Final #1

Mark all correct answers in each of the following questions.

1. We are given an initially empty urn. At each step we put in the earn two additional balls, and then draw at random one out of the balls in the urn. The balls put in the urn at step 1 are labelled by the numbers 1 and 2, the balls put at step 2 – by 3 and 4, and so forth. For $k \geq 1$, let T_k be the number of the step at which ball k is taken out of the urn. (For example, if at step 1 ball 2 is drawn, at step 2 ball 3 is drawn, and at step 3 ball 1 is drawn, then $T_1 = 3, T_2 = 1, T_3 = 2$.)

   (a) $P(T_1 = 1000) = \frac{1}{1001000}$.

   (b) $P(T_1 > 1000 \,|\, T_1 > 500) < \frac{1}{2}$.

   (c) $P(T_{100} > T_{200}) \geq \frac{1}{3}$.

   (d) The random variable $\sqrt{T_1}$ has an expectation, but not variance.

   (e) The random variables T_k and T_l are dependent for every $k, l \geq 1$.

2. From an urn, containing a white and b black balls, we draw without replacement $n_1 + n_2$ balls (where $n_1, n_2 \geq 1$ and $n_1 + n_2 \leq a + b$). Let X be the number of white balls among the first n_1 balls to be drawn and Y the number of white balls among the last n_2 balls.

   (a) For some values of the parameters a, b, n_1, n_2 we have $\frac{E(Y)}{E(X)} > \frac{n_2}{n_1}$, for some values $\frac{E(Y)}{E(X)} = \frac{n_2}{n_1}$, and for some values $\frac{E(Y)}{E(X)} < \frac{n_2}{n_1}$.

   (b) Y is hypergeometrically distributed.

   (c) $\rho(X, Y)$ increases both as a function of n_1 for fixed a, b, n_2, and as a function of n_2 for fixed a, b, n_1.

(d) If $n_2 > n_1$, then $Y - X \sim H(n_2 - n_1, a, b)$.

(e) For $a = 4, b = 3, n_1 = n_2 = 2$ we have $P(X = 1|Y = 1) = \frac{4}{5}$.

3. In a certain production line there are two machines, I and II. The time (in hours) between consecutive breakdowns of machine I is distributed Exp(2), and between consecutive breakdowns of machine II – Exp(3). The machines are independent.

   (a) When we first start the machines, the expected time until the first breakdown (of any machine) is distributed Exp(6).

   (b) Let t be the time since the last failure in one of the machines has been fixed. The expected time until the next failure (in that machine or the other) decreases as a function of t.

   (c) Let T_1 be the total time machine I operates until it fails 1000 times. Chebyshev's inequality implies:

   $$P(450 < T_1 < 550) \geq 0.9.$$

   (d) Let T_2 be the total time machine II operates until it fails 3600 times. Then:

   $$P(1180 < T_2 < 1240) \approx 0.82.$$

   (e) In machine I there is a self-stabilization system, which is activated right after the machine is fixed, and from then on every hour until the next failure of the machine. The number of times the system is activated between consecutive failures of the machine is geometrically distributed.

4. The variable (X, Y) is uniformly distributed in the region:

   $$S = \{(x, y) : 1 \leq x < \infty, 0 \leq y < 1/x^3\}.$$

   (Namely, as area$(S) = 1/2$, if $A \subseteq S$ with area$(A) = a$, then $P((X, Y) \in A) = 2a$.)

(a) The density function f_X of X is given by:

$$f_X(x) = \begin{cases} 2/x^3, & 1 \le x < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

(b) $E(X) = 2\sqrt{2}$.

(c) $V(Y) = \frac{11}{300}$.

(d) $\operatorname{Cov}(X, Y) < 0$.

(e) The correlation coefficient between X and Y is strictly between -1 and 0.

5. Reuven and Shimon hold a backgammon match. Reuven plays first in the odd-numbered games, and Shimon in the even-numbered games. The probability of Reuven to win a game is 0.8 when he plays first and 0.4 when Shimon plays first. Each win is worth 1 point.

(a) Suppose first that the match ends when one of the players first leads by 2 points over his opponent, with the leader at that point being the winner. Let p be the probability for Reuven to win the match. Then:

$$p = \sum_{n=1}^{\infty} (0.8^2)^n.$$

(b) Let Y be the number of games in the match. Then $E(Y) = \frac{50}{11}$.

In the following parts assume that the match goes on indefinitely. Denote by W_n the number of wins of Reuven in the first n games.

(c) $W_n \sim B(n, 0.6)$ for each even n.

(d) Chebyshev's inequality yields:

$$P(|W_{400} - 240| \ge 10) \le 0.4.$$

(e) The sequence $(W_n/n)_{n=1}^{\infty}$ converges in probability to 0.6.

6. (a) If X is a continuous random variable, then (even though the function sin is not one-to-one) so is $\sin X$.

   (b) Let X, Y be independent identically distributed random variables with a variance. Put $Z = Y - X$. When we subtract the variables, errors in each tend to cancel, and in particular $\sigma(Z) < \sigma(X)$.

   (c) Let X, Y, Z, W be random variables with variances. If all three correlation coefficients $\rho(X,Y), \rho(Y,Z), \rho(Z,W)$ are strictly positive, then $\rho(X,W) > -1$.

   (d) If $E(X) = 2$ and $E(X^4) = 16$, then X is a constant (with probability 1).

   (e) Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables with $X_n \sim N(n, n^2)$ for each n. Then the sequence satisfies the weak law of large numbers.

Solutions

1. (a) Note first that, for a positive integer n, for the event $\{T_1 \geq n\}$ to occur, it is required that ball 1 is drawn neither at step 1, nor at step 2, nor at step 3, and so forth up to step $n-1$. Therefore:

$$P(T_1 \geq n) = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdot \ldots \cdot \frac{n-1}{n} = \frac{1}{n}\,. \tag{1}$$

Consequently

$$\begin{aligned} P(T_1 = n) &= P(T_1 \geq n) - P(T_1 \geq n+1) \\ &= \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}\,, \end{aligned} \tag{2}$$

and in particular:

$$P(T_1 = 1000) = \frac{1}{1000 \cdot 1001} = \frac{1}{1001000}\,.$$

   (b) Employing (1), we obtain:

$$P(T_1 > 1000 \,|\, T_1 > 500) = \frac{P(T_1 >\geq 1001)}{P(T_1 >\geq 501)} = \frac{1/1001}{1/501} > \frac{1}{2}\,.$$

(c) For the event $\{T_{100} > T_{200}\}$ to occur, it is required that ball 100 will survive all drawings from step 50, when it is put in the urn, until (and including) step 99, right before ball 200 is put in the urn, and then that ball 200 will be drawn before ball 100. Due to symmetry, the probability for ball 100 to be drawn after ball 200, given that it was not drawn before ball 200 was put in the urn, is 1/2. It follows that:

$$P(T_{100} > T_{200}) = \frac{50}{51} \cdot \frac{51}{52} \cdot \ldots \cdot \frac{99}{100} \cdot \frac{1}{2} = \frac{1}{4}\,.$$

(d) By (2)

$$E\left(\sqrt{T_1}\right) = \sum_{n=1}^{\infty} \frac{1}{n(n+1)} \cdot \sqrt{n} < \sum_{n=1}^{\infty} \frac{1}{n^{3/2}} < \infty.$$

On the other hand,

$$E\left(\sqrt{T_1}^{\,2}\right) = E(T_1) = \sum_{n=1}^{\infty} \frac{1}{n(n+1)} \cdot n = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty.$$

Since the second moment of $\sqrt{T_1}$ does not exist, neither does this variable has a variance.

(e) T_k and T_l are certainly dependent if $k = l$. Now let $k \neq l$, say $k < l$. For $m > l/2$, both events $\{T_k = m\}$ and $\{T_l = m\}$ have positive probabilities, but their intersection, namely $\{T_k = T_l = m\}$, is of 0 probability. It follows that the two events above are dependent, and therefore so are the random variables T_k and T_l.

Thus, (a), (d) and (e) are true.

2. (b) The number of possibilities of choosing the batch of n_2 balls is $\binom{a+b}{n_2}$. The number of possibilities, in which the number of white balls within this batch is y, is $\binom{a}{y}\binom{b}{n_2-y}$. Thus

$$P(Y = y) = \frac{\binom{a}{y}\binom{b}{n_2-y}}{\binom{a+b}{n_2}},$$

so that $Y \sim H(n_2, a, b)$.

(a) Clearly, $X \sim H(n_1, a, b)$. According to the formula for the expectation of a hypergeometric random variable we have

$$E(X) = \frac{n_1 a}{a+b}, \qquad E(Y) = \frac{n_2 a}{a+b},$$

and consequently:

$$\frac{E(Y)}{E(X)} = \frac{n_2}{n_1}.$$

(c) The more white balls are drawn out of the first n_1, the less we expect to see out of the following n_2. Hence it is clear that $\rho(X, Y) < 0$. Now, as n_1 grows, the number of white balls among the first n_1 has a larger effect on the number of white balls among the following n_2. (For example, in the extreme case, where $n_1 = a+b-n_2$, we have $Y = a-X$, so that $\rho(X, Y) = -1$.) Hence we see intuitively that $\rho(X, Y)$ decreases as n_1 and n_2 increase. We shall now verify this claim computationally.

Let $X_i = 1$ if the i-th ball to be drawn is white and $X_i = 0$ otherwise, $1 \leq i \leq n_1$. Define random variables $Y_j, 1 \leq j \leq n_2$, analogously for the following balls. Thus:

$$X = \sum_{i=1}^{n_1} X_i, \qquad Y = \sum_{j=1}^{n_2} Y_j.$$

Hence:

$$E(XY) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} E(X_i Y_j) = n_1 n_2 \frac{a(a-1)}{(a+b)(a+b-1)}\,.$$

A routine calculation yields:

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{-n_1 n_2 ab}{(a+b)^2(a+b-1)}.$$

Using the formula for the variance of a hypergeometric random variable, we obtain:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = -1 \Bigg/ \sqrt{\left(\frac{a+b}{n_1} - 1\right)\left(\frac{a+b}{n_2} - 1\right)}\,.$$

Consequently, $\rho(X, Y)$ decreases as a function of n_1 and n_2.

(d) An $H(n_2 - n_1, a, b)$-distributed random variable may assume only the values 0 through $n_2 - n_1$. The variable $Y - X$ may assume all values between $-n_1$ and n_2. (Sometimes, it may assume values in a shorter interval only, but in many cases all this interval consists of possible values; take, for example, $a = b = 10, n_1 = n_2 = 3$.) Hence $Y - X$ is not $H(n_2 - n_1, a, b)$-distributed.

(e) We have

$$P(X = 1 \,|\, Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{\frac{\binom{4}{1,1,2}\binom{3}{1,1,1}}{\binom{7}{2,2,3}}}{\frac{\binom{4}{1}\binom{3}{1}}{\binom{7}{2}}} = \frac{3}{5}\,.$$

A simpler solution emerges if we notice the following. If Y is given, say $Y = y$, then X behaves again as a hypergeometrically distributed random variable, but where the number of white balls is now $a - y$ instead of a and the number of black balls is $b - (n_2 - y)$. In our particular case:

$$P(X = 1|Y = 1) = \frac{\binom{3}{1}\binom{2}{1}}{\binom{5}{2}} = \frac{3}{5}\,.$$

Thus, only (b) is true.

3. (a) Let T' be the time until machine I fails and T'' be the time until machine II fails. The time T until one of the machines breaks down is $\min(T', T'')$. We have:

$$\begin{aligned} P(T \geq t) &= P(\min(T', T'') \geq t) \\ &= P(T' \geq t, T'' \geq t) \\ &= P(T' \geq t)P(T'' \geq t) \\ &= e^{-2t}e^{-3t} = e^{-5t}. \end{aligned}$$

Hence the time in question is distributed Exp(5).

(b) Due to the memorylessness property of the exponential distribution, at any time when a machine works, the remaining time until it fails is distributed as the time from the point it last started operating. Hence, whenever both machines work, the time until the next failure is distributed as $\min(T', T'')$. In particular, the expected time until the next failure does not depend on t.

(c) T_1 is the sum of 1000 independent Exp(2)-distributed random variables. Hence $E(T_1) = 1000 \cdot 1/2 = 500$ and $V(T_1) = 1000 \cdot 1/2^2 = 250$. By Chebyshev's inequality:

$$\begin{aligned} P(450 < T_1 < 550) &= 1 - P(|T_1 - 500| \geq 50) \\ &\geq 1 - \frac{250}{50^2} = 0.9. \end{aligned}$$

(d) T_2 is the sum of 3600 independent Exp(3)-distributed random variables, say $T_2 = \sum_{i=1}^{3600} T_{2,i}$. Now $E(T_{2,i}) = 1/3$ and $V(T_{2,i}) = 1/9$ for each i. Normalizing T_2, namely taking the random variable

$$T' = \frac{T_2 - 3600 \cdot 1/3}{\sqrt{3600 \cdot 1/9}} \,,$$

we obtain an approximately standard normal random variable. Hence

$$P(1180 < T_2 < 1240) \approx P\left(\frac{1180 - 3600 \cdot \frac{1}{3}}{\sqrt{3600 \cdot 1/9}} < Z < \frac{1240 - 3600 \cdot \frac{1}{3}}{\sqrt{3600 \cdot 1/9}}\right),$$

where $Z \sim N(0,1)$. Thus:

$$\begin{aligned} P(1180 < T_2 < 1240) &\approx P(-1 < Z < 2) \\ &= \Phi(2) - \Phi(-1) = 0.82. \end{aligned}$$

(e) The self-stabilization system is activated between consecutive failures of machine I exactly n times if the time between these failures lies in the interval $[n-1, n)$. Hence the probability in question is

$$P(n - 1 \leq T_1 < n) = e^{-2(n-1)} - e^{-2n} = \left(e^{-2}\right)^{n-1} (1 - e^{-2}).$$

It follows that the number of times is $G(1 - e^{-2})$-distributed.

Thus, (c), (d) and (e) are true.

4. (a) Clearly, $F_X(x) = 0$ for $x < 1$, and

$$F_X(x) = 2\int_1^x \frac{1}{t^3}dt = 1 - \frac{1}{x^2}, \qquad 1 \le x < \infty.$$

By differentiation:

$$f_X(x) = \begin{cases} 2/x^3, & 1 \le x < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

(b) A routine calculation yields:

$$E(X) = \int_1^\infty x \cdot \frac{2}{x^3}dx = 2.$$

(c) To calculate $V(Y)$, we first find the distribution function, and then the density function, of Y. For $0 \le y \le 1$:

$$\begin{aligned} P(Y \ge y) &= 2\int_1^{1/\sqrt[3]{y}} \left(\frac{1}{x^3} - y\right) dx \\ &= \left[-x^{-2} - xy\right]_1^{1/\sqrt[3]{y}} \\ &= \left(-y^{2/3} + 1 - 2y^{2/3} + 2y\right) \\ &= 1 + 2y - 3y^{2/3}\,. \end{aligned}$$

Hence

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ 3y^{2/3} - 2y, & 0 \le y \le 1, \\ 1, & y > 1, \end{cases}$$

and

$$f_Y(y) = \begin{cases} 2y^{-1/3} - 2, & 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus:

$$\begin{aligned} E(Y) &= \int_0^1 y\left(2y^{-1/3} - 2\right) dy \\ &= \int_0^1 \left(2y^{2/3} - 2y\right) dy \\ &= \left[\frac{6}{5}y^{5/3} - y^2\right]_0^1 = \frac{1}{5}\,. \end{aligned}$$

Similarly, $E(Y^2) = 1/12$, and therefore $V(Y) = E(Y^2) - E(Y)^2 = 1/12 - 1/5^2 = 13/300$.

(d) Y assumes values between 0 and $1/X^3$. Hence, the larger the value X assumes is, the smaller is the value we expect Y to assume, so that intuitively we expect that $\mathrm{Cov}(X,Y) < 0$. Now let us calculate this covariance precisely.

Since $E(X)$ and $E(Y)$ have been calculated already, it remains just to calculate $E(XY)$. To this end, we shall calculate the distribution function of XY. Put $T = XY$. Obviously, T assumes only values in the interval $[0,1]$. If $X = x$, then $XY \leq 1/x^2$. Hence, for $0 \leq t \leq 1$, the product XY may assume a value above t only if X is at most $1/\sqrt{t}$. Thus:

$$\begin{aligned} P(T \geq t) =& 2\int_0^{1/\sqrt{t}} \left(\frac{1}{x^3} - \frac{t}{x}\right) dx \\ =& \left[-x^{-1} - 2t\ln x\right]_0^{1/\sqrt{t}} \\ =& 1 - t + \ln t. \end{aligned}$$

Hence

$$F_T(t) = \begin{cases} 0, & t < 0, \\ t - t\ln t, & 0 \leq t \leq 1, \\ 1, & t > 1, \end{cases}$$

and

$$f_T(t) = \begin{cases} -\ln t, & 0 \leq t \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Consequently:

$$\begin{aligned} E(T) &= \int_0^1 -t\ln t dt \\ &= \left[-\frac{t^2}{2}\ln t + \frac{t^2}{4}\right]_0^1 = \frac{1}{4}. \end{aligned}$$

Finally:

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{4} - 2 \cdot \frac{2}{5} = -\frac{3}{20}.$$

(e) To find $\rho(X,Y)$, we need first to calculate $V(X)$. Now

$$E\left(X^2\right) = \int_1^\infty x^2 \cdot \frac{1}{x^3} dx = \infty,$$

so that the variance of X, and thereby the required correlation coefficient, are undefined.

Thus, (a) and (d) are true.

5. (a) The match ends with a win for Reuven if, for some $n \geq 1$, the first $n-1$ pairs of games end with one win for each player, and then Reuven wins both games in the n-th pair. The probability for Reuven to win exactly one out of two consecutive games is $0.8 \cdot 0.6 + 0.2 \cdot 0.4 = 0.56$, and therefore Reuven's probability of winning the match is

$$p = \sum_{n=1}^{\infty} 0.56^{n-1} \cdot 0.8 \cdot 0.4 = \frac{1}{1-0.56} \cdot 0.32 = \frac{8}{11}\,.$$

(b) Given that after several pairs of games Reuven and Shimon have the same number of points, the probability for the game to end by the end of the following two games is $1 - 0.56 = 0.44$. Hence $Y = 2S$, where $S \sim G(0.44)$, so that

$$E(Y) = 2 \cdot \frac{1}{0.44} = \frac{50}{11}\,.$$

(c) Already for $n = 2$ we have $P(W_2 = 2) = 0.8 \cdot 0.4 = 0.32$, while the probability for a $B(2, 0.6)$-distributed random variable to assume the value 2 is $0.6^2 = 0.36$. Hence W_2 is not $B(2, 0.6)$-distributed.

(d) We may write

$$W_n = \sum_{i=1}^{n} M_i,$$

where the M_i's are independent, $M_i \sim B(1, 0.8)$ for odd i and $M_i \sim B(1, 0.4)$ for even i. This yields:

$$E(W_n) = \lceil n/2 \rceil \cdot 0.8 + \lfloor n/2 \rfloor \cdot 0.4 \qquad (3)$$

and

$$V(W_n) = \lceil n/2 \rceil \cdot 0.8 \cdot 0.2 + \lfloor n/2 \rfloor \cdot 0.4 \cdot 0.6. \qquad (4)$$

In particular:

$$E(W_{400}) = 240, \qquad V(W_{400}) = 80.$$

By Chebyshev's inequality:

$$P(|W_{400} - 240| \geq 10) \leq \frac{80}{10^2} = 0.8.$$

(e) According to (3) and (4), for each $n \geq 1$ we have

$$|E(W_n) - 0.6n| \leq 0.2$$

and

$$V(W_n) \leq 0.2n.$$

Employing Chebyshev's inequality, we obtain for any $\varepsilon > 0$:

$$\begin{aligned} P\left(\left|\frac{W_n}{n} - 0.6\right| \geq \varepsilon\right) &= P\left(|W_n - 0.6n| \geq \varepsilon n\right) \\ &\leq P\left(|W_n - E(W_n)| \geq \varepsilon n - 0.2\right) \\ &\leq \frac{0.2n}{(\varepsilon n - 0.2)^2} \underset{n \to \infty}{\longrightarrow} 0. \end{aligned}$$

Thus, (b) and (e) are true.

6. (a) For any number t (between 0 and 1), we may write $\{\sin X = t\} = \cup_{n=1}^{\infty} \{X = a_n\}$, where $\{a_n : n \in \mathbf{N}\}$ is the inverse image of the set $\{t\}$ under the function sin. Since X is continuous,

$$P(\sin X = t) = \sum_{n=1}^{\infty} P(X = a_n) = 0.$$

(b) Since $V(Z) = V(Y - X) = V(Y) + V(X) \geq V(X)$, we have $\sigma(Z) \geq \sigma(X)$.

(c) Viewing the covariance as an inner product on the space of random variables with variance, the correlation coefficient may be viewed as the cosine of the angle between random variables. To construct a counter-example, we need to construct random variables X, Y, Z, W, such that the "angle" between any two consecutive variables in the sequence is acute, but the one between X and W is π. Thus, take, for example, S and T as two uncorrelated random variables with $E(S) = E(T) = 0$ and $V(S) = V(T) = 1$, and let

$$X = S, \qquad Y = S + 2T, \qquad Z = -S + 2T, \qquad W = -S.$$

One readily verifies that

$$\rho(X, Y) = \frac{1}{\sqrt{5}}, \qquad \rho(Y, Z) = \frac{3}{5}, \qquad \rho(Z, W) = \frac{1}{\sqrt{5}},$$

while $\rho(X, W) = -1$.

(d) Recall that, if X is a random variable, then $E(X^2) \geq E(X)^2$, with equality if and only if X is a constant. Employing this twice in our case, we get

$$16 = E(X^4) \geq E(X^2)^2 \geq E(X)^4 = 16.$$

It follows that the two weak inequalities in the chain are in fact equalities, and consequently X is a constant.

(e) Clearly:

$$V(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{n^2} = \frac{(n+1)(2n+1)}{n}.$$

Since the sum of independent normal variables is normal as well, this implies that $\bar{X}_n - \bar{\mu}_n \sim N(0, (n+1)(2n+1)/n)$. Hence for any $\varepsilon > 0$:

$$\begin{aligned}
P\left(\left|\bar{X}_n - \bar{\mu}_n\right| \geq \varepsilon)\right) &= P\left(|Z| > \frac{\varepsilon}{\sqrt{(n+1)(2n+1)/n}}\right) \\
&= 1 - \Phi\left(\frac{\varepsilon}{\sqrt{(n+1)(2n+1)/n}}\right) \\
&\quad + \Phi\left(-\frac{\varepsilon}{\sqrt{(n+1)(2n+1)/n}}\right) \\
&\underset{n\to\infty}{\longrightarrow} 1 - \Phi(0) + \Phi(0) = 1
\end{aligned}$$

(where $Z \sim N(0,1)$). Since the left-hand side does not converge to 0 as $n \to \infty$, the sequence does not satisfy the weak law of large numbers.

Thus, (a) and (d) are true.