- Appello del 10 Luglio 2014 -

Esercizio 1)
Il direttore dell'azienda "YY" di 35 dipendenti ha monitorato nel mese di febbraio quante volte un suo dipendente sia arrivato in ritardo per oltre 10 minuti sul posto di lavoro. Si è ottenuta la segunte statistica:

3	2	0	4	0	3	3	3	0	3
2	1	3	1	1	1	2	1	3	1
3	1	3	2	1	1	2	2	1	3
0	3	1	1	0					

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) se possibile, determini la presenza di outlier.

Esercizio 2)
Il candidato, usando come campione i dati descritti nell'esercizio 1, stimi puntualmente e per intervallo la varianza della V.C.

R: numero di volte che un dipendente dell'azionda "YY" arriva in ritardo in un mese

Il candidato evidenzi e valuti le ipotesi necessarie e proceda al calcolo anche quando queste non siano verificate.

Esercizio 3)
Un medico vuole verificare se l'efficacia di un farmaco è coerente con quanto dichiarato nel il foglietto informativo, pertanto confronta i dati forniti dall'azienda farmaceutica in merito all'esito di un ciclo di trattamento (frequenze teoriche) con quanto ottenuto monitorando un campione formato da pazienti trattati con il farmaco in esame.
Il candiato:

a) determini il numero di osservazioni necessarie affinchè si possa procedere a tale verifica
b) supposto di aver monitorato un campione di 1000 soggetti, indicare se il foglietto informativo può ritenersi corretto basandosi sulle seguenti frequenze relative. Il candidato indichi e verifichi le ipotesi richieste per l'approccio scelto e proceda al calcolo anche qualora queste non siano soddisfatte.

Esito	guarigione senza effetti collaterali	guarigione con effetti collaterali	nessun miglioramento	peggioramento delle condizioni	Morte
Frequenza osservata	60.00%	25.00%	13.00%	2.00%	0.00%
Frequenza teorica	49.99%	30.00%	16.00%	4.00%	0.01%

Esercizio 4)
Si considerino i seguenti eventi dichiarati incompatibili.

E_1: si ottenga z > 1.9 estraendo dalla variabile Z
E_2: si ottenga $x=0$ dove x è estratto da una v.c. distribuita come $Ber(0.8)$

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 10 Luglio 2014 - Svolgimento

Esercizio 1)

a) determini la tipologia del carattere;

Le osservazioni sono rappresentabili con numeri naturali, per cui il carattere è quantitativo discreto.

b) fornisca una rappresentazione grafica dei dati;

Dovendo valutare la presenza di outlier la rappresentazione più sensata è data il box-plot. Per fare ottenere questa rappresentazione è necessario il calcolo dei quartili.

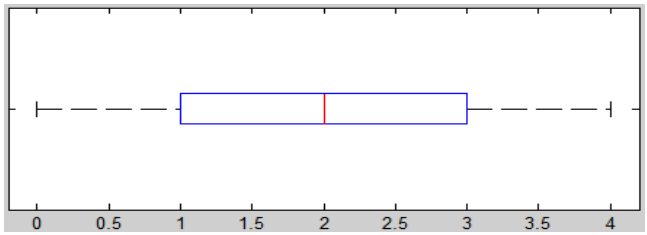
- q_0: osservazione minima $q_0=0$.
- q_1: osservazione che lascia un quarto delle restanti osservazioni alla sua sinistra. Tolto il quartile interessato restano 34 osservazioni; da cui si ricercca il valore precedeuto da 34/4 = 8.5 osservazioni. Poiche il numero non è intero si media la nona e la decima osservazione. Entrambe le osservazioni sono pari a 1, pertanto si ha $q_1=1$.
- q_2: osservazione che lascia metà delle restanti osservazioni alla sua sinistra. Tolto il quartile interessato restano 34 osservazioni; da cui si ricercca il valore precedeuto da 34/2 = 17 osservazioni ovvero la diciottesima osservazione. Si ha $q_2=2$.
- q_3: osservazione che lascia tre quarti delle restanti osservazioni alla sua sinistra. Tolto il quartile interessato restano 34 osservazioni; da cui si ricercca il valore precedeuto da 3*34/4 = 25.5 osservazioni. Poiche il numero non è intero si media la ventiseiesima e la ventisettesima osservazione. Entrambe le osservazioni sono pari a 3, pertanto si ha $q_3=3$.
- q_4: osservazione massima $q_4=4$.

E dei valori adiacenti ottenuti fissando la costante del boxplot ad 1.

$$VAI = max(q_0 ; q_1 - K*(q_3 - q_1)) = max(0 ; 1 - 1*(3-1)) = max(0 ; -1) = 0$$

$$VAS = min(q_4 ; q_3 + K*(q_3 - q_1)) = min(4 ; 3 + 1*(3-1)) = min(4 ; 5) = 4$$

Il grafico ottenuto è il seguente



c) indichi e calcoli tutti gli indici di posizioni adeguati ai dati;

Gli indici di posizioni posizione indicano il valore centrale della statistica e sono:

- *Moda*: modalità avente massima frequenza. Nel caso in esame vale 1.
- *Mediana*: Corrisponde al secondo quartile . Nel caso in esame vale 2.
- *Media*: Somma delle osservazioni fratto la loro numerosità. Utilizzando i conti in tabella questa vale *61/35*.

d) se possibile, determini la presenza di outlier.

Gli outlier sono osservazioni fuorvianti dovute ad errori di misura o intenti maliziosi degli intervistati. Un modo per identificarle è quello di valutare come outlier i valori esterni all'intevallo [VAI ; VAS]. Nel caso in esame non ci sono osservazioni esterne, pertanto si assume che non vi siano outlier.

i	m_i	n_i	F_i	$m_i n_i$	m_i^2	$m_i^2 n_i$
1	0	5	5	0	0	0
2	1	12	17	12	1	12
3	2	6	23	12	4	24
4	3	11	34	33	9	99
5	4	1	35	4	16	16
		35		61		151

Esercizio 2)
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

- il testo fornisce la varibile da utilizzare.
- la grandezza da stimare risulta *Var[X]* il cui stimatore è la varianza campionaria la quale converge in legge per campioni avente numerosità superiore a 30 (ipotesi confermata).
- L'ipotesi di prove i.i.d. è molto debole in quanto le prove sono fatte su un solo mese mentre lavaribile vuole prendere considerazioni sull'intero anno.

La stima puntuale può essere effettuata ricordando che la varianza viene stimata correttamente mediante la varianza campionaria. Il parte del calcolo è già stato effettuato nella tabella riportata nello svolgimento del primo esercizio, ottenendo

$$\hat{Var}[P]=s^2=\frac{n}{n-1}\sigma^2=\frac{n}{n-1}\left(\frac{\sum_{i=1}^{M} n_i(x_i^2)}{n}-\bar{x}^2\right)=\frac{35}{34}\left(\frac{151}{35}-\frac{61^2}{35^2}\right)=\frac{1}{34}\left(151-\frac{3721}{35}\right)=\frac{1}{34}\left(\frac{1564}{35}\right)=1.31$$

Le stime per intervallo sono regolate dal livello di confidenza (*1-α*) che solitamente è fissato da chi svolge l'analisi dei dati. Nel caso in esame una scelta valida è porre *α = 10%*. Determinato il livello di confidenza la stima *I* è data dalla seguente formula:

$$I=\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)};\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]=\left[\frac{34\cdot 1.31}{\chi^2_{0.95}(34)};\frac{34\cdot 1.31}{\chi^2_{0.05}(34)}\right]=\left[\frac{44.7}{50.16};\frac{44.7}{17.93}\right]=[0.89;2.49]$$

Si noti come le tavole non consentano il calcolo del chi-quadro a 34 gradi di libertà. In questo caso si può utilizzare la convergenza in legge della distribuzione chi-quadro a quella normale. Si ha infatti che per un numero di gradi libertà sufficientemente elevato

$$\chi^2(\nu)\sim N(\nu,2\nu)\Rightarrow\chi^2(34)\sim N(34,68)$$

Peranto i valore richiesti posson essere recuperati usando la distribuzione normale. Ovviamente le tavole consento l'uso della sola normale standard. Pertanto si recupera il valore che lascia nella coda della normale la probabilità richiesta e lo si destandardizza. Si ottengono quindi i seguenti valori

$$\chi^2_{0.95}(34)=N_{0.95}(34;68)=Z_{0.95}*Var[N(34;68)]+E[N(34;68)]=1.96*8.2+34=50.16$$

$$\chi^2_{0.05}(34)=N_{0.05}(34;68)=Z_{0.05}*Var[N(34;68)]+E[N(34;68)]=-1.96*8.2+34=17.93$$

Esercizio 3)
L'indagine statistica mira a verificare mediante inferenza se le osservazioni riportate dal ricercatore ben si adattino a quelle descritte nel foglio informativo. 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:

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

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

$$n\ge\frac{5}{0.0001}=50000$$

b) supposto di aver monitorato un campione di 1000 soggetti, indicare se il foglietto informativo può ritenersi corretto basandosi sulle seguenti frequenze relative. 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) il campione abbia una numerosità tale da far convergere lo stimatore e
b) 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 = 5*, e fissato un livello di significatività del 5 % si ha che

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

Per calcolare il valore dello simatore standardizzato è opportuno trasformare le frequenze relative in frequenze assolute. Questa operazione si ha moltiplicando la frequenza relativa per la numerosità del campione.

Esito	guarigione senza effetti collaterali	guarigione con effetti collaterali	nessun miglioramento	peggioramento delle condizioni	Morte
Frequenza osservata	600	250	130	20	0
Frequenza teorica	500	300	160	40	0

$$\sum_{i=1}^{M} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \frac{(600-500)^2}{500} + \frac{(250-300)^2}{300} + \frac{(130-160)^2}{160} + \frac{(20-40)^2}{40} = \frac{10000}{500} + \frac{2500}{300} + \frac{900}{160} + \frac{400}{40} = 43.96$$

Se le ipotesi (a e b) fossero valide potremmo accettare l'ipotesi altenativa (H_1) ovvero che la distribuzione empirica non aderisce a quella teorica ad un livello si significatività del 5%.

Esercizio 4)

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

$P(E_1)$: La probabilità richiesta è quella di estrarre un numero maggiore di 1.9 da una normale standardizzata. Nelle tavole statsistiche viene solitamente riportata $P(Z<z)$ o $P(0<Z<z)$ per diversi valori di *z*. Ricordando il legame fra le due probailità (ovvero che $P(Z<z) = P(0<Z<z) + 0.5$) diviene facile ricavare la probabilità richiesta

$$P(E_1) = P(Z>1.90) = 1 - P(Z<1.9) = 1 - (P(0<Z<1.9) + 0.5) = 0.5 - P(0<Z<1.9) = 0.5 - 0.4713 = 2.87\%$$

$P(E_2)$: L'evento E_2 prevede l'estrazione di uno *0* da una *Ber(0.8)*. In una bernoulliana questa probabilità è chiamata *q* e discende dal parametro della bernoulliana *p* per complemento. Pertano si ha che

$$P(E_2) = q = 1 - p = 1 - 0.8 = 0.2$$

Ricordando che due eventi incompatibili non possono verificarsi contemporaneamente (ovvero hanno probabiltà dell'evento intersezione nulla) le altre probabilità richieste possono essere ricavate utilizzando la probabilità assiomatica

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.2 + 0.0287 - 0 = 0.2287$$

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

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} = \frac{0}{0.0287} = 0$$

b) indichi se gli eventi E_1 ed E_2 possono ritenersi dipendenti.

L'indipendenza statistica si ha quando le pobabilità condizionate corrispondono con le probabilità non condizionate.

$$P(E_1|E_2) = P(E_1) \quad \text{e} \quad P(E_2|E_1) = P(E_2)$$

Poichè questa condizione non è verificata si può asserire che gli eventi sono dipendenti.