Probability Theory 2
Proposed solution of concluding assignment 2018

1. For every $1 \le i \le 72000$, let X_i be the indicator random variable for the event "the outcome of the ith die roll is 1". Note that $X_1, X_2, \ldots, X_{72000}$ are independent, and $\mathbb{E}(X_i) = Pr(X_i = 1) = 1/6$ for every $1 \le i \le 72000$. It is evident that $X = \sum_{i=1}^{72000} X_i$ and thus $\mathbb{E}(X) = \sum_{i=1}^{72000} \mathbb{E}(X_i) = 12000$ holds by the linearity of expectation.

   (a) We aim to apply the following version of Chernoff's inequality which was introduced in Lecture 1.

   Theorem 1 *Let $X_1, \ldots, X_n$ be independent random variables that return values in $[0,1]$, and let $X = \sum_{i=1}^{n} X_i$. Then, for every $t > 0$, it holds that*

   $$Pr(X \ge \mathbb{E}(X) + t) \le e^{-2t^2/n} \quad \textit{and} \quad Pr(X \le \mathbb{E}(X) - t) \le e^{-2t^2/n}.$$

   It follows that

   $$\begin{aligned} Pr(X < 10000 \text{ or } X > 14000) &= Pr(X < 10000) + Pr(X > 14000) \\ &\le Pr(X \le \mathbb{E}(X) - 2000) + Pr(X \ge \mathbb{E}(X) + 2000) \\ &\le 2e^{-2 \cdot 2000^2/72000} = 2e^{-1000/9} < 0.01. \end{aligned}$$

   We conclude that $Pr(10000 \le X \le 14000) = 1 - Pr(X < 10000 \text{ or } X > 14000) \ge 0.99$ as claimed.

   (b) Note that $Var(X_i) = 1/6 \cdot (1 - 1/6) = 5/36$ holds for every $1 \le i \le 72000$.

   $$\begin{aligned} Pr(11900 \le X \le 12100) &= Pr\left(\frac{11900 - 72000 \cdot 1/6}{\sqrt{5/36} \cdot \sqrt{72000}} < \frac{X - 72000 \cdot 1/6}{\sqrt{5/36} \cdot \sqrt{72000}} < \frac{12100 - 72000 \cdot 1/6}{\sqrt{5/36} \cdot \sqrt{72000}}\right) \\ &= Pr\left(\frac{-100}{\sqrt{10000}} < \frac{X - 12000}{\sqrt{10000}} < \frac{100}{\sqrt{10000}}\right) \approx \Phi(1) - \Phi(-1) \\ &= 2\Phi(1) - 1 \approx 0.6826 \le 0.7. \end{aligned}$$

2. We will present a randomized algorithm and then prove that it meets all the requirements of the question.

   Algorithm:

   (i) For every integer $1 \le i \le 3000$ choose an element x_i of A uniformly at random with replacement, all choices being mutually independent.

   (ii) Output the mean of $x_1, \ldots, x_{3000}$.

It is evident that the running time of the algorithm is a constant, that is, it does not depend on n. Indeed, we sample a constant number of elements from A and then calculate the mean of a set of size 3000. It remains to prove that with sufficiently high probability, the output of the algorithm is in the middle third of A.

Let $S = \{x \in A : |\{y \in A : y > x\}| \geq 2n/3\}$, let $L = \{x \in A : |\{y \in A : y < x\}| \geq 2n/3\}$ and let $M = A \setminus (S \cup L)$ (that is, the set S consists of the smallest $\lfloor n/3 \rfloor$ elements of A, the set L consists of the largest $\lfloor n/3 \rfloor$ elements of A, and the set M consists of the remaining "middle" elements of A). For every $1 \leq i \leq 3000$, let Z_S^i (respectively, Z_L^i) be the indicator random variable for the event "$x_i \in S$" (respectively, "$x_i \in L$"). Let $Z_S = \sum_{i=1}^{3000} Z_S^i$ and $Z_L = \sum_{i=1}^{3000} Z_L^i$, and observe that

$$\mathbb{E}(Z_S) = \sum_{i=1}^{3000} \mathbb{E}(Z_S^i) \leq 1000$$

and

$$\mathbb{E}(Z_L) = \sum_{i=1}^{3000} \mathbb{E}(Z_L^i) \leq 1000$$

hold by the linearity of expectation.

We are now in a position to bound from above the probability that the mean of $\{x_1, \ldots, x_{3000}\}$ is not in M. If this mean, denoted henceforth by z, is in L, then $Z_L \geq 1500$ holds by the definition of the mean. Applying Chernoff's inequality (as seen in Theorem 1 above) implies that

$$Pr(z \in L) \leq Pr(Z_L \geq 1500) \leq Pr(Z_L \geq \mathbb{E}(Z_L) + 500) \leq e^{-2 \cdot 500^2/3000} \leq 2^{-101}.$$

An analogous argument shows that

$$Pr(z \in S) \leq Pr(Z_S \geq 1500) \leq Pr(Z_S \geq \mathbb{E}(Z_S) + 500) \leq e^{-2 \cdot 500^2/3000} \leq 2^{-101}.$$

We conclude that $Pr(z \in M) \geq 1 - 2^{-100}$ as required.

3. Fix an arbitrary integer $1 \leq t \leq n/10$ and an arbitrary set $A \subseteq V(G)$ of size t. Then

$$Pr(|E(G[A])| \geq 3t) \leq \binom{\binom{t}{2}}{3t} p^{3t} \leq \left(\frac{et^2}{6tn}\right)^{3t} \leq \left(\frac{t}{2n}\right)^{3t},$$

where $G[A]$ is the subgraph of G with vertex set A and edge set $\{uv \in E(G) : u, v \in A\}$. Applying union bounds over all relevant values of t and all subsets of $V(G)$ of size t then implies that the probability that there exists a set $A \subseteq V(G)$ of size $1 \leq |A| \leq n/10$ for which $|E(G[A])| \geq 3t$ is at most

$$\sum_{t=1}^{n/10} \binom{n}{t} \left(\frac{t}{2n}\right)^{3t} \leq \sum_{t=1}^{n/10} \left(\frac{en}{t} \cdot \frac{t^3}{8n^3}\right)^{t} \leq \sum_{t=1}^{n/10} \left(\frac{t}{n}\right)^{2t} \leq \sum_{t=1}^{\sqrt{n}} \left(\frac{t}{n}\right)^{2t} + \sum_{t=\sqrt{n}}^{n/10} \left(\frac{t}{n}\right)^{2t}$$
$$\leq \sqrt{n} \cdot \left(\frac{\sqrt{n}}{n}\right)^{2} + n \cdot \left(\frac{1}{10}\right)^{2\sqrt{n}} = o(1).$$

4. **(a)** Note that

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{1/2} 4x dx + \int_{1/2}^{1} (4-4x)dx = 2x^2\Big|_0^{1/2} + (4x-2x^2)\Big|_{1/2}^{1}$$
$$= (1/2-0) + [(4-2)-(2-1/2)] = 1.$$

We conclude that f is indeed a density function.

(b) Starting with the cumulative distribution function, note first that

$$F_X(a) = \mathbb{P}(X \le a) = \int_{-\infty}^{a} f(x)dx = \int_{-\infty}^{a} 0 dx = 0$$

whenever $a < 0$ and

$$F_X(a) = \int_{-\infty}^{a} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$$

whenever $a > 1$. Assume next that $0 \le a \le 1/2$. Then

$$F_X(a) = \int_{-\infty}^{a} f(x)dx = \int_0^{a} 4x dx = 2x^2\Big|_0^{a} = 2a^2.$$

Finally, assume that $1/2 < a \le 1$. Then

$$F_X(a) = \int_{-\infty}^{a} f(x)dx = \int_0^{1/2} 4x dx + \int_{1/2}^{a} (4-4x)dx = 2x^2\Big|_0^{1/2} + (4x-2x^2)\Big|_{1/2}^{a}$$
$$= (1/2-0) + [(4a-2a^2)-(2-1/2)] = 4a-2a^2-1.$$

We conclude that

$$F_X(a) = \begin{cases} 0 & \text{if } a < 0 \\ 2a^2 & \text{if } 0 \le a \le 1/2 \\ 4a-2a^2-1. & \text{if } 1/2 < a \le 1 \\ 1 & \text{if } a > 1 \end{cases}$$

Next, we calculate the expectation of X. By definition

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)dx = \int_0^{1/2} 4x^2 dx + \int_{1/2}^{1} (4x-4x^2)dx = 4x^3/3\Big|_0^{1/2} + (2x^2-4x^3/3)\Big|_{1/2}^{1}$$
$$= (1/6-0) + [(2-4/3)-(1/2-1/6)] = 1/2.$$

In order to calculate $Var(X)$, we will first calculate $\mathbb{E}\left(X^2\right)$.

$$\mathbb{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{1/2} 4x^3 dx + \int_{1/2}^{1} (4x^2-4x^3)dx = x^4\Big|_0^{1/2} + (4x^3/3-x^4)\Big|_{1/2}^{1}$$
$$= (1/16-0) + [(4/3-1)-(1/6-1/16)] = 7/24.$$

We conclude that

$$Var(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2 = 7/24 - 1/4 = 1/24.$$