Final #2

Mark the correct answer in each part of the following questions.

4. Two-thirds of the passengers in the Arad-Dimona line are Arad residents and the others are Dimona residents. Passengers in the line are sampled until two in a row in the sample are Arad residents. Let X be the total number of people in the sample, Y – the number of Dimona residents among the same people, $R = X - Y$ and $S = Y/R$. (For example, if the first three people in the sample are from Dimona, the next person from Arad, the next six people from Dimona and the next two from Arad, then $X = 12, Y = 9, R = 3, S = 3$.)

   (a) For any integer $n \geq 2$ we have $P(X = n) =$

   (i) $\frac{1}{3} \cdot \left(\frac{2}{3}\right)^n + 2 \cdot \left(-\frac{1}{3}\right)^n$.
   (ii) $\frac{4}{3} \cdot \left(\frac{2}{3}\right)^n - \frac{4}{3} \cdot \left(-\frac{1}{3}\right)^n$.
   (iii) $\frac{2}{3} \cdot \left(\frac{2}{3}\right)^n + \frac{4}{3} \cdot \left(-\frac{1}{3}\right)^n$.
   (iv) none of the above.

   (b) Consider the distributions of the random variables X, Y.

   (i) X is negative binomial, Y is binomial.
   (ii) X is negative binomial, Y is not binomial.
   (iii) X is not negative binomial, Y is binomial.
   (iv) X is not negative binomial, Y is not binomial.

   (c) Consider the distribution functions F_S and F_R of S and R, respectively.

   (i) F_R has at most finitely many discontinuity points in each finite interval, F_S has at most finitely many discontinuity points both in the interval $[0, 0.1]$ and in the interval $[5, 5.1]$.

(ii) F_R has at most finitely many discontinuity points in each finite interval, while F_S has at most finitely many discontinuity points in the interval $[0, 0.1]$ and infinitely many discontinuity points in the interval $[5, 5.1]$.

(iii) F_R has at most finitely many discontinuity points in each finite interval, while F_S has infinitely many discontinuity points in the interval $[0, 0.1]$ and at most finitely many discontinuity points in the interval $[5, 5.1]$.

(iv) None of the above.

(d) $E(X) =$

(i) $145/36$.

(ii) $149/36$.

(iii) $151/36$.

(iv) none of the above.

(e) $P(X = 2Y + 2) =$

(i) $4/7$.

(ii) $2/3$.

(iii) $4/5$.

(iv) none of the above.

5. A random variable X has a density function f, given by:

$$f(x) = \begin{cases} \frac{1}{x \ln 2}, & 1 \le x \le 2, \\ 0, & \text{otherwise.} \end{cases}$$

(a) $E(X)$ lies in the interval:

(i) $[1, 5/4]$.

(ii) $(5/4, 3/2]$.

(iii) $(3/2, 7/4]$.

(iv) $(7/4, 2]$.

(b) $V(X)$ lies in the interval:

(i) $[0, 0.05]$.

(ii) $(0.05, 0.1]$.

(iii) $(0.1, 0.15]$.

(iv) $(0.15, \infty)$.

(c) X_1, X_2 are independent random random variables, each with the density function f. Let $S = X_1 + X_2$. On the interval $[2, 3]$ (but not necessarily on the interval $[3, 4]$), the density function $f_S(s)$ is given by:

(i) $\frac{2\ln(s-1)}{s\ln^2 2}$.

(ii) $\frac{2\ln s}{s\ln^2 2}$.

(iii) $\frac{2\ln(s+1)}{s\ln^2 2}$.

(iv) none of the above.

(d) Let $X_1, X_2, \ldots$ be a sequence of independent random variables, all with the same density function f. For an integer $n \geq 2$, denote by P_n the probability that for exactly two indices i between 1 and n we have $X_i \in [4/3, 4/3 + 1/n]$. Then

(i) $P_n \underset{n\to\infty}{\longrightarrow} \frac{3}{32\ln 2} e^{-\frac{3}{4\ln 2}}$.

(ii) $P_n \underset{n\to\infty}{\longrightarrow} \frac{3}{32\ln^2} e^{-\frac{3}{4\ln 2}}$.

(iii) $P_n \underset{n\to\infty}{\longrightarrow} \frac{9}{32\ln^2 2} e^{-\frac{3}{4\ln 2}}$.

(iv) none of the above.

(e) Denote by N the least integer such that 100 of the variables $X_1, X_2, \ldots, X_N$ assume values between 3/2 and 2. Then $E(N) =$

(i) $\frac{100}{\ln 3 - \ln 2}$.

(ii) $\frac{100}{2 - \ln 3/\ln 2}$.

(iii) $\frac{100\ln 3}{\ln 2}$.

(iv) none of the above.

6. A die is rolled infinitely many times. Denote by $S_k, k \geq 1$, the sum of results in the first k rolls.

(a) Denote by $N_i, 1 \leq i \leq 6$, the number of times the die shows i out of the first 1000 rolls.

(i) $P\left(\max_{1\le i\le 6} N_i - \min_{1\le i\le 6} N_i = 1\right) = \frac{1000!}{166!^2 167!^4 \cdot 6^{1000}}$.
(ii) $P\left(\max_{1\le i\le 6} N_i - \min_{1\le i\le 6} N_i = 1\right) = \frac{6\cdot 1000!}{166!^2 167!^4 \cdot 6^{1000}}$.
(iii) $P\left(\max_{1\le i\le 6} N_i - \min_{1\le i\le 6} N_i = 1\right) = \frac{15\cdot 1000!}{166!^2 167!^4 \cdot 6^{1000}}$.
(iv) None of the above.

(b) For each positive integer L, denote $X_L = \min\{S_k : k \ge 1, S_k \ge L\}$. (For example, if the results of the first three rolls are $4, 1, 2$, then $X_1 = X_2 = X_3 = X_4 = 4, X_5 = 5, X_6 = X_7 = 7$.) Let $p_L = P(X_L = L)$.

(i) Both of the recursive formulas

$$p_L = \frac{p_{L-1} + p_{L-2} + \ldots + p_{L-6}}{6}, \qquad L \ge 7,$$

and

$$p_L = p_1 p_{L-1} + p_2 p_{L-2} + \ldots + p_6 p_{L-6}, \qquad L \ge 7$$

are correct.
(ii) The first formula above is correct, but the second is not.
(iii) The second formula above is correct, but the first is not.
(iv) Neither of the above formulas is correct.

(c) Let A be the event whereby in the first 1000 rolls each of the two results 3 and 5 occurred exactly 500 times (and no other result occurred even once). Then:

(i) $P\left(S_i \le 4i \forall 1 \le i \le 999 \mid A\right) = \frac{1}{1000}$.
(ii) $P\left(S_i \le 4i \forall 1 \le i \le 999 \mid A\right) = \frac{1}{501}$.
(iii) $P\left(S_i \le 4i \forall 1 \le i \le 999 \mid A\right) = \frac{1}{500}$.
(iv) None of the above.

(d)

(i) The two random variables X_{2000} and $2X_{1000}$ are independent and identically distributed.
(ii) The two random variables X_{2000} and $2X_{1000}$ are dependent and identically distributed.
(iii) The two random variables X_{2000} and $2X_{1000}$ are independent but not identically distributed.

(iv) The two random variables X_{2000} and $2X_{1000}$ are dependent and not identically distributed.

(e) $P(S_{42000} > 147700)$ lies in the interval:

(i) $[0, 1/4]$.

(ii) $(1/4, 1/2]$.

(iii) $(1/2, 3/4]$.

(iv) $(3/4, 1]$.

Solutions

4. (a) If the first person in the sample is from Dimona, then the distribution of the number of remaining people to be sampled is the same as the initial distribution of X. If that person is from Arad, then there is a probability of 2/3 that the next will also be from Arad, in which case the experiment is finished, and a probability of 1/3 that he will turn out to be from Dimona, in which case we revert to the initial situation. Thus:

$$p_n = \frac{1}{3}p_{n-1} + \frac{2}{3}\cdot\frac{1}{3}p_{n-2} = \frac{1}{3}p_{n-1} + \frac{2}{9}p_{n-2}, \qquad n \geq 3.$$

The initial values are easily seen to be:

$$p_1 = 0, \qquad p_2 = \frac{2}{3}\cdot\frac{2}{3} = \frac{4}{9}.$$

The characteristic polynomial of the recurrence is $x^2 - \frac{1}{3}x - \frac{2}{9}$, with roots $\lambda_1 = 2/3, \lambda_2 = -1/3$. Thus,

$$p_n = a\left(\frac{2}{3}\right)^n + b\left(-\frac{1}{3}\right)^n, \qquad n = 1, 2, \ldots$$

for suitable constants a, b. Plugging in the initial conditions, we find that $a = 2/3, b = 4/3$.

Thus, (iii) is true.

(b) Suppose X is negative binomial, say $X \sim \overline{B}(r,p)$ for some r and p. Since X assumes values from 2 and above, we must have $r = 2$. Since $P(X = 2) = p_2 = 4/9$, we must have $p = 2/3$. If this was indeed th case, then we would have

$$P(X = 3) = \binom{3-1}{2-1}\left(\frac{2}{3}\right)^2\left(1-\frac{2}{3}\right)^{3-2} = \frac{8}{27},$$

whereas in fact we have

$$P(X = 3) = p_3 = \frac{1}{3}p_2 + \frac{2}{9}p_1 = \frac{4}{27}.$$

The variable Y assumes all integral non-negative values, and hence cannot possibly be binomial.

Thus, (iv) is true.

(c) R is the number of people from Arad in the sample. Hence the possible values for R are all integers from 2 and above. Thus the discontinuity points of F_R are the points $2, 3, \ldots$, so that each finite interval contains at most finitely many such points.

To characterize the possible values of S, we note that after each resident of Arad in the sample, except for the last two, we have at least one person from Dimona. Hence $Y \geq R - 2$. In other words, the possible values for S are all rationals of the form a/b, where $b \geq 2$ and $a \geq b - 2$. It follows that all rationals from 1 and above are possible values of S, as are all rationals of the forms $1 - 1/b$ and $1 - 2/b$, where $b \geq 2$. In particular, for every $\varepsilon > 0$, there are only finitely many values the variable S may assume which are below $1 - \varepsilon$. In conclusion, F_S has finitely many discontinuity points in the interval $[0, 0.1]$ and infinitely many in the interval $[5, 5.1]$.

Thus, (ii) is true.

(d) We have:

$$\begin{aligned} E(X) &= \textstyle\sum_{n=2}^{\infty} np_n \\ &= \tfrac{2}{3}\textstyle\sum_{n=2}^{\infty} n\left(\tfrac{2}{3}\right)^n + \tfrac{4}{3}\sum_{n=2}^{\infty} n \cdot \left(-\tfrac{1}{3}\right)^n \\ &= \tfrac{2}{3} \cdot \left(\tfrac{2/3}{(1-2/3)^2} - \tfrac{2}{3}\right) + \tfrac{4}{3} \cdot \left(\tfrac{-1/3}{(1+1/3)^2} - \tfrac{-1}{3}\right) = 15/4. \end{aligned}$$

Thus, (iv) is true.

(e) Obviously

$$\{X = 2Y + 2\} = \bigcup_{k=1}^{\infty} A_k,$$

where $A_k = \{X = 2k\} \cap \{Y = k-1\}$. Since the events A_k are pairwise disjoint,

$$\begin{aligned} P(X = 2Y + 2) &= \sum_{k=1}^{\infty} P(A_k) = \sum_{k=1}^{\infty} \left(\frac{1}{3}\right)^{k-1} \left(\frac{2}{3}\right)^{k+1} \\ &= 2 \sum_{k=1}^{\infty} \left(\frac{2}{9}\right)^{k} = \frac{4}{7}. \end{aligned}$$

Thus, (ii) is true.

5. (a) We have:

$$E(X) = \int_1^2 x \cdot \frac{1}{x \ln 2} dx = \int_1^2 \frac{dx}{\ln 2} = \frac{1}{\ln 2}.$$

Thus, (ii) is true.

(b) We first calculate $E(X^2)$:

$$E(X^2) = \int_1^2 x^2 \cdot \frac{1}{x \ln 2} dx = \int_1^2 \frac{x}{\ln 2} dx = \left[\frac{x^2}{2 \ln 2}\right]_1^2 = \frac{3}{2 \ln 2}.$$

It follows that:

$$V(X) = E(X^2) - E^2(X) = \frac{3}{2 \ln 2} - \frac{1}{\ln^2 2}.$$

Thus, (ii) is true.

(c) For $s \in [2, 3]$ we have:

$$f_S(s) = \int_1^{s-1} f(x) f(s-x) dx = \int_1^{s-1} \frac{1}{x \ln 2} \frac{1}{(s-x) \ln 2} dx.$$

Now

$$\frac{1}{x(s-x)} = \frac{1}{s}\left(\frac{1}{x} + \frac{1}{s-x}\right),$$

and therefore

$$f_S(s) = \frac{1}{s \ln^2 2} \left[\ln x - \ln(s-x)\right]_1^{s-1} = \frac{2\ln(s-1)}{s \ln^2 2}.$$

Thus, (i) is true.

(d) In the "interesting" interval $[1, 2]$, the distribution function of X is:

$$F(x) = \int_1^x \frac{1}{t \ln 2} dt = \frac{\ln x}{\ln 2}. \tag{1}$$

Hence for each i:

$$P(4/3 \le X_i \le 4/3+1/n) = \frac{\ln(4/3+1/n) - \ln(4/3)}{\ln 2} = \frac{\ln(1+3/4n)}{\ln 2}.$$

It follows that, denoting by M the number of indices i between 1 and n for which $X_i \in [4/3, 4/3 + 1/n]$, we have $M \sim B(n, \ln(1 + 3/4n)/\ln 2)$. Now recall that $\frac{\ln(1+x)}{x} \underset{x\to 0}{\longrightarrow} 1$. Consequently:

$$n \cdot \frac{\ln(1+3/4n)}{\ln 2} \underset{n\to\infty}{\longrightarrow} \frac{3}{4\ln 2}.$$

Therefore, as $n \to \infty$, the distribution of M converges to $P(3/4\ln 2)$, and in particular

$$P_n \underset{n\to\infty}{\longrightarrow} \frac{(3/4\ln 2)^2}{2!} e^{-3/4\ln 2} = \frac{9}{32\ln^2 2} e^{-3/4\ln 2}.$$

Thus, (iii) is true.

(e) Considering a result between 3/2 and 2 as a success, we may define N as the number of trials until we get 100 successes. By (1), the probability of success is

$$F(2) - F(3/2) = \frac{\ln 2 - \ln(3/2)}{\ln 2} = 2 - \frac{\ln 3}{\ln 2},$$

so that $N \sim \overline{B}\left(100, 2 - \frac{\ln 3}{\ln 2}\right)$. In particular:

$$E(N) = \frac{100}{2 - \ln 3/\ln 2}.$$

Thus, (ii) is true.

6. (a) One checks easily that the event $\max_{1\le i\le 6} N_i - \min_{1\le i\le 6} N_i = 1$ occurs if two of the numbers $1, 2, \ldots, 6$ show up 166 times each and the other four 167 times each. The number of sequences satisfying this condition is

$$\binom{6}{2}\binom{1000}{166, 166, 167, \ldots, 167} = \frac{15 \cdot 1000!}{166!^2 167!^4}$$

and the probability of each is $1/6^{1000}$. Hence:

$$P\left(\max_{1\le i\le 6} N_i - \min_{1\le i\le 6} N_i = 1\right) = \frac{15 \cdot 1000!}{166!^2 167!^4 \cdot 6^{1000}}.$$

Thus, (iii) is true.

(b) Note that p_L is the probability that the sequence of sums we get throughout the process includes the number L. Let $L \ge 7$. If the outcome of the first roll is k (where $1 \le k \le L$), then we get at some point to the sum L if and only if, counting from the second roll on, we get at some point to a sum of $L - k$. By the law of total probability:

$$p_L = \frac{1}{6} \cdot p_{L-1} + \ldots + \frac{1}{6} \cdot p_{L-6}.$$

However, the second formula is false. In fact, $X_1 = 1$ if and only if the result of the first roll is 1, $X_2 = 2$ if and only if the result of the first roll is 2 or the first two rolls are $1, 1$, $X_2 = 3$ if and only if the result of the first roll is 3, or the first two rolls are $1, 2$ or $2, 1$, or the first three rolls are $1, 1, 1$, and so on. Hence, $p_1 = 1/6$, $p_2 = 7/36$ and in general $p_{i+1} > p_i \ge 1/6$ for $1 \le i \le 6$. Hence the second formula cannot possibly be correct.
Thus, (ii) is true.

(c) Given A, the event whereby $S_i \le 4i$ for $1 \le i \le 999$ occurs if and only if at each stage, the number of times an outcome of 3 was obtained is no less than the number of times an outcome of 5 was obtained. Hence this amounts to the ballot problem, and thus

$$P\left(S_i \le 4i \ \ \forall \ \ 1 \le i \le 999 \mid A\right) = \frac{1}{501}.$$

Thus, (ii) is true.

(d) The variable $2X_{1000}$ assumes only even values, whereas X_{2000} may well assume odd values. Hence the two variables are not identically distributed.

The event $\{X_{2000} = 2000\}$ occurs if the outcome of all first 1999 rolls is 1, and therefore the event is of a positive probability. Similarly, $P(2X_{1000} = 500) > 0$. However, $P(X_{2000} = 2000, 2X_{1000} = 500) = 0$, so the variables are dependent.

Thus, (iv) is true.

(e) The outcome of a roll of a die is distributed $U[1, 6]$, with an expected value of $(1 + 6)/2 = 7/2$ and variance $\frac{(6-1+1)^2-1}{12} = \frac{35}{12}$. Hence:

$$\begin{aligned} P(S_{42000} > 147700) &= P\left(\frac{S_{42000}-42000\cdot 7/2}{\sqrt{42000\cdot 35/12}} > \frac{147700-42000\cdot 7/2}{\sqrt{42000\cdot 35/12}}\right) \\ &\approx P(Z > 2), \end{aligned}$$

where Z is a standard normal variable. Therefore:

$$P(S_{42000} > 147700) \approx 1 - \Phi(2) = 0.0228.$$

Thus, (i) is true.