Compito scritto dell'esame di Statistica e analisi dei dati 01 settembre 2022	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 e appunti.
- 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.
Un cappello contiene tre carte: una carta é nera su entrambi i lati; una carta é bianca su entrambi i lati; una carta é nera su un lato e bianca sull'altro. Le carte vengono mescolate nel cappello, poi una viene estratta a caso e appoggiata su di un tavolo.

(a) Se il lato visibile della carta é nero, qual é la probabilitá che l'altro lato sia bianco?

Soluzione: Uno studente potrebbe essere tentato di risolvere in pochi secondi il problema ragionando come segue: la carta di interesse deve essere o la nero / nero o la nero / bianco: queste hanno pari probabilitá, dunque la probabilitá che, osservato il nero, l'altro lato sia bianco é $\frac{1}{2}$.
Lo studente, successivamente, per verificare la sua conclusione, effettua una simulazione del gioco delle carte, ma ottiene il risultato mostrato in Figura 1.

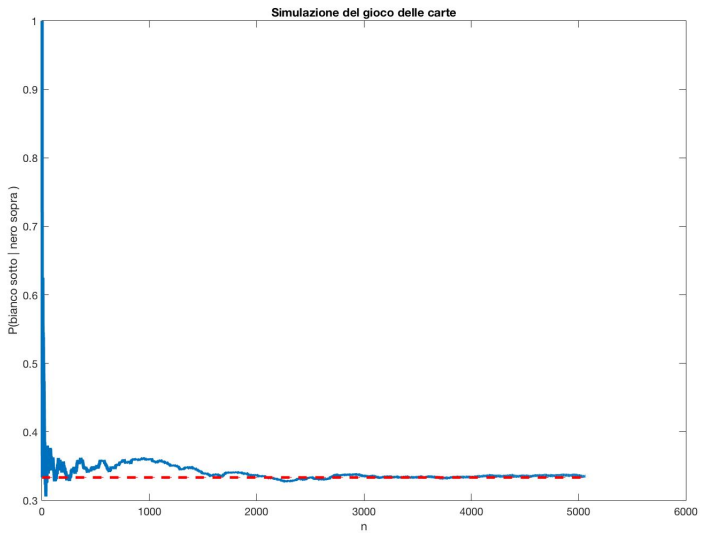

Figure 1: Simulazione del gioco delle tre carte: il valore di $P(\text{“bianco sotto”} \mid \text{“nero sopra”})$ valutato in termini di frequenza relativa, per un numero di prove n crescente, tende al valore di $0.333\cdots \approx \frac{1}{3}$

.

Il valore dell'approssimazione frequentistica di $P(\text{“bianco sotto”} \mid \text{“nero sopra”})$ tende a $\frac{1}{3}$.
Per capire bene come si arriva a questo risultato, mettiamo da parte l'intuizione e definiamo con precisione lo spazio campionario e gli eventi di interesse.
Identifichiamo i lati delle carte

- N_1 e N_2 per la carta nero / nero

- B_1 e B_2 per la carta bianco / bianco
- N_3 e B_3 per la carta nero / bianco

In questo caso lo spazio campionario S é

$$S = \{N_1,N_2, N_3, B_1,B_2,B_3\}$$

L'evento $N =$ "nero sopra" é definibile come $N = \{N_1,N_2, N_3\}$.
L'evento $B =$ "bianco sotto" é definibile come $B = \{N_3, B_1, B_2\}$
Pertanto:

$$P(B \mid N) = \frac{P(B \cap N)}{P(N)}$$

Poiché #("bianco sotto" $\cap$ "nero sopra") $= 1$ e #("nero sopra") $= 3$:

$$P(B \mid N) = \frac{1}{3}.$$

Esercizio 2.
Un robot autonomo controlla il livello di corrosione interna delle tubature dei sistemi di raffreddamento di una centrale nucleare. La corrosione non puó essere osservata direttamente, ma il robot effettua un test che puó fornisce indizi sulla specifica corrosione. Il test non é infallibile: nel 70% dei casi individua una corrosione quando é presente, ma puó anche sbagliare indicando una corrosione inesistente nel 20% dei casi. Risolvere il problema del robot a cui un operatore remoto, sotto l'ipotesi che il 10% delle tubature siano corrose, chiede di:

(a) Comunicargli la probabilitá che una parte delle tubature abbia delle corrosioni interne dopo che il test ha dato esito positivo (presenza di corrosioni). Determinare tale probabilitá.

Soluzione: Definiamo gli eventi $C =$ "la tubatura é corrosa" e $\overline{C} =$ "il test ha identificato la tubatura come corrosa". I dati del problema ci dicono che, a priori,

$$P(C) = 0.1$$

e dunque $P(\sim C) = 1 - 0.1 = 0.9$, dove $\sim C =$ "la tubatura non é corrosa". Inoltre

$$P(\overline{C} \mid C) = 0.7$$
$$P(\overline{C} \mid \sim C) = 0.2$$

Applicando la regola di Bayes:

$$P(C \mid \overline{C}) = \frac{P(\overline{C} \mid C)P(C)}{P(\overline{C})}$$

dove $P(\overline{C}) = P(\overline{C} \mid C)P(C) + P(\overline{C} \mid \sim C)P(\sim C) = 0.25$. Quindi

$$P(C \mid \overline{C}) = \frac{0.7 \times 0.1}{0.25} = 0.28$$

(b) Comunicargli la probabilitá che una parte delle tubature abbia delle corrosioni dopo che il test ha dato esito negativo (nessuna corrosione). Determinare tale probabilitá.

Soluzione: Si ripeta il procedimento precedente per il caso $P(C \mid \sim \overline{C})$

Esercizio 3.
Un sistema di comunicazione consiste di un buffer che ospita i pacchetti provenienti dalla sorgente, e da una linea di comunicazione che recupera i pacchetti dal buffer e li trasmette a un ricevitore. Il sistema opera a coppie di *time-slot*. Nel primo slot il sistema depone nel buffer un certo numero di pacchetti generati dalla sorgente secondo una distribuzione di Poisson di parametro μ. La capacitá del buffer (numero massimo di pacchetti) é pari b_{max}; i pacchetti che arrivano a buffer pieno vengono scartati. Nel secondo slot temporale il sistema trasmette c pacchetti al ricevitore, dove c é un intero $0 < c < b_{max}$, oppure trasmette quelli effettivamente presenti se in numero inferiore a c.

(a) Assumendo che all'inizio del primo slot il buffer sia vuoto, trovare la PMF del numero di pacchetti bufferizzati alla fine del primo slot e la PMF alla fine del secondo slot.

Soluzione: Sia X la V.A. discreta che indica il numero di pacchetti bufferizzati alla fine del primo slot temporale Δt_1. La Figura 2 mostra il suo intervalli di variazione relativamente ai parametri del problema c, b_{max} Per determinare la PMF di X, P_X é sufficiente determinare: 1) la probabilitá di avere bufferizzato $k = 0,1,2, \cdots, b_{max} - 1$ pacchetti; 2) la probabilitá di avere il buffer pieno che accade quando in Δt_1 vengono generati alla sorgente $k \geq b_{max}$ pacchetti.

1. $k = 0,1,2,\cdots, b_{max}-1$:
   La probabilitá di avere esattamente k pacchetti bufferizzati $P_X(X=k)$ é uguale alla probabilitá che k pacchetti siano stati generati alla sorgente durante Δt_1: ció implica, per quanto ci dice la specifica del problema, che X segua una legge di Poisson, $X \sim Pois(\mu)$ e dunque:
   $$P_X(X=k) = \frac{\mu^k}{k!}e^{-\mu} \tag{1}$$
   applicabile per $k = 0,1,2,\cdots, b_{max}-1$.
2. $k \geq b_{max}$:
   La probabilitá di avere il buffer pieno, ovvero che $X = b_{max}$ é la probabilitá che alla sorgente siano stati generati almeno b_{max} pacchetti, ovvero
   $$P_X(X=b_{max}) = \sum_{k=b_{max}}^{\infty} \frac{\mu^k}{k!}e^{-\mu} = 1 - \sum_{k=0}^{b_{max}-1} \frac{\mu^k}{k!}e^{-\mu}$$

Si noti come $P_X(X=0) + P_X(X=1) + \cdots + P_X(X=b_{max}-1) + P_X(X=b_{max}) = 1$, dunque P_X é una PMF correttamente normalizzata.
Chiamiamo ora Y la V.A. discreta che indica il numero di pacchetti bufferizzati alla fine del secondo slot temporale Δt_2. Il suo valore é determinabile come

$$Y = \text{ pacchetti bufferizzati durante } \Delta t_1 - \text{ pacchetti trasmessi durante } \Delta t_2 = X - min\{X,c\}.$$

Infatti se $X \leq c$, allora $min\{X,c\} = X$ e $Y = X - X = 0$. Se $X > c$, $min\{X,c\} = c$, quindi $Y = X - c$.
Per determinare la PMF di Y consideriamo i seguenti casi

1. $Y = 0$: la probabilitá che il buffer sia stato svuotato é uguale alla probabilitá che vi fossero $X \leq c$ pacchetti bufferizzati in Δt_1:
   $$P_Y(Y=0) = P_X(X \leq c) = \sum_{k=0}^{c} P_X(X=k) = \sum_{k=0}^{c} \frac{\mu^k}{k!}e^{-\mu}.$$
2. $0 < Y = k < b_{max} - c$:
   La probabilitá che al termine di Δt_2 vi siano esattamente $Y = k$ pacchetti con $k = 1,2,\cdots, b_{max}-c-1$ é uguale alla probabilitá che in Δt_1 siano stati bufferizzati $X = k + c$ pacchetti, con $k + c < b_{max}$
   $$P_Y(Y=k) = P_X(X=k+c) = \frac{\mu^{k+c}}{(k+c)!}e^{-\mu}$$
3. $Y = b_{max} - c$:
   E' il caso in cui in Δt_1 il buffer śtato riempito con $X = b_{max}$ pacchetti di cui c sono stati inviati in Δt_2:
   $$P_Y(Y = b_{max}-c) = P_X(X=b_{max}) = 1 - \sum_{k=0}^{b_{max}-1} \frac{\mu^k}{k!}e^{-\mu}$$

(b) Qual é la probabilitá che alcuni pacchetti vengano scartati durante il primo slot?
Soluzione: La probabilitá che alcuni pacchetti vengano scartati durante il primo slot é uguale alla probabilitá che piú di b_{max} pacchetti siano stati generati alla sorgente in Δt_1 ovvero

$$P_X(X > b_{max}) = \sum_{k=b_{max}+1}^{\infty} \frac{\mu^k}{k!}e^{-\mu} = 1 - \sum_{k=0}^{b_{max}} \frac{\mu^k}{k!}e^{-\mu}$$

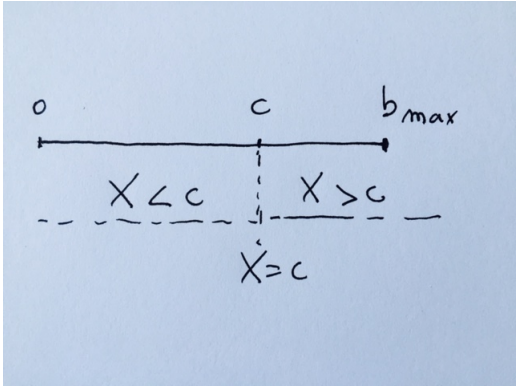

Figure 2: Valori che puó assumere la variabile aleatoria X nel problema del buffer

Esercizio 4.
Una linea di produzione fabbrica resistori da 1000 ohm (Ω). La fabbrica adotta una tolleranza del 10% (ovvero sono scartati i resistori con resistenza $R < 900\ \Omega$ e $R > 1100\ \Omega$).

(a) Sotto l'ipotesi che la resistenza R di un resistore sia una variabile aleatoria distribuita con legge normale di media 1000 Ω e varianza 2500, si calcoli la probabilitá che un resistore preso a caso venga scartato

Soluzione: Sia S l'evento in cui un resistore viene scartato. Allora:

$$S = \{R < 900\} \cup \{R > 1100\}$$

Poiché $\{R < 900\} \cap \{R > 1100\} = \emptyset$, si ha che

$$P(S) = P(\{R < 900\}) \ + P(\{R > 1100\}) = F_X(900) + (1 - F_X(1100)) = \Phi\left(\frac{900-\mu}{\sigma}\right) + \left(1 - \Phi\left(\frac{1100-\mu}{\sigma}\right)\right)$$

Poiché $\mu = 1000$, e $\sigma^2 = 2500 \implies \sigma = 50$, le variabili normalizzate sono

$$\frac{900-1000}{50} = -2$$
$$\frac{1100-1000}{50} = 2$$

.

Sfruttando la proprietá della cumulativa standardizzata $\Phi(-z) = 1 - \Phi(z)$, la precedente diventa:

$$P(S) = 2\,(1 - \Phi(2))$$

.

Accedendo alla tabella della Normale standard per $z = 2$, $\Phi(2) = 0.9772$ da cui $P(S) = 0.0455$

Esercizio 5.
Vengono sottoposti a confronto i consumi delle autovetture A e B alla velocitá costante di $120Km/h$. Si ritiene che i consumi dei due tipi di autovetture possa essere descritto da variabili aleatorie con distribuzione normale con la stessa varianza. L'auto di tipo A in 20 prove consuma mediamente $6.5litri/100Km$, quella di tipo B in 22 prove consuma mediamente $6.6litri/100Km$. Le relative varianze campionarie sono rispettivamente di 0.30 e 0.28.

(a) Possiamo ritenere che le due autovetture abbiano lo stesso consumo medio al livello di significativitá del 5%?

Soluzione:
Abbiamo che:

- $\bar{x}_A = 6.5, s_A^2 = 0.30, n_A = 20$
- $\bar{x}_B = 6.6, s_B^2 = 0.28, n_B = 22$

Siamo nel caso di varianze ignote ma eguali, e la differenza fra medie é valutabile mediante una statistica t-Student, dove i gradi di libertá ν sono dati da $\nu = n_A + n_B - 2 = 40$, $\alpha = 0.05$, per cui, $t_{\nu,\frac{\alpha}{2}} = t_{40,0.025} \approx 2.021$
La deviazione standard pooled vale:

$$S_{pool} = \sqrt{\frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A + n_B - 2}} = 0.5381$$

L'intervallo di confidenza per la differenza fra le medie $\mu_A - \mu_B$ a livello $(100)(1-\alpha) = (100)(1-0.05) = 95\%$ vale pertanto

$$\bar{x}_A - \bar{x}_B \pm t_{40,0.025} S_{pool}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = -0.1 \pm 0.33586$$

ovvero l'IC ha estremi $[-0.43596; 0.23596]$ per cui si puó ragionevolmente ritenere che le due autovetture abbiano differenza in consumo medio trascurabile.
Equivalentemente, si osserva che la statistica della differenza fra medie normalizzata

$$\left|\frac{\bar{x}_A - \bar{x}_B}{S_{pool}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}\right| = 0.612 < t_{40,0.025} = 2.021$$

Esercizio 6. Un costruttore sta considerando l'acquisto di speciali barre metalliche da due diversi fornitori. Un campione di 12 barre di lunghezza dichiarata pari a $127mm$ viene acquistato da ciascuno dei due fornitori e poi misurato. La deviazione standard della lunghezza delle barre del primo fornitore risulta essere $s_1 = 0.13mm$, mentre quella delle barre del secondo fornitore é di $s_2 = 0.17mm$.

(a) Questi dati indicano che la lunghezza di una barra del primo fornitore é soggetta a maggior variabilitá rispetto a quella del secondo fornitore? (Si assuma normalitá e un livello di significativitá 0.05)

Soluzione:
I dati del problema ci dicono che: $n_1 = 12$, $n_2 = 12$, $s_1^2 = (0.13)^2$, $s_2^2 = (0.17)^2$.
I gdl sono $\nu_1 = \nu_2 = 11$
Considerando la statistica $F = \frac{s_1^2}{s_2^2} = 0.585$ (nell'ipotesi $\frac{\sigma_1^2}{\sigma_2^2} = 1$, notiamo immediatamente che é largamente inferiore al valore critico $f_{0.025}(11, 11) = 3.53$ valore oltre il quale la probabilitá che i campioni siano stati generati da distribuzioni con $\sigma_1 = \sigma_2$ (ipotesi nulla) sarebbe molto bassa.
Verifichiamo con il calcolo dell'intervallo di confidenza (IC)
L'IC per il rapporto delle due varianze, con $\alpha = 0.05$ é:

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{0.025}(\nu_1, \nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{0.025}(\nu_2, \nu_1)$$

dove $f_{0.025}(\nu_2, \nu_1) = f_{0.025}(\nu_1, \nu_2) = f_{0.025}(11, 11) = 3.53$
Sostituendo:

$$0.16565 < \frac{\sigma_1^2}{\sigma_2^2} < 2.0642$$

L'intervallo ammette la possibilitá che $\frac{\sigma_1}{\sigma_2} = 1$, dunque i dati indicano che statisticamente non si puó rigettare l'ipotesi che le lunghezze delle barre dei due fornitori abbiano la stessa variabilitá.