Compito scritto dell'esame di Statistica e analisi dei dati 28 giugno 2024	Prof. Giuseppe Boccignone	Corso di Laurea
Cognome:	Nome:	Matricola:

Istruzioni

- Il tempo riservato alla prova scritta e' di 2 ore. Durante la prova e' possibile consultare il formulario ed utilizzare la calcolatrice. Non e' possibile consultare libri, appunti, cellulari.
- Ogni foglio deve riportare il numero di matricola
- In ogni esercizio occorre indicare chiaramente, per ogni risposta, il numero della domanda corrispondente
- Riportare lo svolgimento degli esercizi per esteso (quando l'esercizio richiede piu' passaggi di calcolo, non sara' preso in considerazione se riporta solo le soluzioni). Se una serie di calcoli coinvolge una o piu' frazioni semplici (numeratore e denominatore interi), per chiarezza, si svolgano i calcoli mantenendo tali numeri in forma frazionaria fin dove possibile (non li si converta nelle loro approssimazioni con virgola e decimali: solo il risultato finale sara' eventualmente rappresentato in quest'ultima forma).

Problemi

Esercizio 1. Nell'arco di una giornata (= tra le 8:00 del mattino e mezzanotte) Aldo riceve in media 10 sms.

(a) Qual é la probabilitá che Aldo riceva 3 sms tra le 12:00 e le 13:30 (= evento A)?

Soluzione: Poiché il numero medio di sms é basso, possiamo assumere che il numero X_T di sms ricevuti da Aldo in un periodo di T ore segua la legge di Poisson $X_T \sim Pois(\mu = \lambda T)$, dove l'intensitá o rate λ rappresenta il numero medio di sms ricevuti in un'ora. Dai dati risulta che

$$\lambda = 10/16 = 0.625\,[n^o sms/ora]$$

perché tra le 8:00 del mattino e mezzanotte trascorrono 16 ore. Tra le le 12:00 e le 13:30 passa un periodo di $T = 1.5$ ore, quindi

$$P(A) = Pois(X_{1.5} = 3; \mu = \lambda T) = \frac{\mu^3}{3!}e^{-\mu} = \frac{(0.625 \times 1.5)^3}{3!}e^{-0.625\times 1.5} \approx 5.38\% \quad (1)$$

(b) Qual é la probabilitá che Aldo non riceva alcun sms prima di mezzogiorno (= evento B)?

Soluzione:
Prima di mezzogiorno significa tra le 8:00 e mezzogiorno, quindi é $T = 4$ ore e si ha

$$P(B) = Pois(X_4 = 0; \mu = \lambda T) = e^{-2.5} = 8.21\%$$

Esercizio 2.
Un professore sostiene che il punteggio medio in un certo test é stato almeno 83. Si assuma che il punteggio al test si distribuisca normalmente. Gli studenti ritengono che invece il punteggio medio sia inferiore ad 83, per cui decidono di chiedere ad un campione casuale di studenti il loro voto e ne risultano i seguenti voti:

$$82, 77, 85, 76, 81, 91, 70, 82$$

(a) Verificare mediante un test di ipotesi se sia lecito dubitare dell'affermazione del professore ad un livello di significativitá del 5%?

Soluzione:
Abbiamo che la taglia del campione é piccola: $n = 8$.
La media campionaria é:

$$\bar{x} = \frac{82 + 77 + 85 + 76 + 81 + 91 + 70 + 82}{8} = 80.5$$

La varianza empirica:

$$s^2 = \frac{(82 - 80.5)^2 + (77 - 80.5)^2 + \cdots}{(8 - 1)} = 39.71429$$

e dunque $s = \sqrt{39.71429} = 6.3$

Usiamo un test di ipotesi unilaterale sulla media

$$\begin{cases} H_0 & : \mu \geq \mu_0 = 83 \\ \\ H_1 & : \mu < \mu_0. \end{cases} \tag{2}$$

E' un test effettuato sulla media: distribuzione normale, varianza teorica non nota, campione piccolo.
Dunque si utilizza la statistica test T:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{80.5 - 83}{6.3/\sqrt{8}} = -1.122$$

La regola di decisione é: rifiuto H_0 se

$$T < t_{\alpha,n-1}.$$

Con $\nu = n - 1 = 7$ gdl e $\alpha = 0.05$, il quantile inferiore $t_{\alpha,n-1}$ di Student é $t_{0.05,7} = -1.895$
Vediamo che $T = -1.122 > -1.895$.
Dunque, non possiamo rigettare H_0, ovvero l'ipotesi che il punteggio medio nel test sia stato almeno 83

Esercizio 3.
Due gemelli possono essere veri gemelli, e in tal caso sono sempre dello stesso sesso, oppure pseudo-gemelli, e in questo caso la probabilitá che abbiano lo stesso sesso é del 50%. Le coppie dello stesso sesso rappresentano i 2/3 di tutte le coppie di gemelli.

(a) Qual é la probabilitá che due gemelli siano veri gemelli?

Soluzione: Definiamo gli eventi
$V =$ "veri gemelli"
$\sim V =$ "pseudo-gemelli"
$S =$ "gemelli dello stesso sesso"
Qui S rappresenta il dato osservabile.
Le informazioni a disposizione sono:

$$P(S \mid V) = 1$$
$$P(S \mid \sim V) = 1/2$$
$$P(S) = 2/3$$

Si vuole calcolare $P(V)$.
Esplicitando $P(S)$ (noto) mediante il teorema della probabilitá totale

$$\frac{2}{3} = P(S) = P(S \mid V)P(V) + P(S \mid \sim V)P(\sim V) = P(V) + \frac{1}{2}(1 - P(V))$$

da cui $P(V) = \frac{1}{3}$

(b) Qual é la probabilitá che due gemelli, se sono dello stesso sesso, siano veri gemelli?

Soluzione: Applicando la regola di Bayes si inferisce che:

$$P(V \mid S) = \frac{P(S \mid V)P(V)}{P(S)} = \frac{1/3}{2/3} = \frac{1}{2}$$

Esercizio 4.
Un ricercatore, incaricato di stimare la percentuale di famiglie italiane che hanno più di un computer, dopo aver rilevato che il 27% di un campione costituito da 492 famiglie ha dichiarato di possedere più di un computer, fornisce l'intervallo di confidenza $[0.2308, 0.3092]$, ma omette di dire il livello di confidenza.

(a) Qual é il livello di confidenza associato a questo intervallo?

Soluzione: Un intervallo di confidenza per la proporzione p (approssimato per grandi campioni) é

$$\hat{p} \pm z_{\frac{\alpha}{2}} \cdot ES$$

con ES errore standard pari a

$$ES = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

L'errore e non puó superare il semi-intervallo $e = |z_{\frac{\alpha}{2}} \cdot ES|$, da cui

$$z_{\frac{\alpha}{2}} = \frac{e}{ES} \tag{3}$$

La proporzione stimata $\hat{p}$ é ovviamente il centro dell'intervallo $[0.2308, 0.3092]$, pertanto

$$\hat{p} = \frac{0.2308 + 0.3092}{2} = 0.27$$

.

Possiamo quindi calcolare i termini dell'equazione 4:

$$e = |0.3092 - 0.27| = 0.0392$$

$$ES = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.27 \times 0.73}{492}} = 0.02$$

e dunque

$$z_{\frac{\alpha}{2}} = \frac{0.0392}{0.02} = 1.96 \tag{4}$$

Riconosciamo a colpo come tale valore critico sia quello riconducibile al livello di confidenza del 95%.
Infatti, piú formalmente:

$$1 - \frac{\alpha}{2} = P(Z < 1.96) = 0.975 \implies \alpha = 0.05 \implies (100)(1-\alpha)\% = 95\%$$

Esercizio 5.
Un'azienda produce un modello di auto la cui percorrenza X (in km con 1 litro di benzina) ha distribuzione normale, media 25 km/l e deviazione standard 2 km/l. Supponiamo di avere un campione casuale di quattro auto prodotte in serie.

(a) La percorrenza media campionaria che distribuzione ha?

Soluzione: Il testo ci dice che

$$X \sim \mathcal{N}(\mu = 25, \sigma = 2)$$

Il campione é formato da quattro misure di percorrenza (X_1, X_2, X_3, X_4) indipendenti e identicamente distribuite come X.
Quindi la percorrenza media campionaria $\bar{X} = \frac{1}{4}\sum_{i=1}^{4} X_i$ ha distribuzione normale con $\mu_{\bar{X}} = E\left[\bar{X}\right] = \mu = 25$ e $\sigma^2_{\bar{X}} = var(\bar{X}) = \frac{\sigma^2}{n} = \frac{4}{4} = 1$. Dunque:

$$\bar{X} \sim \mathcal{N}(\mu_{\bar{X}} = 25, \sigma^2_{\bar{X}} = 1)$$

(b) Qual é la probabilitá che la percorrenza media sia superiore a 26 km/l?

Soluzione:
Dobbiamo calcolare

$$P(\bar{X} > 26)$$

Riportiamoci ad una normale standard $\mathcal{N}(0,1)$ in modo da poter consultare le tavole numeriche. Sia $Z_{\bar{X}} \sim \mathcal{N}(0,1)$.

$$P(\bar{X} > 26) = P(Z_{\bar{X}} > \frac{26 - 25}{1}) = P(Z_{\bar{X}} > 1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$$

Esercizio 6. Viene eseguita una rilevazione per studiare il numero di nascite maschili in famiglie con 5 figli. Esaminando in totale 1000 famiglie si ottiene il seguente risultato:

Num. Maschi	0	1	2	3	4	5
Frequenza	30	140	316	309	174	31

(a) Verificare se i dati empirici ottenuti si adattano ad una distribuzione binomiale ad un livello di significatività del 5%.

Soluzione: Definiamo la v.a. discreta $X = \{0,1,2,3,4,5\}$ che indica il numero di figli maschi in una famiglia con 5 figli.
Per comoditá notazionale riscriviamo la tabella come

x_i	0	1	2	3	4	5
O_i	30	140	316	309	174	31

con O_i la frequenza assoluta della classe i, $i = 1, \cdots, K = 6$, osservata nel campione di $N = \sum_{i=1}^{K} O_i = 1000$ famiglie. Il modello binomiale teorico di interesse é la legge che regola la distribuzione di X successi (num. figli maschi) su $n = 5$ prove (num. figli) :

$$X \sim Bin(x; n = 5, p)$$

(numero di teste su 5 lanci di una moneta o su 5 monete lanciate in parallelo). In questo caso il parametro p non é noto e dovrá essere stimato empiricamente dai dati.
La bontá dell'adattamento di tale modello binomiale ai dati empirici si puó verificare con un test Chi-quadro, dove le ipotesi sono

$$\begin{cases} H_0 & : \text{il modello Bin(x; n= 5, p) si adatta} \\ H_1 & : \text{il modello non si adatta.} \end{cases} \tag{5}$$

La statistica da utilizzare é

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}, \tag{6}$$

dove O_i, E_i sono rispettivamente le frequenze assolute osservate e le frequenze attese sulla base della distribuzione binomiale $Bin(n = 5,p)$. $K = 6$ é il numero di classi in tabella
La regione critica é quella a destra del parametro critico $\chi^2_{\nu,\alpha}$. Qui $\alpha = 0.05$ e i g.d.l. sono

$$\nu = K - 1 - m = 6 - 1 - 1 = 4$$

dove $m = 1$ perché il parametro teorico p non é noto e dovrá essere stimato dai dati. Dalla tavola della distribuzione Chi-quadro, $\chi^2_{4,0.05} = 9.488 \approx 9.49$.
Si rifiuta H_0 se

$$\chi^2 > \chi^2_{4,0.05} = 9.49$$

Per la $Bin(x; n = 5, p)$ sappiamo che il valore atteso si scrive

$$E[X] = np \implies p = \frac{E[X]}{n} = \frac{E[X]}{5}$$

Il campione é ampio ($N = \sum_{i=1}^{K} O_i = 1000$) e senza alcun problema $E[X]$ puó essere stimato dai dati in approssimazione Gaussiana cioé con la media campionaria

$$E[X] \approx \bar{x} = \frac{1}{N} \sum_{i=1}^{K} x_i \cdot O_i = 2.55,$$

da cui

$$p = \frac{2.55}{5} = 0.51$$

La distribuzione teorica si scrive pertanto:

$$P(X = x) = Bin(x; n = 5, p = 0.51) = \binom{5}{x} (0.51)^x (0.49)^{5-x}$$

Possiamo allora calcolare $P(X = 0) = \binom{5}{0}(0.51)^0(0.49)^5 = 0.0282$ e cosí i successivi

$$P(X = 1) = 0{,}1470; P(X = 2) = 0.3060; P(X = 3) = 0{,}3185; P(X = 4) = 0.1657; P(X = 5) = 0.0345.$$

Per le varie $i = 1, \cdots$,6 calcoliamo le frequenze teoriche come

$$E_i = P(X = x_i) \cdot N;$$

per esempio:

$$E_1 = P(X = 0) \cdot 1000 = 28.2 \approx 28,$$

Otteniamo cosí la tabella completa

x_i	0	1	2	3	4	5
O_i	30	140	316	309	174	31
E_i	28	147	306	318	166	35

Nessuna delle E_i é minore di 5, dunque non vi sono classi da accorpare.
Calcolando la statistica come in Equazione (6) si ottiene

$$\chi^2 = \frac{(30-28)^2}{28} + \frac{(140-147)^2}{147} + \cdots + \frac{(31-35)^2}{35} = 1.9$$

Essendo $\chi^2 = 1.9 < 9.49$ la statistica non appartiene alla regione di rifiuto, quindi si puó concludere che le fluttuazioni che osserviamo nei dati rispetto ai valori attesi teorici sono sostanzialmente dovute al caso e i dati si adattano bene al modello $Bin(x; n = 5, p = 0.51)$ ad un livello di significatività del 5%.