Metodi Matematici e Informatici per la Biologia
a.a. 2023-2024
(Agliari - Panati - Presilla - Simonella)
25 settembre 2024
Prova di Verifica III

Nome: **Cognome:** **Matricola:**

Esercizio 1

Tutte le unità di un grande allevamento avicolo sono state vaccinate e, dalle esperienze relative ad analoghi allevamenti, si stima che l'1.25% della popolazione si ammalerà nonostante la vaccinazione.

1.1 Considerato un campione di 120 unità estratto dall'allevamento, si indichi una ragionevole distribuzione di probabilità per il numero di unità, appartenenti a questo campione, che si ammaleranno nonostante la vaccinazione e se ne riporti l'espressione matematica esplicita.

Il numero di animali malati nel campione può essere considerato una variabile aleatoria di Poisson, che indichiamo con X, di media $\lambda = 120 \times 1.25/100 = 1.5$.
La forma esplicita è data da

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda} = \frac{1.5^n}{n!} e^{-1.5}$$

1.2 Si stimi la probabilità che nel campione prelevato nessuna unità si ammalerà.

$$P(X = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-1.5} \approx 0.22 \to 22\%$$

1.3 Si stimi la probabilità che, nel campione prelevato, al più due unità si ammaleranno.

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= e^{-\lambda} \left[\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} \right] \approx 0.22 \times [1 + 1.5 + 1.125] \approx 0.798 \to 79.8\% \end{aligned}$$

1.4 Si stimi la probabilità che, nel campione prelevato, almeno quattro unità si ammaleranno.

$$\begin{aligned} P(X \geq 4) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3) \\ &= 1 - [P(X = 0) + P(X = 1) + P(X = 2)] - \frac{\lambda^3}{3!} e^{-\lambda} \\ &\approx 1 - 0.798 - 0.124 \approx 0.078 \to 7.8\% \end{aligned}$$

Potremmo anche usare una distribuzione Binomiale.

Sia $X_i = \begin{cases} 1 & \text{con probabilità } p = 0.0125 \\ 0 & \text{" "} \quad 1-p \end{cases}$

la variabile aleatoria che descrive la possibilità di ammalarsi $(X_i = 1)$ dell'animale $i = 1, \dots, n = 120$.

Posto $S_n = \sum_{i=1}^{n} X_i$ abbiamo

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$P(S_n = 0) = \binom{n}{0} p^0 (1-p)^n = (1-p)^n \simeq 0.2210$$

$$P(S_n \geq 4) = 1 - \left[P(S_n = 0) + P(S_n = 1) + P(S_n = 2) + P(S_n = 3) \right]$$

$$P(S_n = 1) = n\, p\, (1-p)^{n-1} \simeq 0.3357$$

$$P(S_n = 2) = \frac{n(n-1)}{2} p^2 (1-p)^{n-2} \simeq 0.2528$$

$$P(S_n = 3) = \frac{n(n-1)(n-2)}{6} p^3 (1-p)^{n-3} \simeq 0.1259$$

$$P(S_n \geq 4) \simeq 1 - \left[0.2210 + 0.3357 + 0.2528 + 0.1259 \right]$$

$$= 0.064$$

Questi risultati sono molto simili a quelli ottenuti con la distribuzione di Poisson con $\lambda = np$ dove $n \gg 1$ e $p \ll 1$ poiché vale il seguente teorema

Legge degli eventi rari (teorema del limite di Poisson)

Sia p_n una successione in $[0,1]$ tale che $\lim_{n\to\infty} n p_n = \lambda > 0$ allora

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^k =$$

$$= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}(1+o(1))\right)^k \left(1-\frac{\lambda}{n}(1+o(1))\right)^{n-k}$$

$$= \lim_{n\to\infty} \frac{n^k + O(n^{k-1})}{k!} \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}(1+o(1))\right)^n$$

$$= \lim_{n\to\infty} \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}(1+o(1))\right)^n$$

$$= \lim_{n\to\infty} \frac{\lambda^k}{k!} e^{-\lambda}$$

Esercizio 2

La misurazione della concentrazione di particelle in sospensione eseguite in un complesso petrolchimico in 40 diversi momenti, fornisce i valori seguenti (misurati in mg/m^3)

13	108
29	148
85	59
35	27
0	32
8	5
2	31
19	42
45	31
3	17
37	205
73	11
46	97
17	37
29	68
14	50
0	21
2	89
31	64
2	33

2.1 Si calcolino media campionaria $\bar{X}$, mediana e deviazione standard S, ricordandone le definizioni.

Per il campione fornito, avendo $n = 40$ data, si ha:

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 41.6250\ mg/m^3$,

Mediana $= \frac{X_{20}+X_{21}}{2} = 31\ mg/m^3$, avendo ordinato i dati in maniera crescente $(X_1 \leq ... \leq X_{40})$

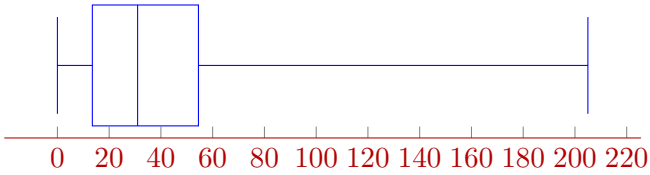
$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 42.4032\ mg/m^3$

2.2 Si dica se si tratta di un campione approssimativamente normale, fornendo *un* argomento per giustificare la risposta.

Il campione non appare approssimativamente normale.
Possibili argomenti:
l'istogramma presenta una evidente asimmetria a destra
il normal-plot presenta una concavità positiva
la regola 69-95-99.7 è violata
...
(Il campione è stato estratto da una distribuzione esponenziale)

2.3 Si rappresenti il boxplot del campione di dati, specificando tutti i relativi valori di riferimento.

Primo quartile: $Q_1 = 13.5\ mg/m^3$, Mediana: $31\ mg/m^3$, Terzo quartile: $Q_3 = 54.5\ mg/m^3$
Minimo: $= 0\ mg/m^3$, Massimo $= 205\ mg/m^3$



2.4 Si individui un intervallo per il valore assunto dalla variabile aleatoria in esame, entro cui si concentra il 68% circa dei dati.

Osservando che la variabile aleatoria assume solo valori positivi e che il 68-esimo percentile corrisponde a 44.1 mg/m^3, un possibile intervallo (in unità mg/m^3) è $[0, 44]$.
Attenzione: in questo caso l'intervallo $[\bar{X} - S, \bar{X} + S] = [-0.7782, 84.0282]$ corrisponderebbe ad una probabilità di circa l'86%.

2.5 Si calcolino, illustrando il procedimento, il 10−imo ed il 90−esimo percentile del campione di dati.
Calcolo la posizione del decimo percentile: $k = n \times \frac{10}{100} = 4$, essendo il valore intero il 10−imo sarà $\frac{X_k + X_{k+1}}{2} = \frac{2+2}{2} = 2mg/m^3$.
Analogamente, per il novantesimo percentile: $k = n \times \frac{90}{100} = 36$, essendo il valore intero il 90−esimo sarà $\frac{X_k + X_{k+1}}{2} = \frac{89+97}{2} = 93mg/m^3$.

Esercizio 3

I dati seguenti mettono in relazione x, il numero di incidenti stradali in diversi capoluoghi italiani, con y, il numero di veicoli immatricolati nello stesso capoluogo.

Numero incidenti (x)	Numero veicoli
28	1300
30	1400
30	1407
31	1700
31	1490
36	1400
40	1400
38	1560
31	1310
30	1335

3.1 Si calcoli il coefficiente di correlazione lineare e si valuti qualitativamente (e.g., forte, debole) la relazione tra le due variabili.

Coefficiente di correlazione lineare: $R = 0.2014$; poiché per definizione $R \in [-1, +1]$, il valore trovato indica una relazione relativamente debole. Il coefficiente è calcolato tramite

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{j=1}^{n}(Y_j - \bar{Y})^2}} \tag{1}$$

3.2 Quale percentuale di variabilità delle y è spiegata dal modello di regressione?

Il coefficiente di determinazione è pari a $R^2 = 0.0528$, da cui, per definizione, la percentuale cercata è pari a 5.28%.

3.3 Si individui il punto che presenta il residuo massimo e, nel caso in cui tale punto venisse escluso dall'analisi, si valuti la conseguente variazione percentuale del coefficiente angolare della retta di regressione.

Il punto che presenta il residuo massimo è $(31, 1700)$, il cui residuo vale circa 280.77 da confrontare con una media dei valori assoluti dei residui pari a circa 88.49. Includendo questo punto nell'analisi, si ottiene un coefficiente angolare per la retta di regressione che vale $m = 7.1207$, mentre escludendo tale punto si ottiene $m' = 10.552$, pertanto la variazione percentuale risulta $(m' - m)/m \approx 0.4819$.

3.4 Si può concludere che il modello di regressione è un buon modello per questi dati? Si fornisca *una* giustificazione a supporto delle risposta.

No, non è un buon modello in quanto abbiamo un coefficiente di determinazione relativamente basso, inoltre escludendo un dato si ha una variazione di circa il 50% nel valore del coefficiente angolare.

Esercizio 4

I dati che seguono mettono in relazione la frazione di minatori di carbone che mostra sintomi di pneumoconiosi con il numero di anni lavorati in miniera.

Anni di lavoro	Frazione di malati
5	0.0000
10	0.0090
15	0.0185
20	0.0672
25	0.1542
30	0.1720
35	0.1840
40	0.2105
45	0.3570
50	0.4545

4.1 Si determini l'equazione della retta di regressione e si calcoli il coefficiente di determinazione.

Retta di regressione: $y = 0.0095x - 0.0991$; $R^2 = 0.908$

4.2 Osservando il "qqplot" dei residui, si stabilisca se la distribuzione dei residui presenta una asimmetria e, in caso positivo, di che tipo. Si giustifichi la risposta.

La distribuzione non presenta particolari asimmetrie: i dati nel qqpplot sono allineati sulla bisettrice.

4.3 Si riporti valor medio, mediana e deviazione standard dei residui e si commenti se tali risultati sono compatibili con l'affermazione che il modello lineare è un buon modello per rappresentare la dipendenza di y da x.

La media è circa 0, la mediana è -8×10^{-4}, la deviazione standard è circa 0.05. La distribuzione dei residui suggerisce che il modello lineare sia un buon modello.

4.4 Si stimi la probabilità che un minatore di carbone, che ha lavorato per 42 anni, mostri sintomi della malattia.

La probabilità si può stimare come $y(42) \approx 0.30$, ovvero 30%.

Esercizio 5

Per condurre un esperimento è stato selezionato un campione casuale semplice costituito da 3000 persone, di cui 2000 non fumatori e 1000 fumatori. Queste persone sono state seguite per 10 anni, tenendo traccia di quanti, tra loro, hanno sviluppato un cancro ai polmoni. I risultati sono riportati nella seguente tabella:

	Fumatori	Non fumatori
Cancro ai polmoni	6	2
Niente cancro ai polmoni	994	1998

5.1 Si verifichi l'ipotesi nulla secondo cui il cancro ai polmoni ed il fumo sono indipendenti, fissando il livello di significatività all'1%. Si riportino i principali passaggi, con i relativi risultati intermedi.

Si utilizza il test del chi-quadro.
Sotto l'ipotesi nulla che cancro ai polmoni e fumo sono indipendenti, la statistica del test vale

$$\chi^2_{\text{test}} \simeq \frac{(6-2.67)^2}{2.67} + \frac{(2-5.33)^2}{5.33} + \frac{(994-997.33)^2}{997.33} + \frac{(1998-1994.67)^2}{1994.67} \simeq 6.26.$$

Siccome per 1 grado di libertà risulta $P(\chi^2 \geq 6.26) \simeq 0.0123$ (ovvero, siccome $0.01 = P(\chi^2 \geq 6.635)$ e $\chi^2_{\text{test}} < 6.635$), non possiamo rifiutare, a questo livello di significatività, che se una persona a caso contrae un tumore ai polmoni questo sia indipendente dal fatto che fumi o meno.

5.2 Si dica se il risultato cambia fissando il livello di significatività al 5%.

La statistica del test assume lo stesso valore, $\chi^2_{\text{test}} \simeq 3.841$, ma ora va confrontata con $\chi^2_{0.05,1} = 3.841$. Pertanto, a questo livello di significatività, possiamo rifiutare l'ipotesi che se una persona a caso contrae un tumore ai polmoni questo sia indipendente dal fatto che fumi o meno. Il calcolo numerico CHIDIST(3.841;1)=0.50013, conferma che possiamo rifiutare l'ipotesi H_0 a un livello di significatività del 5%.