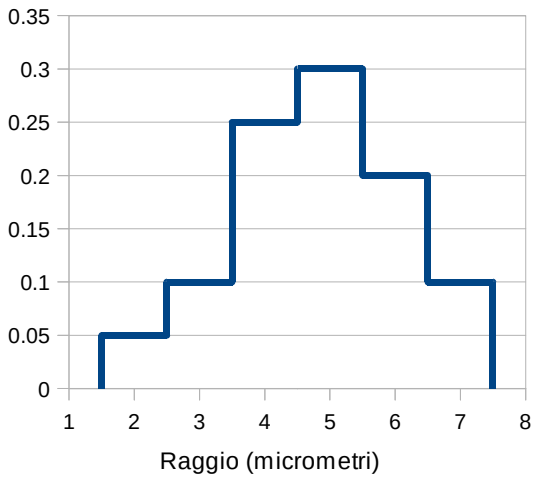
Matematica e Statistica: Modulo di Statistica - Prof. Federico Di Palma

- Appello del 12 Febbraio 2014 -

Esercizio 1)

In una ricerca si è interessati a verificare le dimensioni in micrometri di un granulocita neutrofilo. A tale scopo si sono misurati N=2000 campioni ottenendo la distribuzione (a classi) il cui istogramma è riportato 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.

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 2, stimi puntualmente e per intervallo il valore atteso della glicemia al risveglio 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 $x < 0$ dove x è distribuita come una normale avente $E[X] = 2$ e $Var[X]=4$
E_2: y = 1 dove y è distribuita come una binomiale con $n = 2$ e $p = 0.5$.

a) Il candiato 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 12 Febbraio 2014 -
Svolgimento

Esercizio 1)

a) Determinare la tipologia del carattere.

Il carattere è di tipo quantitativo (in quanto espresso da numeri) continuo (in quanto concettualmente una lunghezza come un raggio può assumere qualsiasi valore)

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

Per risovere questo punto, risulta opportuno esplicitare la relazionefra le grandezze riporatate nell'istogramma e le frequenze assolute. In un istogramma le ordinate riportano la densità di frequenza (d_i) delle classi mentre le ascisse gli estremi delle classi che compongono la distribuzione. . Ricordando che d_i è definita come il rapporto fra le frequenze relative (f_i) e l'ampiezza della classe cui sono riferite (sup_i - inf_i); e che le frequenze relative son il rapporto fra le frequenze assolute (n_i) ed il totale delle osservazioni (N), si ha:

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

Pertanto i dati richiesti possono essere ricavati applicando la formula testè ricavata. I risultati sono stati raccolti nella tabella ad entrata semplice (Tabella 1) riportata in calce.

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% (5*0 % - F_{i^*-1}*) 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^*} = 4.5) ricavando il valore di 4.83.

Lo stesso risultato poteva essere ottenuto applicando la seguente formula che riassume il procedimento appena descritto

$$Me = inf_{i^*} + (0.5 - F_{i^*-1}) / d_{i^*} = 4.5 + (0.5 - 0.4)/0.3 = 4.5 + 0.1/0.3 = 4.83$$

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

$$\sigma_x^2 = \left(\sum_{i=1}^{M} f_i * c_i^2\right) - \bar{x}^2 = \left(\sum_{i=1}^{M} f_i * c_i^2\right) - \left(\sum_{i=1}^{M} f_i * c_i\right)^2 = 24.7 - (4.8)^2 = 24.7 - 23.04 = 1.66$$

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	1.5	2.5	2	100	0.050	0.050	0.1000	4	0.2
2	2.5	3.5	3	200	0.100	0.150	0.3000	9	0.9
3	3.5	4.5	4	500	0.250	0.400	1.0000	16	4
4	4.5	5.5	5	600	0.300	0.700	1.5000	25	7.5
5	5.5	6.5	6	400	0.200	0.900	1.2000	36	7.2
6	6.5	7.5	7	200	0.100	1.000	0.7000	49	4.9
			Totali		1		4.8000		24.7000

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.

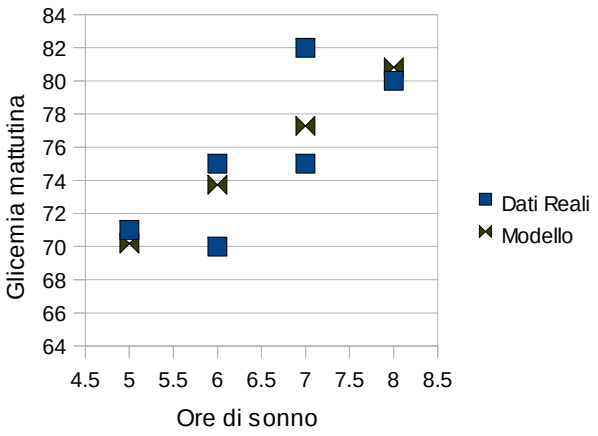


Figura 1) Rappresentazione dei dati dell Es 2

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 una glicemia prevista di 137.55 mg/dl.
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: glicemia al risveglio in un soggetto dibatico*.
- 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).
- L'ipotesi di prove i.i.d. è molto debole in quanto le prove essendo estratte dallo stesso soggetto saranno fortemente correllate. Questa considerazione rissume il fatto che difficilmente analizzando un solo soggetto è possibile trarre conclusioni su tutta la popolazione.

La stima puntuale si ottiene semplicemente dall'applicazione dello stimatore, pertanto ricordano quanto calcolato nell'esercizio precedente

$$\hat{E[X]} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 75.5$$

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

$$\hat{E}[X] \in \left[\bar{x} - z_{0.975}\sqrt{\frac{Var[X]}{n}} ; \bar{x} + z_{0.975}\sqrt{\frac{Var[X]}{n}} \right]$$

Dove il valore della normale si ricava dalle tavole:

$$z_{0.975} = 1.96$$

La varianza della popolazione non è nota pertanto essa viene stimata utilizzando la varianza campionaria. Ricordando i calcoli effettuati in precedenza si ha che:

$$\hat{Var[X]} = s^2 = \sigma^2 \frac{n}{n-1} = \frac{113.5}{6}\frac{6}{5} = \frac{113.5}{5} = 22.7$$

Infine si ottiene la stima richiesta:

$$\hat{E}[X] \in \left[75.5 - 1.96\sqrt{\frac{22.7}{6}} ; 75.5 + 1.96\sqrt{\frac{22.7}{6}} \right] = [75.5 - 3.8 ; 75.5 + 3.8] = [71.7 ; 79.3]$$

Esercizio 4)
a) Il candiato calcoli le seguenti Probabilità: P(E₁); P(E₂); P(E₁ U E₂) P(E₁ | E₂).
L'evento E_1 è dato dalla probabilità di estrarre un numero negativo da una normale con valore atteso due e varianza quattro. Per definire tale probabilità ci si deve riportare alla normale standardizzatata, standardizzando il valore *x = 0*

$$z_0 = \frac{x_0 - E[X]}{\sqrt{Var[X]}} = \frac{0-2}{\sqrt{4}} = -1$$

Ricordando che le tavole assegnate riportano gli integrali della normale fra 0 ed un numero positivo si ha che

$$P(E_1) = P(X < 0) = P(z < -1) = 0.5 - P(0 < z < 1) = 0.5 - 0.3413 = 0.1587$$

L'evento E_2 è dato dalla probabilità di avere un esito unitario negativo in una prova di Binomiale con n=2 e p =0.5.
La prova binomiale è data dalla somma di *n* prove di Bernoulli i.i.d. dove la generica prova b_i può avere esito pari a 1 o 0.

$$y=\sum_{i=1}^{n} b_i = b_1 + b_2$$

Nel caso in esame l'unico modo di ottenre y = 1 è con le che si verifichi uno dei due casi

$$E': (b_1=0)\cap(b_2=1) \qquad E'': (b_1=1)\cap(b_2=0)$$

che fra loro sono incompatbili. Pertanto

$$P(E_2)=P(E'\cup E'')=P(E')+P(E'')$$

Essendo gli eventi legati alle varibili *b* indipendenti la probabilità dell'evento intersezione è data dal prodotto delle probabilità, pertanto risulta facile calcolare la probabilità richiesta:

$$P(E_2)=P(E'\cup E'')=P(E')+P(E'')=P(b_1=0)\,P(b_2=1)+P(b_1=1)\,P(b_2=0)=0.5*0.5+0.5*0.5=0.5$$

La stessa conclusione poteva essere raggiunta più agevolmente ricordando che da distribuzione di probabilità di una binomiale è data dalla seguente:

$$P(y=k)=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

Da cui

$$P(E_2)=P(y=1)=\binom{2}{1}0.5^1(1-0.5)^{2-1}=\frac{2*1}{1*(2-1)}0.5^1(1-0.5)^{2-1}=0.5$$

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 candiato 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.