Compito scritto dell'esame di Statistica e analisi dei dati 17 gennaio 2023	**Prof. Giuliano Grossi**	**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.
Dovete partecipare ad un torneo di scacchi in cui giocate contro altri tre giocatori (una sola partita con ciascuno). Conoscete le probabilitá di vincere con ognuno di loro e potete scegliere l'ordine in cui incontrarli. Si vince il torneo se si vincono due partite consecutive.

(a) Volendo massimizzare la probabilitá di vittoria, si mostri che l'ordinamento ottimale é quello che vi fa incontrare alla seconda partita il giocatore piú debole dei tre, mentre l'ordine in cui incontrate i restanti due giocatori non ha rilevanza.

Soluzione:
Denotiamo con

$$P(\texttt{"vittoria contro il giocatore incontrato alla partita i -esima"}) = P(i) = p_i$$

con $i = 1,2,3$.
La descrizione del problema ci dice che si vince la partita se si gioca la seconda partita con il giocatore piú debole - dunque con $p_2 \geq p_1, p_2 \geq p_3$ - e, **congiuntamente**, si vince con l'uno **oppure** l'altro giocatore incontrati alla partita 1 o 3. Quindi, formalmente, la probabilitá $P(V)$ di vincere il torneo é

$$P(V) = P(2 \cap (1 \cup 3)) = P(2)P(1 \cup 3) = P(2)\left[P(1) + P(3) - P(1) \cap P(3)\right] =$$
$$= p_2(p_1 + p_3 - p_1p_3).$$

L'ordinamento é ottimale se tale probabilitá non é inferiore alle probabilitá calcolate utilizzando i due ordinamenti alternativi $P(1 \cap (2 \cup 3)) = p_1(p_2 + p_3 - p_2p_3)$ e $P(3 \cap (2 \cup 1)) = p_3(p_2 + p_1 - p_2p_1)$, ovvero:

$$p_2(p_1 + p_3 - p_1p_3) \geq p_1(p_2 + p_3 - p_2p_3), \tag{1}$$
$$p_2(p_1 + p_3 - p_1p_3) \geq p_3(p_2 + p_1 - p_2p_1). \tag{2}$$

Infatti, la prima disuguaglianza é equivalente alla condizione

$$p_2 \geq p_1,$$

mentre la seconda disuguaglianza si riduce a

$$p_2 \geq p_3.$$

Le due condizioni (**??**) e (**??**) ci dicono che la probabilitá di vincere alla seconda partita deve essere massima, ovvero: per massimizzare la probabilitá di vincere il torneo il secondo giocatore da incontrare deve essere il piú debole dei tre.

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. Si dispone di un campione di 100 misure di una variabile temporale X di una popolazione di cui non conosciamo la distribuzione, ma la cui deviazione standard é nota e vale $\sigma_X = 120$ secondi. .

(a) Qual é la probabilitá che la media campionaria differisca per piú di 3 secondi dal valore atteso teorico (incognito) dei tempi misurati?

Soluzione:
Siano:

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

la media campionaria;

$$\mu_X = E[X]$$

il valore atteso teorico.
Il problema ci chiede di determinare la probabilitá

$$P_X\left(\left|\frac{S_n}{n} - \mu\right| > 3\right)$$

Non conosciamo la legge di probabilitá secondo cui si distribuisce la popolazione, ma il Teorema Centrale Limite (TCL) ci assicura che per n sufficientemente grande la distribuzione degli scarti fra media campionaria $\frac{S_n}{n}$ e valore atteso teorico μ_X segue una legge Gaussiana:

$$P\left(a \leq \frac{\sqrt{n}}{\sigma_X} \cdot \left(\frac{S_n}{n} - \mu\right) \leq b\right) \approx \Phi(b) - \Phi(a).$$

Nel nostro caso,

$$\frac{\sigma_X}{\sqrt{n}} = \frac{120}{\sqrt{100}} = 12$$

é la deviazione standard del campione.
Usando il TCL con $a = -3, b = 3$ e passando alla variabile standardizzata $Z_n \sim N(0,1)$

$$Z_n = \frac{\frac{S_n}{n} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}},$$

il problema iniziale si riduce a calcolare la seguente:

$$P_X\left(\left|\frac{S_n}{n} - \mu_X\right| > 3\right) = P_Z\left(\frac{\sigma_X}{\sqrt{n}}|Z_n| > 3\right) = P_Z(|Z_n| > 0.25).$$

Nel caso Gaussiano, $P_Z(|Z_n| > z) = 2(1 - \Phi(z))$, dunque :

$$P_Z(|Z_n| > 0.25) = 2(1 - \Phi(0.25)).$$

Usando le tabelle della CDF standard Φ leggiamo che

$$\Phi(0.25) = 0.5987.$$

Pertanto,

$$P_X\left(\left|\frac{S_n}{n} - \mu\right| > 3\right) = 2(1 - 0.5987) = 0.8026.$$

Esercizio 4.
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 5. Un produttore di batterie per auto garantisce che il suo prodotto dura in media 3 anni con una deviazione standard di 1 anno. Si campionano casualmente 5 batterie e ne risulta che abbiano una durata di $1.9, 2.4, 3.0, 3.5, 4.2$ anni.

(a) Assumendo che la vita della popolazione di batterie sia distribuita con legge normale, si tragga una conclusione con un livello di confidenza al 95% sull'asserzione del produttore che $\sigma = 1$.

Soluzione:
Utilizziamo l'intervallo di confidenza

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

con $\sigma^2 = \sigma = 1$
Determiniamo la varianza del campione s^2 con

$$s^2 = \frac{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n(n-1)}$$

Per $n = 5$:

$$\sum_{i=1}^{5} x_i = 1.9 + 2.4 + 3.0 + 3.5 + 4.2 = 15$$

e

$$\sum_{i=1}^{5} x_i^2 = 1.9^2 + 2.4^2 + 3.0^2 + 3.5^2 + 4.2^2 = 48.26$$

Pertanto

$$\begin{aligned} s^2 &= \frac{n\sum_{i=1}^{5} x_i^2 - (\sum_{i=1}^{5} x_i)^2}{n(n-1)} \\ &= \frac{5 \cdot 48.26 - (15)^2}{5(5-1)} \\ &= \frac{16.3}{20} \\ &= 0.815 \end{aligned}$$

Determiniamo $\alpha/2$. Da $95\% = 100(1-\alpha)\%$, si ha:

$$\begin{aligned} (100)(1-\alpha) &= 95 \\ 1-\alpha &= 0.95 \\ \alpha &= 0.05 \\ \alpha/2 &= 0.025 \end{aligned}$$

I g.d.l. sono $\nu = n - 1 = 4$. Usando la tabella della distribuzione Chi-quadro, il valore critico per 0.025 con 4 g.d.l. é $\chi^2_{0.025} = 11.143$, e $\chi^2_{1-0.025} = \chi^2_{0.975} = 0.484$.
L'intervallo di interesse é

$$\begin{aligned} \frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \\ \frac{(5-1)\cdot 0.815}{\chi^2_{0.025}} < \sigma^2 < \frac{(5-1)\cdot 0.815}{\chi^2_{0.975}} \\ \frac{4 \cdot 0.815}{11.143} < \sigma^2 < \frac{4 \cdot 0.815}{0.484} \\ 0.29 < \sigma^2 < 6.74 \end{aligned}$$

Poiché 1 cade nell'intervallo di confidenza al 95% si puó concludere che quanto dichiarato dal produttore é plausibile statisticamente.

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.