# Homework 2

Shixin Tian, Xiaochen Yang
Large Data Com S 535

October 25, 2017

# 1 Task 1

## 1.1 (a)

Let $x_i$ denote the outcome of the $i$'th toss. It's easy to see that $X = \sum_{0 \le i < n} x_i$.

$$\begin{aligned} E(X) &= E(\textstyle\sum_{0 \le i < n} x_i) \\ &= \textstyle\sum_{0 \le i < n} E(x_i) \\ &= \textstyle\sum_{0 \le i < j < n} (1 * \frac{1}{9} + 2 * \frac{1}{3} + 3 * \frac{1}{9} + 4 * \frac{1}{6} + 5 * \frac{1}{9} + 6 * \frac{1}{6}) \\ &= \frac{10n}{3} \end{aligned}$$

## 1.2 (b)

Let $y_i$ be 1 if the $i$'th outcome is even and 0 otherwise. We have $Y = \sum_{i=0}^{n} y_i$.

$$\begin{aligned} E(Y) &= E(\textstyle\sum_{0 \le in} y_i) \\ &= \textstyle\sum_{0 \le i < n} E(y_i) \\ &= \textstyle\sum_{0 \le i < j < k < n} (0 * \frac{1}{9} + 1 * \frac{1}{3} + 0 * \frac{1}{9} + 1 * \frac{1}{6} + 0 * \frac{1}{9} + 1 * \frac{1}{6}) \\ &= \frac{2n}{3} \end{aligned}$$

# 2 Task 2

## 2.1 (a)

Let $u_i$ be 1 if $i$ has only one inverse with respect to $f$ and 0 otherwise. It's easy to see that $|U| = \sum_{i=0}^{n} u_i$.

Firstly, we calculate $E(u_i)$. Let's assume that $u_i = 1$ and the only inverse of $i$ is $k$.

$$\begin{aligned} Pr(u_i = 1, f(k) = i) &= Pr(f(k) = i) * \textstyle\prod_{j \ne k} Pr(f(j)! = i) \\ &= \frac{1}{n}(\frac{n-1}{n})^{n-1} \end{aligned}$$

Then, we can calculate the $Pr(u_i = 1)$

$$\begin{aligned} Pr(u_i = 1) &= \sum_{k=0}^{n} Pr(u_i = 1, f(k) = i) \\ &= \frac{1}{n}(\frac{n-1}{n})^{n-1} * n \\ &= (\frac{n-1}{n})^{n-1} \end{aligned}$$

$$\begin{aligned} E(|U|) &= E(\sum_{i=0}^{n} u_i) \\ &= \sum_{i=0}^{n} E(u_i) \\ &= \frac{(n-1)^{n-1}}{n^{n-2}} \end{aligned}$$

### 2.2 (b)

Let $v_i$ be 1 if $f(i) = i$ and 0 otherwise. It's easy to see that $|V| = \sum_{i=0}^{n} v_i$.

$$\begin{aligned} E(|V|) &= E(\sum_{i=0}^{n} v_i) \\ &= \sum_{i=0}^{n} E(v_i) \\ &= \sum_{i=0}^{n} 1 * Pr(f(i) = i) \\ &= \sum_{i=0}^{n} \frac{1}{n} \\ &= 1 \end{aligned}$$

## 3 Task 3

Let $c_{x,y}$ be 1 if $< x, y > \in C_h$,i.e., $h(x) = h(y)$ and 0 otherwise. It's easy to see that $|C_h| = \sum_{x<y} c_{x,y}$. Easy to see $E(c_{x,y}) = \frac{1}{N^2}$ and $E(C_h) = \frac{N-1}{2N}$.

$$\begin{aligned} Var(c_{x,y}) &= E(c_{x,y}^2) - (E(c_{x,y}))^2 \\ &= (1^2 * Pr(c_{x,y} = 1) + 0^2 * Pr(c_{x,y} = 0)) - (E(c_{x,y}))^2 \\ &= \frac{1}{N^2} - \frac{1}{N^4} \\ &= \frac{N^2-1}{N^4} \end{aligned}$$

Since $c_{x,y}$ is independent with each other, we have $Var(C_h) = \sum_{x<y} Var(c_{x,y}) = \frac{N(N-1)}{2} Var(c_{x,y})$.

$$\begin{aligned} Pr[|C_h| \geq 1] &= Pr[|C_h| - E(|C_h|) \geq 1 - E(|C_h|)] \\ &\leq Pr[||C_h| - E(|C_h||) \geq 1 - E(|C_h|)] \\ &\leq \frac{Var(|C_h|)}{(1-E(|C_h|))^2} \\ &= \frac{2(N^2-1)(N-1)}{N(N+1)^2} \\ &\leq 2 \end{aligned}$$

Thus, we can get a trivial upper bound 2.

## 4 Task 4

Let $x_i$ denote the outcome of the n'th roll and $X = \sum x_i$. We are interested in calculating an upper bound of $Pr(X \leq 2n)$.

## 4.1 Markov's Inequality

Note that in Markov's inequality, it is calculating an upper bound of $Pr(X > v)$. Thus, we need to create a new variable to fit in the Markov's inequality.

Let $y_i = 7 - x_i$ and $Y = \sum_{y_i}$. Easy to see that $E(y_i) = 7 - E(x_i) = \frac{7}{2}$

$$\begin{aligned} Pr[X \leq 2n] &= Pr[\sum_{x_i} \leq 2n] \\ &= Pr[\sum_{7-y_i} \leq 2n] \\ &= Pr[\sum_{y_i} \geq 5n] \\ &= Pr[\sum_{y_i} > 5n - 1] \\ &= Pr[Y > 5n - 1] \\ &\leq \frac{E(Y)}{5n-1} \\ &= \frac{7n}{(5n-1)2} \end{aligned}$$

Thus, we can get an upper bound $\frac{7n}{(5n-1)2}$. This is approximated to $\frac{7}{10}$ when $n$ is large enough.

## 4.2 Chebyshev's Inequality

Firstly we calculate the variance of $X$.

$$\begin{aligned} Var(x_i) &= E(x_i^2) - (E(x_i))^2 \\ &= \frac{91}{6} - \frac{49}{4} \\ &= \frac{35}{12} \end{aligned}$$

Since each outcome is independent, $Var(X) = \sum Var(x_i)$

$$\begin{aligned} Pr[X \leq 2n] &= Pr[-X \geq -2n] \\ &= Pr[\frac{7n}{2} - X \geq \frac{7n}{2} - 2n] \\ &\leq Pr[|X - E(X)| \geq \frac{3n}{2}] \\ &\leq \frac{Var(X)}{(\frac{3n}{2})^2} \\ &= \frac{35n}{12} * (\frac{2}{3n})^2 \\ &= \frac{35}{27n} \end{aligned}$$

Thus, we got an upper bound $\frac{35}{27n}$.

## 4.3 Chernoff's Bound

Note that the variable in Chernoff's inequality should be between 0 and 1. We can make up a new variable $z_i = \frac{6-x_i}{5}$. Apparently, $z_i$ takes values between 0 and 1 and $E(z_i) = \frac{1}{2}$.

$$\begin{aligned} Pr[X \leq 2n] &= Pr[\sum x_i \leq 2n] \\ &= Pr[\sum 6 - 5z_i \leq 2n] \\ &= Pr[\sum z_i \geq \frac{4n}{5}] \\ &= Pr[\frac{Z}{n} \geq \frac{4}{5}] \\ &= Pr[\frac{Z}{n} - \frac{1}{2} \geq \frac{3}{10}] \\ &\leq Pr[|\frac{Z}{n} - \frac{1}{2}| \geq \frac{3}{10}] \\ &= Pr[|\frac{Z}{n} - \frac{1}{2}| \geq \frac{1}{2}\frac{3}{5}] \\ &\leq 2e^{-\frac{9n}{100}} \end{aligned}$$

Thus, we got an upper bound $2e^{-\frac{9n}{100}}$.

# 5 Task 5

$$\begin{aligned}
Pr[p - 0.1 \le X/2000 \le p + 0.1] &= Pr[|X/2000 - p| \le 0.1] \\
&= 1 - Pr[|X/2000 - p| \ge 0.1] \\
&= 1 - Pr[|X/2000 - p| \ge p * (0.1/p)] \\
&\ge 1 - 2e^{-(0.1/p)^2 * 2000p/2} \\
&= 1 - 2e^{-10/p} \\
&\ge 1 - 2e^{-10} \\
&\ge 1 - \frac{1}{1000}
\end{aligned}$$