterni all'intervallo *[-5; 5]* la predente diviene

$$\int_{-5}^{+5} f_Y(\tau)\,d\tau = 1$$

ricordando che la d.d.p. è costante si ha che

$$\int_{-5}^{+5} C\,d\tau = (5-(-5))C = 10C = 1$$

da cui si ha che C= 1/10= 0.1. Il calcolo della probabilità richiesta diviene ora facile

$$P(E_2)=P(y>2)=\int_2^5 \frac{1}{10} d\tau=(5-2)\frac{1}{10}=0.3$$

Propedeutico al calcolo delle altre due probabilità e il calcolo della probabilità dell'evento intersezione (ovvero che i due eventi si verifichino contemporaneamente). Per eventi indipendenti essa è il prodotto delle due probabilità ovvero

$$P(E_1 \cap E_2)=P(E_1)P(E_2)=1\cdot 0.3=0.3$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=1+0.3-0.3=1 \qquad P(\bar{B})=1-P(B)=1-0.3=0.7$$

b) Il candidato fornisca la definizione dei seguenti eventi notevoli: eventi statisticamente indipendenti, evento certo ed evento impossibile.

Due eventi si dicono

- *statisticamente indipendenti* se il verificarsi di un evento non altera la probabilità del verificarsi dell'altro. Se A e B sono due eventi statisticamente indipendenti si ha che

  $$P(A|B) = P(A) \qquad P(B|A) = P(B)$$

- *certo*: un evento certo è un evento che si verifica sempre. Se E è un evento certo si ha che *P(E) =1*.
- *impossibile*: un evento impossibile è un evento che non si verifica mai. Se E è un evento impossibile si ha che *P(E) =0*.