MTH 201: Probability and Statistics

Section A End Semester Exam

13/06/2023

Sanjit K. Kaul

No books, notes, or devices are allowed. Just a pen/ pencil and eraser. Your own work alone must be in your sheet. Explain your answers. Show your steps. Don't waste time on algebra unless straightforward. You have about 120 minutes.

Question 1. 25 **marks** A teaching assistant (TA) collects a stack of sheets for evaluation. The stack has a total of k sheets. The TA doesn't know the total number of sheets in the stack and must count the sheets in the stack. The TA proceeds sheet by sheet from top of the stack to its bottom. The TA accidentally skips counting any sheet in the stack with probability p independently of other sheets. Answer the following questions.

(a) (10 marks) The TA reports a count of the number of sheets after counting the stack. Derive the PMF of the count.

(b) (5 marks) Suppose the TA repeats the process of counting sheets in the stack, one or more times, till the TA gets the correct count. Derive the PMF of the total number of times the TA carries out the process of counting sheets in the stack.

(c) (10 marks) Suppose the TA decides to carry out the process of counting the number of sheets in the stack a total of $m > 1$ times. Having obtained m counts, the TA chooses the maximum of the m counts to be the correct count of sheets in the stack. Derive the PMF of this maximum.

Question 2. 25 **marks** Lazy must spend 8 hours at work. He decides to spend 3 of the 8 hours on social media (SM) and 5 hours waiting for the weekend (WW). At the beginning of any hour, Lazy chooses either SM or WW randomly from the number of SM and WW hours remaining in the day. Let us index the hours at work as $1, 2, \ldots, 8$. The RV $X_i = 1$, $i \in \{1, 2, \ldots, 8\}$, if Lazy chooses WW in the i^{th} hour. Else, $X_i = 0$. Answer the following questions.

(a) (5 marks) Derive the marginal PMFs of the RVs X_i, $i = 1, 2, \ldots, 8$. [*Hint*: Think counting/ combinatorics]

(b) (5 marks) Derive the joint PMFs of the RVs X_i and X_j, for $i \neq j$.

(c) (2.5 marks) Are the RVs $X_1, X_2, \ldots, X_8$ mutually independent? Justify your answer.

(d) (2.5 marks) Derive the expected value of the sum $X_1 + X_2 + \ldots + X_8$.

(e) (5 marks) Derive the variance of the sum $X_1 + X_2 + \ldots + X_8$.

(f) (5 marks) Suppose Lazy chooses to spend at least one hour on SM before spending the first hour on WW. Calculated the conditional expected value of X_3.

Question 3. 20 **marks** A virus is known to mutate into variants over months. Any person is infected by the virus within 0 or more integer number of months of the virus having been first detected. The number of months is distributed as a Poisson random variable with expected value $\alpha = 1$ and PMF

$\alpha^k e^{-\alpha}/k!$, $k = 0, 1, \ldots$. A person infected by the virus within m months, $m = 0, 1, \ldots$, takes an additional exponentially distributed months with rate $\lambda_m = m + 1$ to show symptoms of infection. The PDF of the exponential RV is $\lambda_m e^{-\lambda_m x}$, $x \geq 0$, and its expected value is $1/\lambda_m$. Derive the expected value of the months, since the virus is first detected, any person takes to show symptoms of infection by the virus.

Question 4. 10 **marks** You arrive at a bus stop knowing that the time you must wait for a bus is distributed as an exponential RV with PDF $f_X(x) = \lambda e^{-\lambda x}, x \geq 0$. The bus hasn't arrived for the first s seconds of your wait. Derive the conditional CDF of X. Repeat the above under the assumption that you must wait for a bus for a time that is uniformly distributed over $(0, 120)$. Note that $s \in (0, 120)$.

Question 5. 20 **marks** An external agency is visiting IIIT-Delhi to evaluate students at the university. A student's evaluation results in a score that is a sum of the marks obtained by the student in humanities and the marks obtained in sciences. The agency would like the average score, where the average is calculated over scores of all students in the institute. Given the large number of students, the agency decides to instead estimate the average. The agency is able to conduct the sciences exam for 50 students and the humanities exam for 25 students. They ask IIIT-Delhi to provide lists of randomly chosen students that will take the two exams.

Assume that the marks of any student i in the sciences exam is given by RV X_i and marks in the humanities exam is given by Y_i. Marks obtained by students in sciences are independent of other marks and identically distributed as the RV X. Marks obtained by students in humanities are independent of other marks and identically distributed as the RV Y. Propose an unbiased estimator for the average. You must show that your estimator is unbiased.

Assume that $\text{Var}[X] = \text{Var}[Y] = 25$. Apply the Chebyshev's inequality to derive the confidence interval $2c$ for a confidence of 95%. The Chebyshev's inequality states that for any RV Z, $P[|Z - E[Z]| \geq c] \leq \text{Var}[Z]/c^2$, for any $c > 0$.

Question 1. 25 **marks** A teaching assistant (TA) collects a stack of sheets for evaluation. The stack has a total of k sheets. The TA doesn't know the total number of sheets in the stack and must count the sheets in the stack. The TA proceeds sheet by sheet from top of the stack to its bottom. The TA accidentally skips counting any sheet in the stack with probability p independently of other sheets. Answer the following questions.

(a) (10 marks) The TA reports a count of the number of sheets after counting the stack. Derive the PMF of the count.

(b) (5 marks) Suppose the TA repeats the process of counting sheets in the stack, one or more times, till the TA gets the correct count. Derive the PMF of the total number of times the TA carries out the process of counting sheets in the stack.

(c) (10 marks) Suppose the TA decides to carry out the process of counting the number of sheets in the stack a total of $m > 1$ times. Having obtained m counts, the TA chooses the maximum of the m counts to be the correct count of sheets in the stack. Derive the PMF of this maximum.

(a) The TA skips a sheet with probability p independently of other sheets. The total no. of sheets is k.

The TA will come up with a count in the set $\{0, 1, 2, \ldots, k\}$

Let the count be X.

$$S_X = \{0, 1, 2, \ldots, k\}$$ At most 1

$$P[X=0] = P[\text{TA skips all } k \text{ sheets}] = p^k.$$

$$P[X=1] = {}^kC_1\, p^{k-1}(1-p)$$

$\vdots$

Some examples At most 2

We have

$$P[X=x] = \begin{cases} {}^kC_x\, p^{x}(1-p)^{k-x} & x \in \{0,1,2,\ldots,k\} \\ 0 & \text{otherwise.} \end{cases}$$ At most 10

As you may have guessed, X is a Binomial$(k, (1-p))$ RV. At most 1.

(b) The TA must count the stack again and again till the first time the count is k.

The success event of interest occurs when the TA doesn't skip any sheet when counting.

The probability of the event is $(1-p)^k$.

Let Y be the no. of times the TA must count the stack. Let $q = (1-p)^k$.

$$S_Y = \{1, 2, \ldots\}$$ At most 1

$$P[Y=1] = q$$

$$P[Y=2] = (1-q)\,q$$

$\vdots$

Some examples At most 1

$$P[Y=y] = \begin{cases} (1-q)^{y-1} q & y = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \left(1-(1-p)^k\right)^{y-1}(1-p)^k & y = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$ At most 5

Y is geometric$((1-p)^k)$ At most 1

(c) Let $Z_1, Z_2, \ldots, Z_m$ be the m counts. Each is a Binomial$(k, (1-p))$ RV. We want the PMF of

$$Z = \max(Z_1, Z_2, \ldots, Z_m)$$ At most 2

Note that Z has a range space

$$S_Z = \{0, 1, \ldots, k\}.$$ 1

$$P[Z=0] = P[\max(Z_1, Z_2, \ldots, Z_m) = 0]$$

$$= P[Z_1 = 0, Z_2 = 0, \ldots, Z_m = 0]$$

$$= (P[Z_1=0])(P[Z_2=0])\cdots(P[Z_m=0])$$ $\because Z_i$ are indep. RVs

$$= (p^k)(p^k)\cdots(p^k)$$

$$= (p^k)^m.$$

$$P[Z=1] = P[\max(Z_1, Z_2, \ldots, Z_m) = 1]$$

$= P[(m-1)$ of the $Z(s)$ are ≤ 1 & one is exactly $1]$

$$= {}^mC_1\, (P[Z_i \leq 1])^{m-1} P[Z_i = 1]$$

$\vdots$

Some examples of prob calculations (if provided) 3

$$P[Z=z] = \begin{cases} {}^mC_1\, (P[Z_j \leq z])^{m-1} P[Z_i = z] & z = 1, 2, \ldots, k \\ (P[Z_i = 0])^m = (p^k)^m & z = 0 \\ 0 & \text{otherwise} \end{cases}$$ 4 1 At most 8

Question 2. 25 **marks** Lazy must spend 8 hours at work. He decides to spend 3 of the 8 hours on social media (SM) and 5 hours waiting for the weekend (WW). At the beginning of any hour, Lazy chooses either SM or WW randomly from the number of SM and WW hours remaining in the day. Let us index the hours at work as $1, 2, \ldots, 8$. The RV $X_i = 1$, $i \in \{1, 2, \ldots, 8\}$, if Lazy chooses WW in the i^{th} hour. Else, $X_i = 0$. Answer the following questions.

(a) (5 marks) Derive the marginal PMFs of the RVs X_i, $i = 1, 2, \ldots, 8$. [*Hint*: Think counting/ combinatorics]

(b) (5 marks) Derive the joint PMFs of the RVs X_i and X_j, for $i \neq j$.

(c) (2.5 marks) Are the RVs $X_1, X_2, \ldots, X_8$ mutually independent? Justify your answer.

(d) (2.5 marks) Derive the expected value of the sum $X_1 + X_2 + \ldots + X_8$.

(e) (5 marks) Derive the variance of the sum $X_1 + X_2 + \ldots + X_8$.

(f) (5 marks) Suppose Lazy chooses to spend at least one hour on SM before spending the first hour on WW. Calculated the conditional expected value of X_3.

(a)

$P[X_i = 1] = P[$Lazy chooses WW in the ith hour and the other hours have a total of 4 WW and 3 SM$]$

No. of ways in which lazy can assign WW or SM to the hours with the ith hour fixed to WW is $^7C_3$.

Total no. of ways of assigning WW or SM to the hours is $^8C_3$.

$$\therefore P[X_i = 1] = \frac{^7C_3}{^8C_3} = \left(\frac{\lfloor 7}{\lfloor 3 \lfloor 4}\right)\frac{\lfloor 3 \lfloor 5}{\lfloor 8} = \frac{5}{8}$$

2.5 + 2.5

Numerator Denominator.

Of course alternate methods may be used.

Valid arguments for why 5/8 is the prob are okay too.

Thus the PMF of X_i is

$$P[X_i = x] = \begin{cases} 3/8 & x = 0 \\ 5/8 & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that this is true for all i in $\{1, 2, \ldots, 8\}$ since the prob is independent of i.

(b)

Consider

$P[X_i = 1, X_j = 1]$

$= P[i$th & jth hour are spent on WW and the remaining 3 WW & 3 SM$]$

$$= \frac{^6C_3}{^8C_3} = \frac{\lfloor 6}{\lfloor 3 \lfloor 3} \frac{\lfloor 3 \lfloor 5}{\lfloor 8} = \frac{\lfloor 6}{\lfloor 3}\frac{\lfloor 5}{\lfloor 8} = \frac{(5 \times 4)\lfloor 3}{(3 \times 2)(8 \times 7)} = \left(\frac{5}{8}\right)\left(\frac{4}{7}\right)$$ — 1

At least 1 for any attempt

$$P[X_i = 0, X_j = 1] = \frac{^6C_2}{^8C_3} = \left(\frac{3}{8}\right)\left(\frac{5}{7}\right)$$ — 1

$$P[X_i = 1, X_j = 0] = \frac{^6C_3}{^8C_3} = \left(\frac{3}{8}\right)\left(\frac{5}{7}\right)$$ — 1

$$P[X_i = 0, X_j = 0] = \frac{^6C_1}{^8C_3} = \left(\frac{3}{8}\right)\left(\frac{2}{7}\right)$$ — 1

The PMF constitutes the above four probabilities. Of course for any other pair of values, the probability is 0.

(c) It is easy to show that the RV(s) are not independent. Pick any one of the joint probabilities in (b).

Say $P[X_i = 0, X_j = 0] = \left(\frac{3}{8}\right)\left(\frac{2}{7}\right)$ — 1 [If in line with what was done in (b)]

We know from (a) that $P[X_i = 0]\, P[X_j = 0] = \left(\frac{3}{8}\right)\left(\frac{3}{8}\right)$ — 1 [If in line with what was done in (a).]

Thus, the RV(s) are not independent — 0.5 [Final result]

(d) $E[X_1 + X_2 + \cdots + X_8]$

$= E[X_1] + E[X_2] + \cdots + E[X_8]$ → At least 1 for the expansion

$E[X_i] = (1)(5/8) + (0)(3/8) = 5/8$.

$\therefore E[X_1 + \cdots + X_8] = 5$.

2.5 if done right. Wrong prob calc doesn't impact marks for this part.

Using intuition & getting the correct answer is fine too. Note that the sum $X_1 + \cdots + X_8 = 5$ w.p. 1.

(e) $Var[X_1 + \cdots + X_8]$

$=$ Sum of elements of the covariance matrix.

$$= Var[X_1] + \cdots + Var[X_8] + Cov[X_1 X_2] + \cdots + Cov[X_1 X_8] + \cdots + Cov[X_8 X_1] + \cdots + Cov[X_8 X_7]$$

Upto 2

$$= n\, Var[X_1] + (n^2 - n)\, Cov[X_1 X_2] \quad [n = 8]$$

[Note that this simplification is because PMF of X_i is the same for all i & so are the joint PMF(s) of the pairs]

$$Var[X_1] = E[X_1^2] - (E[X_1])^2 = (1)^2\left(5/8\right) - \left(5/8\right)^2 = \frac{5}{8}\left(3/8\right).$$ Upto 1

$Cov[X_1 X_2] = E[X_1 X_2] - E[X_1]E[X_2]$

$E[X_1 X_2] = (1)(1)\, P[X_1 = 1, X_2 = 1] = (1)(1)(5/8)(4/7) = \frac{20}{56}$.

$E[X_1]\, E[X_2] = \left(\frac{5}{8}\right)(5/8) = \frac{25}{64}$.

Upto 1

$\therefore Var[X_1 + \cdots + X_8]$

$$= 8\,(5/8)(3/8) + 8(8-1)\left[\frac{20}{56} - \frac{25}{64}\right] = \frac{15}{8} + \left[20 - \frac{25 \times 7}{8}\right] = \frac{15}{8} + \left[\frac{160 - 175}{8}\right] = 0.$$

Upto 1

Using intuition & getting the correct answer is fine too. Since the sum $X_1 + \cdots + X_8 = 5$ always. $Var[C] = 0$.

(f) Let A be the event that: Lazy chooses to spend at least one hour on SM before spending an hour on WW.

$E[X_3 | A] = ?$

$E[X_3 | A] = (1)\, P[X_3 = 1 | A] + (0)\, P[X_3 = 0 | A]$

$= P[X_3 = 1 | A]$ — Upto 1

$\frac{P[X_3 = 1, A]}{P[A]} =$ — Upto 1

$P[X_3 = 1, X_1 = 0, X_2 = 0]$ → First WW is 3rd hour

$+ P[X_3 = 1, X_1 = 0, X_2 = 1]$ → First WW is 2nd hour and third is also WW.

over

$P[X_1 = 0, X_2 = 1]$ → First WW is 2nd hour

$+ P[X_1 = 0, X_2 = 0, X_3 = 1]$ → First WW is 3rd hour

$+ P[X_1 = 0, X_2 = 0, X_3 = 0]$ → First WW is 4th hour

Expansion of Numerator 1

Denom 1

Correctness 1

Other ways are likely.

$$= \frac{\left(\frac{3}{8}\right)\left(\frac{2}{7}\right)\left(\frac{5}{6}\right) + \left(\frac{3}{8}\right)\left(\frac{5}{7}\right)\left(\frac{4}{6}\right)}{\left[\left(\frac{3}{8}\right)\left(\frac{5}{7}\right) + \left(\frac{3}{8}\right)\left(\frac{2}{7}\right)\left(\frac{5}{6}\right) + \left(\frac{3}{8}\right)\left(\frac{2}{7}\right)\left(\frac{1}{6}\right)\right]}$$

Question 3. 20 **marks** A virus is known to mutate into variants over months. Any person is infected by the virus within 0 or more integer number of months of the virus having been first detected. The number of months is distributed as a Poisson random variable with expected value $\alpha = 1$ and PMF $\alpha^k e^{-\alpha}/k!$, $k = 0, 1, \ldots$. A person infected by the virus within m months, $m = 0, 1, \ldots$, takes an additional exponentially distributed months with rate $\lambda_m = m + 1$ to show symptoms of infection. The PDF of the exponential RV is $\lambda_m e^{-\lambda_m x}$, $x \geq 0$, and its expected value is $1/\lambda_m$. Derive the expected value of the months, since the virus is first detected, any person takes to show symptoms of infection by the virus.

Let X be the total months for a person to show symptoms

Let M be the no. of months within which any person is infected.

$$E[X] = \sum_{m=0}^{\infty} E[X \mid M=m]\, P[M=m]$$

Up to 5

$$E[X \mid M=m] = \underbrace{m}_{\text{This to infection}} + \underbrace{\frac{1}{\lambda_m}}_{\text{Average to showing symptoms once infected}} = m + \frac{1}{m+1}$$

The conditional Up to 10

Similar split in case people go the route of calculating the PDF of X.

$$E[X] = \sum_{m=0}^{\infty} \left(m + \frac{1}{m+1}\right) \frac{e^{-1}}{m!}$$

$$= \underbrace{\sum_{m=0}^{\infty} m \frac{e^{-1}}{m!}}_{\text{Expected value of Poisson }(\alpha=1)} + \sum_{m=0}^{\infty} \frac{e^{-1}}{(m+1)!}$$

Up to 5

$$= 1 + e^{-1}\left[1 + \frac{1}{2!} + \frac{1}{3!} + \cdots\right]$$

$$= 1 + e^{-1}\left[e - 1\right]$$

$$= 1 + \left(1 - \frac{1}{e}\right) = 2 - \frac{1}{e}.$$

Question 4. 10 **marks** You arrive at a bus stop knowing that the time you must wait for a bus is distributed as an exponential RV with PDF $f_X(x) = \lambda e^{-\lambda x}, x \geq 0$. The bus hasn't arrived for the first s seconds of your wait. Derive the conditional CDF of X. Repeat the above under the assumption that you must wait for a bus for a time that is uniformly distributed over $(0, 120)$. Note that $s \in (0, 120)$.

We want

$$P[X \leq x \mid X > s]$$
$$= 1 - P[X > x \mid X > s]$$

$$P[X > x \mid X > s] = \begin{cases} \dfrac{P[X > x]}{P[X > s]} & x > s \\ 1 & \text{otherwise} \end{cases}$$

Up to 2

When $f_X(x)$ is exponential,

$$P[X > x \mid X > s] = \begin{cases} \dfrac{e^{-\lambda x}}{e^{-\lambda s}} & x > s \\ 1 & \text{otherwise} \end{cases}$$

$$\therefore \; P[X \leq x \mid X > s] = \begin{cases} 1 - e^{-\lambda(x-s)} & x > s \\ 0 & \text{otherwise} \end{cases}$$

Up to 4

When $f_X(x)$ is uniform $(0, 120)$

$$P[X > x \mid X > s] = \begin{cases} \dfrac{120 - x}{120 - s} & x > s \\ 1 & \text{otherwise} \end{cases}$$

$$\Rightarrow P[X \leq x \mid X > s] = \begin{cases} 1 - \dfrac{(120 - x)}{120 - s} & x > s \\ 0 & \text{otherwise} \end{cases}$$

Up to 4

Question 5. 20 marks An external agency is visiting IIIT-Delhi to evaluate students at the university. A student's evaluation results in a score that is a sum of the marks obtained by the student in humanities and the marks obtained in sciences. The agency would like the average score, where the average is calculated over scores of all students in the institute. Given the large number of students, the agency decides to instead estimate the average. The agency is able to conduct the sciences exam for 50 students and the humanities exam for 25 students. They ask IIIT-Delhi to provide lists of randomly chosen students that will take the two exams.

Assume that the marks of any student i in the sciences exam is given by RV X_i and marks in the humanities exam is given by Y_i. Marks obtained by students in sciences are independent of other marks and identically distributed as the RV X. Marks obtained by students in humanities are independent of other marks and identically distributed as the RV Y. Propose an unbiased estimator for the average. You must show that your estimator is unbiased.

Assume that $\text{Var}[X] = \text{Var}[Y] = 25$. Apply the Chebyshev's inequality to derive the confidence interval $2c$ for a confidence of 95%. The Chebyshev's inequality states that for any RV Z, $P[|Z - E[Z]| \geq c] \leq \text{Var}[Z]/c^2$, for any $c > 0$.

We want to estimate $E[X+Y]$.
The unbiased estimator is

$$Z = M_{n_1}(X) + M_{n_2}(Y) \qquad \begin{cases} n_1 = 50 \\ n_2 = 25 \end{cases}$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_1} X_i + \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

This can be seen by taking its expectation.

$$E[M_{n_1}(X) + M_{n_2}(Y)]$$

$$= \frac{1}{n_1}\sum_{i=1}^{n_1} E[X] + \frac{1}{n_2}\sum_{i=1}^{n_2} E[Y]$$

$$= E[X] + E[Y].$$

5

5

$$P[|Z - E[Z]| \geq c] \leq \frac{\text{Var}[Z]}{c^2}$$

$$\text{Var}[Z] = \text{Var}[M_{n_1}(X) + M_{n_2}(Y)]$$

$$= \text{Var}[M_{n_1}(X)] + \text{Var}[M_{n_2}(Y)]$$

$$= \frac{1}{n_1}\text{Var}[X] + \frac{1}{n_2}\text{Var}(Y)$$

$$= \frac{25}{50} + \frac{25}{25}$$

5

We want 95% confidence.

That is choose c such that

$$\frac{\text{Var}[Z]}{c^2} \leq 0.05.$$

$$c^2 = \frac{\text{Var}[Z]\,(100)}{5} = \left(\frac{1}{2} + 1\right)(20)$$

$$= 30$$

satisfies the requirement.

$$c = \sqrt{30}.$$

5