CS203 (2023) – First assignment

Total marks: 40

- **Note.** Answers without clear and concise explanations will not be taken into account. Use of immoral means will get severe punishment.

Name: Siddhant Suresh Jakhotiya

Roll No: 211030

Questions

1. **(10 marks)** Let $\{A_i\}_{i=1}^n$ be a family of sets indexed from 1 to n. Let $I \subseteq [n]$ be a subset of index set. Let B be the event when only and all A_i's from I have happened, i.e.,

$$B = (\cap_{i\in I} A_i) \cap (\cap_{i\notin I} A_i^c) .$$

Notice that B is subset of $\cap_{i\in I} A_i$, but need not be equal to it. Show that,

$$P(B) = \sum_{J\supseteq I} (-1)^{|J|-|I|} P(\cap_{i\in J} A_i).$$

Hint: can you define some new sets in terms of A_i, such that, this problem looks like inclusion-exclusion?

Solution:
B = $(\cap_{i\in I} A_i) \cap (\cap_{i\notin I} A_i^c)$
B = $(\cap_{i\in I} A_i) \cap (\cup_{i\notin I} A_i)^c$ (By Demorgan's Law)
Define C = $(\cup_{i\notin I} A_i)^c$ ⇒ B = $(\cap_{i\in I} A_i) \cap C$

Any set T in the sample space S can be defined as T = T∩S
Now, for any set Q in S, S = Q∪Q^c ⇒ T = T∩S = T∩(Q∪Q^c) = (T∩Q) ∪ (T∩Q^c)
Since the sets are equal, their probability must be equal ⇒ P(T) = P((T∩Q)∪(T∩Q^c)). Notice that T∩Q and T∩Q^c are disjoint sets.
⇒P((T∩Q)∪(T∩Q^c)) = P(T∩Q) + P(T∩Q^c)
Consider T as $\cap_{i\in I} A_i$ and Q as $(\cup_{i\notin I} A_i)^c$
⇒ $P(\cap_{i\in I} A_i) = P(\cap_{i\in I} A_i \cap (\cup_{i\notin I} A_i)^c) + P(\cap_{i\in I} A_i \cap (\cup_{i\notin I} A_i))$
P($\cap_{i\in I} A_i$) = P(B) + P($\cap_{i\in I} A_i \cap (\cup_{i\notin I} A_i)$) —(1)
Note that $\cap_{i\in I} A_i \cap (\cup_{i\notin I} A_i) = \cup_{j\notin I}(A_j \cap (\cap_{i\in I} A_i))$
(Distribution of union over the intersection operation)

Now, by the inclusion-exclusion principle,
P($\cup_{j\notin I}(A_j \cap (\cap_{i\in I} A_i))) = \sum_{S\subseteq I^c, S\neq\phi}(-1)^{|S|+1}P(\cap_{j\in S}(A_j(\cap_{i\in I}A_i)))$
$= \sum_{S\subseteq I^c, S\neq\phi}(-1)^{|S|+1}P((\cap_{j\in S}A_j) \cap (\cap_{i\in I}A_i))$
$= \sum_{S\subseteq I^c, S\neq\phi}(-1)^{|S|+1}P(\cap_{i\in S\cup I}A_i)$
For any $S \subseteq I^c$, define $J = I \cup S \Rightarrow$ |J| = |I| + |S| since I$\cap S = \phi$
$\Rightarrow$ P($\cup_{j\notin I}(A_j \cap (\cap_{i\in I} A_i))) = \sum_{S\subseteq I^c}(-1)^{|S|+1}P((\cap_{j\in S}A_j) \cap (\cap_{i\in I}A_i)) - (-1)^1 P(\cap_{i\in I}A_i)$
(adding the case of S=ϕ to the summation and subtracting it explicitly)
$\Rightarrow$ P($\cup_{j\notin I}(A_j \cap (\cap_{i\in I} A_i))) = \sum_{I\subseteq J}(-1)^{|J|-|I|+1}P(\cap_{i\in J}A_i) + P(\cap_{i\in I}A_i)$
From (1), P(B) = P($\cap_{i\in I}A_i$) - P($\cup_{j\notin I}(A_j \cap (\cap_{i\in I}A_i))$)
$\Rightarrow$ P(B) = -$\sum_{I\subseteq J}(-1)^{|J|-|I|+1}P(\cap_{i\in J}A_i)$

$$\Rightarrow \text{P(B)} = \sum_{I\subseteq J}(-1)^{|J|-|I|}P(\cap_{i\in J}A_i)$$

□

2. **(15 marks)** To study the efficacy of two tests for a set of three diseases, ICMR conducted trials with 10000 patients having one of these diseases. We restrict our universe to set of people having one of these diseases. The number of people who had these diseases, and the outcome of the result of these two tests on those patients is given in the table below.

Disease	Numbers having this disease	Result ++	Result +-	Result -+	Result –
d_1	3215	2110	301	704	100
d_2	2125	396	132	1187	410
d_3	4660	510	3568	73	509

Assume that these number accurately reflect the probability of people having a certain disease and efficacy of the tests. For a new patient (assuming he has one of the diseases), given test results, we want to estimate the probability of having d_1,d_2 or d_3. In other words fill out the following table.

Outcome	d_1	d_2	d_3
+ +			
+ -			
- +			
- -			

Solution:

Define the events as follows:

- D: a person has one of the diseases
- D_i: a person has disease i
- PP: both the tests return a positive result

- PN: test 1 returns positive, test 2 returns negative
- NP: test 1 returns negative, test 2 returns positive
- NN: both the tests return a negative result

Note that our universe has been restricted to people having one of the 3 diseases
⇒ P(D) = 1.
We have been given the test results of 10k people that we are to assume as a representative of the entire population of the people.

So, $P(D_1) = \frac{3215}{10000}$, $P(D_2) = \frac{2125}{10000}$ and $P(D_3) = \frac{4660}{10000}$.

From the table containing the test results, $P(PP|D_1) = \frac{P(PP \cap D_1)}{P(D_1)} = \frac{2110/10000}{3215/10000} = \frac{422}{643}$
Similarly, $P(PP|D_2) = \frac{P(PP \cap D_2)}{P(D_2)} = \frac{396}{2125}$
And, $P(PP|D_3) = \frac{P(PP \cap D_3)}{P(D_3)} = \frac{51}{466}$

$P(PN|D_1) = \frac{P(PN \cap D_1)}{P(D_1)} = \frac{301/10000}{3215/10000} = \frac{301}{3215}$
$P(PN|D_2) = \frac{P(PN \cap D_2)}{P(D_2)} = \frac{132}{2125}$
$P(PN|D_3) = \frac{P(PN \cap D_3)}{P(D_3)} = \frac{3568}{4660}$

$P(NP|D_1) = \frac{P(NP \cap D_1)}{P(D_1)} = \frac{704/10000}{3215/10000} = \frac{704}{3215}$
$P(NP|D_2) = \frac{P(NP \cap D_2)}{P(D_2)} = \frac{1187}{2125}$
$P(NP|D_3) = \frac{P(NP \cap D_3)}{P(D_3)} = \frac{73}{4660}$

$P(NN|D_1) = \frac{P(NN \cap D_1)}{P(D_1)} = \frac{100/10000}{3215/10000} = \frac{20}{643}$
$P(NN|D_2) = \frac{P(NN \cap D_2)}{P(D_2)} = \frac{82}{425}$
$P(NN|D_3) = \frac{P(NN \cap D_3)}{P(D_3)} = \frac{509}{4660}$

By the partition rule, $P(PP) = P(PP|D_1) \times P(D_1) + P(PP|D_2) \times P(D_2) + P(PP|D_3) \times P(D_3)$
$(sinceD_1 \cup D_2 \cup D_3 = D$, the sample space and D_i's are disjoint)

$$P(PP) = \frac{2110}{3215} \times \frac{3215}{10000} + \frac{396}{2125} \times \frac{2125}{10000} + \frac{510}{4660} \times \frac{4660}{10000}$$

$\Rightarrow P(PP) = 0.3016$
We can similarly model the probability of the events PN, NP and NN.

$$P(PN) = \frac{301}{3215} \times \frac{3215}{10000} + \frac{132}{2125} \times \frac{2125}{10000} + \frac{3568}{4660} \times \frac{4660}{10000}$$

$\Rightarrow P(PN) = 0.4001$

$$P(NP) = \frac{704}{3215} \times \frac{3215}{10000} + \frac{1187}{2125} \times \frac{2125}{10000} + \frac{73}{4660} \times \frac{4660}{10000}$$

$\Rightarrow P(NP) = 0.1964$

$$P(NN) = \frac{100}{3215} \times \frac{3215}{10000} + \frac{410}{2125} \times \frac{2125}{10000} + \frac{509}{4660} \times \frac{4660}{10000}$$

$\Rightarrow P(NN) = 0.1019$

By Baye's Theorem, $P(D_1|PP) = \dfrac{P(PP|D_1) \times P(D_1)}{P(PP)}$

$$\Rightarrow P(D_1|PP) = \frac{\frac{2110}{3215} \times \frac{3215}{10000}}{\frac{2110}{10000} + \frac{396}{10000} + \frac{510}{10000}} = 0.69960$$

Now, $P(D_2|PP) = \dfrac{P(PP|D_2 \times P(D_2)}{P(PP)}$

$$\Rightarrow P(D_2|PP) = \frac{\frac{396}{2125} \times \frac{2125}{10000}}{\frac{2110}{10000} + \frac{396}{10000} + \frac{510}{10000}} = 0.13130$$

And, $P(D_3|PP) = \dfrac{P(PP|D_3 \times P(D_3)}{P(PP)}$

$$\Rightarrow P(D_3|PP) = \frac{\frac{510}{4660} \times \frac{4660}{10000}}{\frac{2110}{10000} + \frac{396}{10000} + \frac{510}{10000}} = 0.16910$$

And similarly for all others,

$P(D_1|PN) = \dfrac{P(PN|D_1) \times P(D_1)}{P(PN)}$

$$\Rightarrow P(D_1|PN) = \frac{\frac{301}{3215} \times \frac{3215}{10000}}{\frac{301}{10000} + \frac{132}{10000} + \frac{3568}{10000}} = 0.07523$$

$P(D_2|PN) = \dfrac{P(PN|D_2) \times P(D_2)}{P(PN)}$

$$\Rightarrow P(D_2|PN) = \frac{\frac{132}{2125} \times \frac{2125}{10000}}{\frac{301}{10000} + \frac{132}{10000} + \frac{3568}{10000}} = 0.03299$$

$$P(D_3|PN) = \frac{P(PN|D_3) \times P(D_3)}{P(PN)}$$

$$\Rightarrow P(D_3|PN) = \frac{\frac{3568}{4660} \times \frac{4660}{10000}}{\frac{301}{10000} + \frac{132}{10000} + \frac{3568}{10000}} = 0.89178$$

$$P(D_1|NP) = \frac{P(NP|D_1) \times P(D_1)}{P(NP)}$$

$$\Rightarrow P(D_1|NP) = \frac{\frac{704}{3215} \times \frac{3215}{10000}}{\frac{704}{10000} + \frac{1187}{10000} + \frac{73}{10000}} = 0.35845$$

$$P(D_2|NP) = \frac{P(NP|D_2) \times P(D_2)}{P(NP)}$$

$$\Rightarrow P(D_2|NP) = \frac{\frac{1187}{2125} \times \frac{2125}{10000}}{\frac{704}{10000} + \frac{1187}{10000} + \frac{73}{10000}} = 0.60438$$

$$P(D_3|NP) = \frac{P(NP|D_3) \times P(D_3)}{P(NP)}$$

$$\Rightarrow P(D_3|NP) = \frac{\frac{73}{4660} \times \frac{4660}{10000}}{\frac{704}{10000} + \frac{1187}{10000} + \frac{73}{10000}} = 0.03717$$

$$P(D_1|NN) = \frac{P(NN|D_1) \times P(D_1)}{P(NN)}$$

$$\Rightarrow P(D_1|NN) = \frac{\frac{100}{3215} \times \frac{3215}{10000}}{\frac{100}{10000} + \frac{410}{10000} + \frac{509}{10000}} = 0.09814$$

$$P(D_2|NN) = \frac{P(NN|D_2) \times P(D_2)}{P(NN)}$$

$$\Rightarrow P(D_2|NN) = \frac{\frac{410}{2125} \times \frac{2125}{10000}}{\frac{100}{10000} + \frac{410}{10000} + \frac{509}{10000}} = 0.40236$$

$$P(D_3|NN) = \frac{P(NN|D_3) \times P(D_3)}{P(NN)}$$

$$\Rightarrow P(D_3|NN) = \frac{\frac{509}{4660} \times \frac{4660}{10000}}{\frac{100}{10000} + \frac{410}{10000} + \frac{509}{10000}} = 0.49951$$

To summarize,

Outcome	d_1	d_2	d_3
+ +	**0.69960**	**0.13130**	**0.16910**
+ -	**0.07523**	**0.03299**	**0.89178**
- +	**0.35845**	**0.60438**	**0.03717**
- -	**0.09814**	**0.40236**	**0.49951**

□

3. **(5 + 10 marks)** Your friend is trying to invest in NFTs. In particular, he is investing in a scheme where you get a random NFT from a set of n NFTs by giving Rs 50. The probability of getting a particular NFT is $\frac{1}{n}$. If you get a new NFT that wasn't in your collection already, then it is added to your collection, otherwise that buy is a waste. The company has an offer where they give Rs. 100 for each NFT if you have the complete set, and hence "double" the money. Let X be the random variable which counts the number of times your friend has to request NFT's so that he has a complete set. You have decided to show your friend that this is a losing deal and save him from the scam. Calculate the expectation and variance of X.

Solution:
Define a random variable as follows:
X_i: The number of **additional** buys to get the i^{th} distinct NFT after having obtained i - 1 distinct NFTs.
P(X_j = k) $= (\frac{j-1}{n})^{(k-1)}(1 - \frac{(j-1)}{n})$
Thus, E[X_j] $= \sum_{r=1}^{\infty} r \times P(X_j = k)$

$\Rightarrow$ E[X_j] $= \sum_{r=1}^{\infty} r \times (\frac{k-1}{n})^{r-1} \times (1 - \frac{k-1}{n}) = (1 - \frac{k-1}{n}) \sum_{r=1}^{\infty} r \times (\frac{k-1}{n}^{r-1}$
This is an AGP. Upon solving it, we get
E[X_j] $= \frac{1}{1 - \frac{k-1}{n}}$
Now, define a function of random variable X $= X_1 + X_2 + ... + X_n$
X is thus the total number of buys required to get the entire collection of n stickers.
E[X] = E[$X_1 + X_2 + ... + X_n$] = E[X_1] + E[X_2] + ... + E[X_n] (By linearity of Expectation)
So, the expected number of total buys required to complete the collection is

E[X] $= \sum_{i=1}^{n} \frac{1}{1-\frac{i-1}{n}} = n\sum_{i=1}^{n}\frac{1}{i}$

Now the expected cost incurred to complete 50× E[X], since each buy costs my friend Rs. 50.
Now, define a random variable
Also, the reward after collecting the collection is 100n, irrespective of the number of buys required to complete the collection. Thus my friend has an expected earning of

100n - 50E[X]

after having completed the collection.
⇒The expected profit is 50n× (2 - $\sum_{i=1}^{n}\frac{1}{i}$)

For values of n≥8, the profit is negative. Whenever the expected profit is negative, the deal shouldn't be made. Thus it is a losing deal.

Before finding the variance of X, we will start by proving a claim:
If X_1, X_2, ... , X_n be independent variables then

$$\text{Var}[X_1 + X_2 + ... + X_n] = \text{Var}[X_1] + \text{Var}[X_2] + ...\ \text{Var}[X_n]$$

<u>Proof</u>:
Var$[\sum_i X_i]$ = E$[\sum_i (X_i)^2]$ - E$[\sum_i X_i]^2$
$= \sum_i E[X_i^2] + \sum_{i\neq j} E[X_iX_j] - \sum_i E[X_i]^2 - \sum_{i\neq j} E[X_i]E[X_j]$
$= (\sum_i E[X_i^2] - \sum_i E[X_i]^2) + \sum_{i\neq j}(E[X_iX_j] - E[X_i]E[X_j])$
Since $X_i and X_j$ are independent for $i \neq j$, we have,
$E[X_iX_j] = E[X_i]E[X_j] \forall i \neq j \Rightarrow Var[\sum_i X_i] = \sum_i Var(X_i)$
Hence Proved.

Now, we need to calculate Var(X_k).
$Var(X_k) = E[X_k^2] - E[X_k]^2$. We know E[$X_k$].
E$[X_k^2] = \sum_{r=1}^{\infty} r^2 \times P(X_k = r)$
$= (1 - \frac{k-1}{n})\sum_{r=1}^{\infty} r^2(\frac{k-1}{n})^{r-1}$
Solving this AGGP in a similar fashion as above we get,
E$[X_k^2] = \frac{1+\alpha}{(1-\alpha)^2}$, where $\alpha = \frac{k-1}{n}$
Var$[X_k] = \frac{n(k-1)}{(n-k+1)^2}$. Thus, Var(X) = (X_k)
Var$[X_k] = n^2\sum_{k=1}^{n}\frac{1}{k^2} - n\sum_{k=1}^{n}\frac{1}{k}\ \frac{n^2\pi^2}{6} - nln(n)$ □