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

- Appello del 26 Febbraio 2014 -

Esercizio 1)

In una ricerca si è interessati a verificare il tempo di incubazione (tempo che intercorre dall'inoculazione alla comparsa del primo sintomo) di un agente patogeno. A tale scopo si sono infettate delle cavie e se ne osservato il tempo di incubazione (espresso in giorni) ottenendo la statistica riportata a lato.

Il candidato

a) determini la tipologia del carattere;
b) fornisca una rappresentazione tabellare dei dati opportuna;
c) se possibile, calcoli la mediana;
d) se possibile, calcoli un indice di variabilità;
e) indichi se sono presenti outlier.

10	*11*	*12*	*9*	*11*
8	*7*	*11*	*16*	*9*
9	*7*	*7*	*7*	*16*
6	*24*	*10*	*11*	*11*
6	*12*	*11*	*10*	*9*
11	*10*	*9*	*9*	*7*
8	*9*	*9*	*8*	*11*
10	*8*	*9*	*9*	*11*
5	*20*	*8*	*11*	*8*

Esercizio 2)

Si vuole verificare la validità di una legge genetica (fittizia) che impone la combinazione di due geni A e B. Il candidato,

a) determini il numero di osservazioni necessarie affinchè si possa procedere a tale verifica;
b) supposto di aver monitorato 1000 combinazioni dei geni indicati, indichi se la legge può ritenersi valida. Il candidato indichi e verifichi le ipotesi richieste per l'approccio scelto e proceda al calcolo anche qualora queste non siano soddisfatte.

Gene risultante	A	B	Mutazione (ne A ne B)	Morte dell'embrione (nessun gene da analizzare)
Frequenza relativa teorica	71.50%	26.50%	1.90%	0.10%
Frequenza assoluta osservata (N= 1000)	729	255	16	0

Esercizio 3)

Il candidato, utilizzando i dati dell'Esercizio 1, stimi puntualmente e per intervallo il valore atteso del tempo di incubazione dell'agente patogeno descritto nell'Esercizio 1.

Esercizio 4)

Si considerino i seguenti eventi considerati incompatibili:

E_1: si abbia $x < 1$ dove x è distribuita come una normale avente $E[X] = 2$ e $Var[X]=9$
E_2: y $= 0$ dove y è distribuita come una binomiale con $n = 2$ e $p = 0.4$.

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) Ricalcolare le probabilità precedenti ipotizzando che gli eventi siano indipendenti invece che incompatibili.

- Appello del 26 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 un tempo può assumere qualsiasi valore di tipo continuo)

b) Fornisca una rappresentazione tabellare dei dati opportuna..

Una rappresentazione tabellare opportuna è fornita dalla tabella ed entrata semplice. Questa può essere realizzata in maniera canonica o raccogliendo in classi di modalità i dati. In questo caso si è deciso di utilizzare un tabella ad entrata semplice senza raccogliere i dati in classi di modalità (Tabella 1).

c) Se possibile, calcoli la mediana.

La mediana è il valore che bipartisce le osservazioni ordinate. Considerando le frequenze cumulate in Tabella 1, si osserva come la mediana sia 9 (ovvero la prima misurazione ad avere una frequenza comulata superiore a 0.5).

d) Se possibile, un indice di variabilità.

Avendo calcolato la mediana come indice di posizione, diviene naturale utlizzare la distanza interquartile come indice di posizione. Essa è data data dalla differenza fra il terzo ed il primo quartile.

Ricordando che il ptimo quartile è l'osservazione preceduta da un quarto dei restanti dati ordinati si vede come chiaramente essa sia la quarta modalità (8). Infatti la sua frequenza cumulata è la prima ad essere superiore o uguale a 0.25. Analogamente il terzo quartile è l'osservazione seguita da un quarto dei restanti dati ordinati si vede come chiaramente essa sia la 11. Infatti la sua frequenza cumulata è la prima ad essere superiore o uguale a 0.75.
Risulta quindi chiaro che la distanza interqualtile (D) sia

$$D = q_3 - q_1 = 11 - 8 = 3$$

e) indichi se sono presenti outlier

La presenza degli outlier viene rilevata quando vi sono delle osservazioni esterne ai limiti fissati nel valore adicianete superiore (VAS) ed inferiore (VAI). Data una costante k tipicamente compresa fra 1 e 3 i summenzionati limiti sono dati dalla seguente

$$VAI = q_1 - k\,D \qquad VAS = q_2 + k\,D$$

ponendo $k = 1.5$, si ha che

$$VAI = q_2 - k\,D = 8 - 4.5 = 3.5 \qquad VAS = q_3 + k\,D = 11 + 4.5 = 15.5$$

Pertanto vi sono 4 outlier {16 16 20 24}

i	m_i	n_i	f_i	F_i	$c_i * d_i$	$(x_i-\overline{m})^2$	$(x_i-\overline{m})^2 * f_i$
1	5	1	0.022	0.022	0.1111	25	0.5556
2	6	2	0.044	0.067	0.2667	16	0.7111
3	7	5	0.111	0.178	0.7778	9	1.0000
4	8	6	0.133	0.311	1.0667	4	0.5333
5	9	10	0.222	0.533	2.0000	1	0.2222
6	10	5	0.111	0.644	1.1111	0	0.0000
7	11	10	0.222	0.867	2.4444	1	0.2222
8	12	2	0.044	0.911	0.5333	4	0.1778
9	16	2	0.044	0.956	0.7111	36	1.6000
10	20	1	0.022	0.978	0.4444	100	2.2222
11	24	1	0.022	1.000	0.5333	196	4.3556
Totali		45	1		10		11.6000

Tabella 1) analisi dati Esercizio 1

Esercizio 2)

L'indagine statistica mira a verificare mediante inferenza se osservazioni riportate dal ricercatore ben si adattano al quelle imposte dalla teoria descritta. Pertanto viene richiesto di utilizzare il test di adderenza della distribuzione empirica.

a) determinare il numero di osservazioni necessarie affinchè si possa procedere a tale verifica
Generalente si considera attendibile un test di adattamento alla distribuzione empirica se il campione produce delle frequenze assolute teoriche almeno pari a 5. Ricordando che le frequenze assolute sono date dalle frequenze relative (in questo caso coincidenti con le probabilità teoriche) moltiplicate per la numerosità del campione si ha che:

$$\hat{n}_i = \hat{f}_i * n \Rightarrow 5 \leq \hat{f}_i * n \Rightarrow \frac{5}{\hat{f}_i} \leq n$$

che assume valore massimo per la minima frequenza teorica relativa. Si ottiene pertanto che

$$n \geq \frac{5}{0.001} = 5000$$

b) supposto di aver monitorato 1000 combinazioni dei geni indicati, indichi se la legge può ritenersi valida. Il candidato indichi e verifichi le ipotesi richieste per l'approccio scelto e proceda al calcolo anche qualora queste non siano soddisfatte.

Le tecniche di stima viste nel corso prevedono che:

a) che il campione abbia una numerosità tale da far convergere lo stimatore e
b) che le prove siano indipendenti ed identicamente distribuite (i.i.d.).

La prima ipotesi è stata dimostrata al punto precedente non essere soddisfatta. Il testo non fornisce indicazioni sufficienti per capire se l'esperimento sia stato disegnato usando accorgimenti sufficienti a garantire l'indipendenza delle misurazioni.

Il test in esame, sceglie fra due possibili ipotesi:

$$H_0 : \textit{la distribuzione è quella attesa} \qquad H_1 : \textit{la distribuzione non è quella attesa}$$

Questo test utilizza lo stimatore la di pizzetti-pearson e nel caso ci fosse convergenza dello stimatore questo si distribuirebbe come un chi quadro avente un numero di gradi di libertà pari alle modalità in esame meno uno.

$$\sum_{i=1}^{M} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \sim \chi^2(M-1)$$

Determinata la distribuzione limite dello stimatore, è possibile determinare la regione di accettazione *A*. Il test in esame è di tipo unilaterale destro. Ricorando che *M = 4*, e fissato un livello di significatività del 5 % si ha che

$$A = \left[0 ; \chi^2_{1-\alpha}(M-1)\right] \Rightarrow A = \left[0 ; \chi^2_{0.95}(3)\right] \Rightarrow A = [0 ; 7.81]$$

Per calcolare il valore dello simatore standardizzato è opportuno trasformare le frequenze relative in frequenze assolute.

Gene risultante	A	B	Mutazione (ne A ne B)	Morte dell'embrione (nessun gene da analizzare)
Frequenza assoluta teorica	715	265	19	1
Frequenza assoluta osservata (N= 1000)	729	255	16	0

$$\sum_{i=1}^{M} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \frac{(729-715)^2}{715} + \frac{(255-265)^2}{265} + \frac{(16-19)^2}{19} + \frac{(0-1)^2}{1} = \frac{196}{715} + \frac{100}{265} + \frac{9}{19} + 1 = 2.12$$

Se le ipotesi (a e b) fossero valide potremmo accettare l'ipotesi nulla (H_0) ovvero che la distribuzione empirica aderisce a quella teorica ad un livello si significatività del 5%.

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

- è possibile descrivere l'esperimento mediante la seguente variabile casuale *X: del tempo di incubazione dell'agente patogeno*.
- 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. non è confermabile in quanto non abbiamo indicazioni sul protocollo di test seguito dallo sperimentatore.

La stima puntuale si ottiene semplicemente dall'applicazione dello stimatore, pertanto in base ai conti in tabella 1 si ha

$$\hat{E[X]} = \bar{x} = \sum_{i=1}^{M} f_i x_i = 10$$

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:

$$\hat{Var[X]} = s^2 = \frac{N}{N-1}\sigma_{\bar{x}}^2 = \frac{N}{N-1}\sum_{i=1}^{M} f_i \left(x_i - \bar{x}\right)^2 = \frac{45}{44} 11.6 = 11.86$$

Infine si ottiene la stima richiesta:

$$\hat{E[X]} \in \left[10 - 1.96\sqrt{\frac{11.86}{45}} ; 10 + 1.96\sqrt{\frac{11.86}{45}} \right] = [10 - 1.01 ; 10 + 1.01] = [8.99 ; 11.01]$$

Esercizio 4)
a) Il candiato calcoli le seguenti Probabilità: $P(E_1)$; $P(E_2)$; $P(E_1 \cup E_2)$ $P(E_1 \mid E_2)$.
L'evento E_1 è dato dalla probabilità di estrarre un numero negativo da una normale con valore atteso due e varianza nove. Per definire tale probabilità ci si deve riportare alla normale standardizzatata, standardizzando il valore *x = 0*

$$z_0 = \frac{x_0 - E[X]}{\sqrt{Var[X]}} = \frac{1-2}{\sqrt{9}} = -\frac{1}{3}$$

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

$$P(E_1) = P(X < 0) = P(z < -1/3) = 0.5 - P(0 < z < 1/3) = 0.5 - 0.1293 = 0.3707$$

L'evento E_2 è dato dalla probabilità di avere due esito negativi (ovvero nessun esito positivo) 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 verfichi

$$E_2 : (b_1 = 0) \cap (b_2 = 0)$$

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:

$$P(E_2)=P(b_1=0)P(b_2=0)=(1-p)*(1-p)=0.6*0.6=0.36$$

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

$$P(E_2)=P(y=0)=\binom{2}{0}0.5^0(1-0.5)^{2-0}=\frac{2*1}{1*(2*1)}0.4^1(1-0.4)^2=0.36$$

Essendo gli eventi incompatibili la probabilità dell'evento intersezione è nulla

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

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1\cup E_2)=P(E_1)+P(E_2)-P(E_1\cap E_2)=0.37+0.36-0=0.73 \qquad P(E_1|E_2)=P\frac{(E_1\cap E_2)}{P(E_2)}=\frac{0}{0.36}=0$$

b) Ricalcolare le probabilità precedenti ipotizzando che gli eventi siano indipendenti invece che incompatibili.
Le probiltà degli eventi E1 ed E2 restano invariate. Cambia la probabilità dell'evento intersezione che diviene pari al prodotto delle probabilità.

$$P(E_1\cap E_2)=P(E_1)P(E_2)=0.37*0.36=0.1332$$

In ragione di questo si possono ricalcolare la restanti probabilità riapplicando le definzioni assiomatiche:

$$P(E_1\cup E_2)=P(E_1)+P(E_2)-P(E_1\cap E_2)=0.37+0.36-0.13=0.60$$

$$P(E_1|E_2)=P\frac{(E_1\cap E_2)}{P(E_2)}=\frac{P(E_1)P(E_2)}{P(E_2)}=P(E_1)=0.37$$