

- Appello del 20 Febbraio 2013 -

**Esercizio 1)**

Un ricercatore vuole verificare la pericolosità dell'uso dei telefonini cellulari di nuova generazione. Per far ciò ha provato a fare una telefonata di 3 minuti ad un soggetto ed ha valutato la penetrazione delle onde nel cervello mediante un esame non invasivo. Il risultato della penetrazione delle onde è stato suddiviso in tre categorie (Alto, Medio e Basso). L'esperimento condotto su 40 soggetti ha conseguito i seguenti risultati:

A A M B A M M A B A M B A M A A M B M B  
A B A M M B A M M M A M A M A M A B M A

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) Descriva e calcoli gli indici di variabilità adeguati ai dati.
- e) Dopo aver descritto quali caratteristiche siano riassunte dagli indici sintetici di forma, calcoli quelli adatti ai dati in esame.

**Esercizio 2)**

Ai 40 soggetti descritti nel precedente esercizio è stato chiesto il numero intero di ore in cui usano il cellulare in media al giorno. Si sono ottenute *ordinatamente*<sup>1</sup> le seguenti osservazioni:

0 2 1 1 2 2 2 0 1 2 2 1 1 1 0 1 0 2 1 1  
0 2 1 0 1 0 0 1 2 1 2 2 0 0 1 0 1 0 2 1

Il candidato usi queste osservazioni per stimare puntualmente e per intervallo il numero di ore in cui un generico soggetto usa il cellulare. Il candidato indichi le ipotesi necessarie e proceda al calcolo anche se queste non fossero soddisfatte.

**Esercizio 3)**

Il ricercatore vuole utilizzare i dati raccolti negli esercizi 1 e 2 per verificare se i due caratteri misurati siano tra di loro legati. Il candidato

- a) raccolga i dati in una tabella a doppia entrata
- b) indichi il carattere della serie bi-variata ottenuta
- c) Se possibile calcoli un indice sintetico di posizione
- d) Il candidato verifichi se esiste un legame fra le ore di uso quotidiano del cellulare ed il grado di penetrazione delle onde, evidenziando le necessarie ipotesi. Il candidato proceda al calcolo anche se queste non fossero verificate.
- e) Qualora un legame esistesse, si indichi la natura del legame.

**Esercizio 4)**

Si considerino i seguenti eventi dichiarati indipendenti.

$E_1$ : si ottenga  $x < 1$  dove  $x$  è estratto da una v. c. distribuita come una  $N(2, 1)$ .

$E_2$ : si ottenga  $y > 0$  dove  $y$  è estratto da una v. c.  $Unif(-2; 1)$ .

- a) Il candidato calcoli le seguenti Probabilità:  $P(E_1)$ ;  $P(E_2)$ ;  $P(E_1 \cup E_2)$   $P(E_1 | E_2)$ .
- b) Il candidato fornisca la definizione dei seguenti eventi notevoli: eventi statisticamente indipendenti, evento certo ed evento impossibile.

---

<sup>1</sup> Il termine ordinatamente indica che le osservazioni sono state raccolte interrogando le stesse unità statistiche nel medesimo ordine e che quindi le osservazioni relative alla stessa unità statistica occupano la stessa posizione. Pertanto se si volesse costruire le osservazioni relative alla bi-variata si esse sarebbero (A;0), (A;2), (M,1), ..., (A;1).

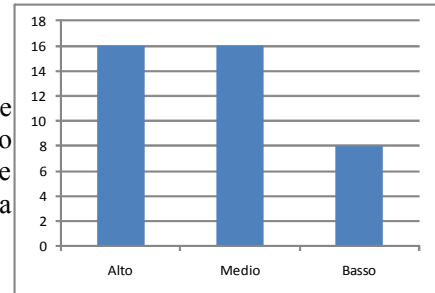
- Tema d'esame del 20 Febbraio 2013 -  
*Svolgimento*

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) *Fornisca una rappresentazione grafica opportuna.*

Per caratteri qualitativi non ordinabili tipicamente si utilizza come rappresentazione grafica 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 relativo ai dati in oggetto.



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

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 solo i primi due.

- La moda (ovvero la modalità cui corrisponde la modalità maggiore). In questo caso la frequenza assoluta non ha un massimo unico ma doppio. Si dice quindi che la statistica è bimodale e le due mode sono "Alto" e "Medio".
- La Mediana (ovvero l'osservazione che bipartisce le osservazioni ordinate). In questo caso si hanno 40 osservazioni pertanto la mediana deve lasciarne  $(40-1)/2 = 19.5$  alla propria destra ed altrettante alla propria sinistra. Poiché questo non è possibile, generalmente si mediano la 20-sima e la 21-sima osservazione. Normalmente la media di due osservazioni qualitative non sarebbe possibile ma considerando le frequenze assolute e cumulate riportate a lato si evince che entrambe le osservazioni sono pari ad M, quindi la mediana non può che essere "M".

|       | $n_i$ | $F_i$ |
|-------|-------|-------|
| Alto  | 16    | 16    |
| Medio | 16    | 32    |
| Basso | 8     | 40    |

d) *Descriva e calcoli gli indici di variabilità adeguati ai dati..*

Per dati come qualitativi non è possibile determinare alcun indice di variabilità.

e) *Dopo aver descritto quali caratteristiche siano riassunte dagli indici sintetici di forma, calcoli quelli adatti ai dati in esame.*

Gli indici di forma sono volti a misurare due caratteristiche

- La simmetria* ovvero il fatto che le osservazioni si distribuiscano in modo simmetrico rispetto al valore centrale. Gli indici utilizzabili a questo scopo sono il momento centrale terzo, il momento centrale terzo normato (o standardizzato), il coefficiente di skewness (o di asimmetria) di Pearson.
- La curtosi* ovvero il fatto che le osservazioni si distribuiscano in modo simile ad una distribuzione di tipo gaussiano (o normale). Gli indici utilizzabili a questo scopo sono il momento centrale quarto, il momento centrale quarto normato (o standardizzato), l'eccesso curtosi,

Nessuno degli indici summenzionati può essere calcolato per caratteri qualitativi quale quello in esame.

Esercizio 2)

Possiamo modellare il processo di generazione dei dati del presentato nell'esercizio come il frutto di diverse estrazioni dalla variabile casuale

$P$  : numero di ore giornaliere in cui una persona usa il cellulare.

avente distribuzione ignota. Si sono effettuate  $n = 40$  osservazioni i. i. d. in cui si sono rilevate  $M = 3$  modalità.

Continuando con il modello precedentemente definito l'esercizio richiede di stimare  $E[P]$  puntualmente e per intervallo.

La stima è possibile solo se le estrazioni sono indipendenti ed identicamente distribuite e se la dimensione del campione è pari ad almeno 30 unità. Nel caso in esame entrambe le ipotesi si possono considerare verificate.

Lo stimatore puntuale del valore atteso è la media campionaria che, utilizzando i conti riportati in tabella produce la

seguente stima

$$E[\hat{P}] = \bar{y} = \sum_{i=1}^M f_i * y_i = 3$$

Per ottenere la stima per intervallo occorre fissare un livello di confidenza  $\alpha$ . Posto  $\alpha = 90\%$ , la stima per intervallo è in caso che la varianza della popolazione sia ignota è data dalla

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

ricordando la formula del calcolo della varianza campionaria ed utilizzando i conti riportati in tabella si ha che:

$$s^2 = \sigma^2 \frac{n}{n-1} = \left( \sum_{i=1}^M f_i * y_i^2 - \bar{y}^2 \right) \frac{n}{n-1} = (1.6 - 1^2) \frac{40}{39} = 0.6153 \quad z_{\frac{\alpha}{2}} = 1.96$$

Pertanto l'intervallo richiesto è:  $E[X] \in \left[ 1 - 1.96 \sqrt{\frac{0.6156}{40}} ; 1 + 1.96 \sqrt{\frac{0.6156}{40}} \right] = [0.876 ; 1.124]$

| $y_i$         | $n_i$     | $f_i$    | $n_i * f_i$ | $y_i^2$ | $y_i^2 * f_i$ |
|---------------|-----------|----------|-------------|---------|---------------|
| 0             | 12        | 0.300    | 0.000       | 0       | 0             |
| 1             | 16        | 0.400    | 0.400       | 1       | 0.4           |
| 2             | 12        | 0.300    | 0.600       | 4       | 1.2           |
| <b>Totali</b> | <b>40</b> | <b>1</b> | <b>1</b>    |         | <b>1.6</b>    |

### Esercizio 3)

Unendo i dati dei primi due esercizi si ottiene la seguente bivariata (x, y) dove

*X: grado di penetrazione delle onde, Y: ore giornaliere in cui una persona usa il cellulare*

a) raccolga i dati in una tabella a doppia entrata

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

| X: grado di penetrazione  | Y: Ore giornaliere |           |           | Totale complessivo |
|---------------------------|--------------------|-----------|-----------|--------------------|
|                           | 0                  | 1         | 2         |                    |
| Alto                      | 6                  | 6         | 4         | <b>16</b>          |
| Medio                     | 4                  | 6         | 6         | <b>16</b>          |
| Basso                     | 2                  | 4         | 2         | <b>8</b>           |
| <b>Totale complessivo</b> | <b>12</b>          | <b>16</b> | <b>12</b> | <b>40</b>          |

b) indichi il carattere della serie bi-variata ottenuta

La serie bivariata è data dall'unione di un carattere qualitativo ed un quantitativo pertanto non presenta un carattere preciso.

c) Se possibile calcoli un indice sintetico di posizione

L'unico indice di posizione calcolabile per la serie in esame è la moda corrispondente alla modalità avente frequenza relativa maggiore. Questo caso vi sono 4 modalità che soddisfano questo requisito.

(Alto; 0)      (Alto; 1)      (Medio; 1)      (Medio; 2)

d) Il candidato verifichi se esiste un legame fra le ore di uso quotidiano del cellulare ed il grado di penetrazione delle onde, evidenziando le necessarie ipotesi. Il candidato proceda al calcolo anche se queste non fossero

verificare.

E' possibile rispondere al quesito mediante un particolare test di ipotesi: il test di indipendenza. Il test sfrutta 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 risultano essere le seguenti:

| X: grado di penetrazione  | Y: Ore giornaliere |           |           | Totale complessivo |
|---------------------------|--------------------|-----------|-----------|--------------------|
|                           | 0                  | 1         | 2         |                    |
| Alto                      | 4.8                | 6.4       | 4.8       | <b>16</b>          |
| Medio                     | 4.8                | 6.4       | 4.8       | <b>16</b>          |
| Basso                     | 2.4                | 3.2       | 2.4       | <b>8</b>           |
| <b>Totale complessivo</b> | <b>12</b>          | <b>16</b> | <b>12</b> | <b>40</b>          |

A questo punto è possibile valutare la convergenza dello stimatore di Pizzetti Pearson, possibile solo se tutte le frequenze teoriche sono superiori a 5. La condizione non è verificata ma il testo richiede comunque di continuare il procedimento.

Poiché entrambi i caratteri della bi-variata hanno 3 modalità ( $M_x = M_y = 3$ ), la regione di accettazione per un test al 5 % è la seguente.

$$A = [0; \chi^2_{0.95}((M_x - 1) * (M_y - 1))] = [0; \chi^2_{0.95}(4)] = [0; 9.488]$$

Per valutare l'ipotesi di indipendenza non rimane che da calcolare il valore dello stimatore e verificare se appartiene alla regione di accettazione. Il valore dello stimatore è

$$\sum_{i=1}^3 \sum_{j=1}^3 \frac{(n_{i,j} - \hat{n}_{i,j})^2}{\hat{n}_{i,j}} = \frac{(6-4.8)^2}{4.8} + \frac{(6-6.4)^2}{6.4} + \frac{(4-4.8)^2}{4.8} + \frac{(4-4.8)^2}{4.8} + \frac{(6-6.4)^2}{6.4} + \frac{(6-4.8)^2}{4.8} + \frac{(2-2.4)^2}{2.4} + \frac{(4-3.2)^2}{3.2} + \frac{(2-2.4)^2}{2.4} = 1.25$$

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

e) *Qualora un legame esistesse indichi la natura del legame.*

Poiché i caratteri sono indipendenti non vi è nessun legame.

#### Esercizio 4)

Nel testo si considerano come eventi elementari delle estrazioni da due vv. cc. continue. Si ricorda che per una v.c. continua  $X$ , la probabilità di ottenere una realizzazione  $x$  compresa fra due estremi  $x_{inf}$  ed  $x_{sup}$  è data dall'integrale (area sottesa) dalla densità di probabilità  $f_X(x)$  fra  $x_{inf}$  ed  $x_{sup}$ . In simboli

$$P(x \in [x_{inf}; x_{sup}]) = \int_{x_{inf}}^{x_{sup}} f_X(\tau) d\tau$$

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

Per il calcolo di  $P(E_1)$ , si deve calcolare l'area sottesa dalla curva della densità di probabilità di una normale con valore atteso 2 varianza unitaria da meno infinito a 1. Purtroppo integrare le d.d.p. di una normale qualsiasi non è agevole solitamente si ricorre alla variabile normale standardizzata  $Z(N(0;1))$ . Pertanto si deve trasporre l'evento di estrazione da  $X$  all'estrazione da  $Z$  mediante la standardizzazione dei valori coinvolti.

$$x_{inf} = -\infty \rightarrow z_{inf} = \frac{x_{inf} - E[X]}{\sqrt{Var[X]}} = \frac{-\infty - 2}{1} = -\infty \quad x_{sup} = -1 \rightarrow z_{sup} = \frac{x_{sup} - E[X]}{\sqrt{Var[X]}} = \frac{-1 - 2}{1} = -3$$

Pertanto ha la seguente equivalenza

$$P(E_1) = P(x \in [-\infty; 1]) = P(z \in [-\infty; -1]) = \int_{-\infty}^{-1} f_z(\tau) d\tau$$

Infine ricordando che, la d.d.p. di una qualsiasi gaussiana è simmetrica rispetto al valore atteso e che sulle tavole periodiche si riportano solo gli integrali fra 0 e un valore positivo si ha che

$$P(E_1) = \int_{-\infty}^{-1} f_z(\tau) d\tau = \int_{-\infty}^0 f_z(\tau) d\tau - \int_{-1}^0 f_z(\tau) d\tau = \int_{-\infty}^0 f_z(\tau) d\tau - \int_0^1 f_z(\tau) d\tau = 0.5 - 0.3413 = 0.1587$$

La probabilità dell'evento  $E_2$  è pari alla probabilità di estrarre un numero  $y$  positivo da una v. c. uniforme fra -2 ed 1. 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  $[-2; 1]$  la predente diviene

$$\int_{-2}^{+1} f_Y(\tau) d\tau = 1$$

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

$$\int_{-2}^{+1} C d\tau = (1 - (-2))C = 3C = 1$$

da cui si ha che  $C = 1/3$ .

Il calcolo della probabilità richiesta diviene ora facile

$$P(E_2) = P(y > 0) = \int_0^{+1} \frac{1}{3} d\tau = (1 - 0) \frac{1}{3} = 0.3333$$

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

$$P(E_1 \cap E_2) = P(E_1)P(E_2) = 0.1587 \cdot 0.3333 = 0.0529$$

Le restanti probabilità possono essere ricavate utilizzando la definizione assiomatica

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.1587 + 0.3333 - 0.0529 = 0.4391 \quad P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{0.0529}{0.3333} = 0.1587$$

*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  $P(A|B) = P(A)$  e  $P(B|A) = P(B)$
- *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  $P(E) = 0$ .