# Esame Scritto del Sesto Appello

Tempo a disposizione: 2 ore

Riportare il numero di matricola **all'inizio** di ogni foglio. La soluzione di ogni esercizio deve essere scritta in modo chiaro e ordinato, e può non essere valutata se la calligrafia è illeggibile. Le soluzioni degli esercizi devono essere riportate sul foglio protocollo nell'ordine proposto, la soluzione di ogni esercizio deve iniziare in una nuova pagina.
Le soluzioni devono includere il procedimento dettagliato che porta alle risposte. Risposte corrette non adeguatamente motivate saranno penalizzate.
Non è permesso l'uso di note, appunti, manuali o materiale didattico di alcun tipo, al di fuori del formulario e delle tavole statistiche fornite assieme al compito. Non è permesso l'uso di dispositivi elettronici ad esclusiva eccezione di una calcolatrice non programmabile. L'infrazione di queste regole o la comunicazione con altri comportano l'annullamento del compito.

1. Per ognuna delle seguenti affermazioni si determini se essa è VERA oppure FALSA, motivando rigorosamente le risposte.

   (a) La correlazione di due variabili aleatorie $X, Y$ è invariante per trasformazioni lineari, ovvero non cambia se le si rimpiazza con le variabili $aX + b, cY + d$, $a, c > 0$ e $b, d \in \mathbb{R}$.
   VERO: come segue dalla definizione $\rho(aX+b, cY+d) = \mathrm{Cov}(aX+b, cY+d)/\sqrt{\mathrm{Var}(aX+b)\,\mathrm{Var}(cY+d)}$ e dalle relazioni $\mathrm{Cov}(aX+b, cX+d) = ac\,\mathrm{Cov}(X,Y)$ e $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$, $\mathrm{Var}(cY+d) = c^2\,\mathrm{Var}(Y)$.

   (b) Se la densità $f$ di una variabile aleatoria $X$ è pari, cioè $f(x) = f(-x)$, allora $X$ ha distribuzione Gaussiana standard.
   FALSO: è vero che la densità Gaussiana standard è pari, ma esistono altre densità pari, ad esempio la densità uniforme su $[-1, 1]$.

   (c) Data una variabile aleatoria doppia $(X, Y)$, se le sue componenti soddisfano la relazione $X + Y = 0$ allora sono indipendenti.
   FALSO: Infatti si ha $Y = -X$ e dunque $\mathrm{Cov}(X, Y) = -\mathrm{Var}(X)$, quindi a meno che $X$ non sia costante la covariazione non è nulla e dunque le variabili non sono indipendenti.

   (d) Si lanciano due monete (non truccate): le variabili di Bernoulli

   $$X = \begin{cases} 1 & \text{la prima è testa} \\ 0 & \text{altrimenti} \end{cases}, \quad Y = \begin{cases} 1 & \text{la seconda è testa} \\ 0 & \text{altrimenti} \end{cases}, \quad Z = \begin{cases} 1 & \text{solo una è testa} \\ 0 & \text{altrimenti} \end{cases},$$

   sono indipendenti.
   FALSO: infatti gli eventi $A = \{X = 1\}$, $B = \{Y = 1\}$ e $C = \{Z = 1\}$ soddisfano

   $$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2},$$

   quindi non sono indipendenti.

   (e) Se lo spazio campionario $\Omega$ è finito non può esistere una variabile aleatoria $X : \Omega \to \mathbb{N}$ con distribuzione geometrica.
   VERO: infatti se $\Omega$ è finito, allora $X$ può assumere solo un numero finito di valori, mentre una v.a. geometrica assume (con probabilità positiva) ogni valore intero positivo.

(f) Se $X_1, \dots, X_n$ è un campione Gaussiano $N(\mu, 1)$,

$$f(X_1, \dots, X_n) = \left(\frac{X_1 + \cdots + X_n}{n}\right)^2 - \frac{1}{n}$$

è uno stimatore corretto di $\mu^2$.
VERO: infatti $\bar{X}_n = \frac{X_1+\cdots+X_n}{n}$ ha distribuzione Gaussiana $N(\mu, 1/n)$, quindi

$$\mathbb{E}[f(X_1, \dots, X_n)] = \mathbb{E}[\bar{X}^2] - \frac{1}{n} = \mathrm{Var}(\bar{X}) + \mathbb{E}[\bar{X}]^2 - \frac{1}{n} = \frac{1}{n} + \mu^2 - \frac{1}{n} = \mu^2.$$

2. Si considerino 4 dadi equilibrati, ossia per ognuno la probabilità di cadere su ogni faccia è la stessa e i lanci dei quattro dadi sono tra loro indipendenti. I primi due dadi hanno le facce numerate da 1 a 6 come usuale, indichiamo con $A, B$ le variabili aleatorie che descrivono l'esito del loro lancio. Il terzo dado ha le facce numerate con 1,2,2,3,3,4 e il quarto ha le facce numerate con 1,3,4,5,6,8, indichiamo con $C, D$ le variabili aleatorie che descrivono l'esito del loro lancio.

(a) Calcolare le funzioni di massa $p_C, p_D$ degli esiti del terzo e quarto dado, e le funzioni di massa $p_{A+B}, p_{C+D}$ delle somme dei risultati di $A, B$ e $C, D$ (sono diverse tra loro?).
Per $p_C$ basta usare la frazione casi favorevoli su casi possibili,

$$p_C(1) = 1/6,\ p_C(2) = 2/6 = 1/3,\ p_C(3) = 2/6 = 1/3,\ p_C(4) = 1/6,$$

mentre $p_D$ è uniforme sugli esiti,

$$p_D(1) = p_D(2) = p_D(3) = p_D(4) = p_D(5) = p_D(6) = 1/6.$$

Poiché i lanci sono tutti indipendenti si applica la formula della convoluzione discreta,

$$p_{A+B}(n) = \sum_{k=1}^{n} p_A(k) p_B(n-k), \quad n = 2, \dots, 12,$$

da cui si ottiene con semplici calcoli

| $n$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_{A+B}$ | 1/36 | 2/36 | 3/36 | 4/36 | 5/36 | 6/36 | 5/36 | 4/36 | 3/36 | 2/36 | 1/36 |

Si procede similmente per la somma $C + D$, che può avere esiti $1, \dots, 12$ (il massimo risultato è 4+8) e dunque dalla formula

$$p_{C+D}(n) = \sum_{k=1}^{n} p_C(k) p_D(n-k), \quad n = 1, \dots, 12$$

si ricavano i valori di $p_{C+D}$, che risulta coincidere esattamente con $p_{A+B}$.

(b) Calcolare la probabilità che la somma $X = A + B$ dei risultati dei primi due dadi sia uguale alla somma $Y = C + D$ dei risultati degli altri due.
Vogliamo calcolare la probabilità dell'evento

$$\{X = Y\} = \{X = 1, Y = 1\} \cup \cdots \cup \{X = 12, Y = 12\},$$

in cui l'unione a destra coinvolge eventi tra loro disgiunti, per cui

$$\mathbb{P}(X=Y)=\mathbb{P}(X=1,Y=1)+\cdots+\mathbb{P}(X=12,Y=12)=p_X(1)p_Y(1)+\cdots+p_X(12)p_Y(12),$$

in cui il secondo passaggio usa l'indipendenza di $X,Y$. Dal punto precedente sappiamo che $p_X(k)=p_Y(k)$ per ogni $k$ e i relativi valori, sostituendo nella formula si ottiene $\mathbb{P}(X=Y)=73/648\simeq 11\%$.

(c) Si lancia il terzo dado $C$ per $N=120$ volte, con i seguenti risultati:

| esito | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| frequenza assoluta | 17 | 40 | 44 | 19 |

Si testi l'ipotesi nulla che il dado sia equilibrato.
Dobbiamo applicare un test $\chi^2$ di adattamento, per la distribuzione discreta $\tilde{p}(k)=\Pr(\text{esito } k)$, ipotesi nulla $\tilde{p}=p_C$ con $p_C$ al punto (a). Calcoliamo le frequenze assolute attese

| esito $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $E_k=Np_c(k)$ | 20 | 40 | 40 | 20 |

in particolare $Np_c(k)\geq 5$ e possiamo applicare il test $\chi^2$. La statistica di test è ($O_k$ sono le frequenze assolute osservate)

$$Q=\sum_{k=1}^{4}\frac{(O_k-E_k)^2}{E_k},$$

con distribuzione sotto $H_0$ approsimativamente $\chi^2$ a $4-1=3$ gradi di libertà. La regione critica a livello $\alpha$ è $C=\{Q>\chi^2_{3,1-\alpha}\}$, il $p$-value dei dati è $P(Q>v)$, con $v=0.9$ valore assunto dalla statistica con i nostri dati. In particolare, dalle tavole notiamo che il $p$-value deve essere maggiore di 0.8 (il software dà 0.82), perciò si accetta $H_0$ a ogni ragionevole livello.

3. Si vuole misurare il livello di concentrazione di PM2.5 (in $\mu g/m^3$) nell'aria. Lo strumento usato ha un errore di misurazione con distribuzione gaussiana di media pari alla concentrazione reale e deviazione standard pari a 1 $\mu g/m^3$. Su 100 misurazioni, risulta una concentrazione media di 10.5 $\mu g/m^3$.

(a) Fornire un intervallo di fiducia, di livello 0.95, per la concentrazione reale di PM2.5.
Stiamo cercando un intervallo di fiducia per la media di una popolazione gaussiana, deviazione standard nota $\sigma=1$. L'invervallo di fiducia cercato è ($n=100$, $\alpha=0.05$)

$$\left[\bar{X}\pm\frac{\sigma}{\sqrt{n}}q_{1-\alpha/2}\right]=\left[\bar{X}\pm 0.196\right].$$

Inserendo il valore $\bar{x}=10.5$, troviamo $[10.304, 10.696]$.

(b) Sulla base dei dati, c'è evidenza che la concentrazione reale sia superiore a 10 $\mu g/m^3$? Formulare un opportuno test di ipotesi di livello 0.01 (con $H_0$: concentrazione non superiore a 10) e applicarlo ai dati del campione.
Siamo in presenza di un test di ipotesi sulla media $m$ di una popolazione gaussiana, con deviazione standard $\sigma=1$ nota. L'ipotesi nulla è $H_0: m\leq 10 (=m_0)$, contro $H_1: m>32$. La regione critica di livello $\alpha=0.01$ è ($n=100$)

$$C=\{\frac{\sqrt{n}}{\sigma}(\bar{X}-m_0)>q_{1-\alpha}\}=\{(\bar{X}-10)>0.233\}$$

Applichiamo il test al dato $\bar{x} = 10.5$: i dati cadono nella regione critica $C$, quindi c'è evidenza, a livello 0.01, per concentrazione superiore a 10.
In alternativa, si può calcolare il p-value relativo a $\bar{x} = 10.5$, che è (chiamando $Z$ una v.a. normale standard)

$$\bar{\alpha} = P\left(Z > \frac{\sqrt{n}}{\sigma}(\bar{x} - m_0)\right) = 1 - \Phi(5) \approx 0,$$

quindi c'è evidenza a ogni livello ragionevole per una concentrazione superiore a 100.

(c) Lo strumento a nostra disposizione si è guastato, e per via di tagli al bilancio il nuovo strumento comprato in sostituzione è meno efficiente, in particolare ha deviazione standard pari a 2. Con il nuovo strumento, quante misurazioni dobbiamo fare per ottenere una stima con la stessa precisione del vecchio? cioè, fissato ad esempio il livello di fiducia 0.95, quante misurazioni è necessario fare con il nuovo strumento per produrre un intervallo di fiducia con la stessa ampiezza di cui al punto (a)?
La semi-ampiezza di un intervallo di fiducia di livello $\alpha$ è $\frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}$. Quindi, nel caso delle misurazioni con il vecchio strumento ($\sigma = 1$, $n = 100$), la semi-ampiezza è $0.1 \cdot q_{1-\alpha/2}$, mentre nel caso delle misurazioni con il nuovo strumento ($\sigma = 2$), la semi-ampiezza è $2/\sqrt{n} \cdot q_{1-\alpha/2}$. Dobbiamo quindi avere $2/\sqrt{n} = 0.1$, cioè $n = 20^2 = 400$ misurazioni.