Problem 1 (25 points)

(a) Consider the random process of rolling one die. Let $X$ be the number appearing on a fair die. Find $\mathbb{V} [X]$. Use Chebyshev’s inequality to bound $\mathbb{P} (|X - \frac{7}{2}| \geq \frac{1}{2})$.

(b) Suppose we roll a standard fair die 100 times independently. Let $Y = \sum_{i=1}^{100} X_i$, where $X_i$ is the number appearing on the die at the $i$th roll. Find $\mathbb{E} [Y]$ and $\mathbb{V} [Y]$. Use Chebyshev’s to bound $\mathbb{P} (|Y - 350| \geq 50)$.

(c) Suppose we roll a standard fair die 100 times independently. Let $Z = \frac{1}{100} \sum_{i=1}^{100} X_i$, where $X_i$ is the number appearing on the die at the $i$th roll. Use Chebyshev’s to bound $\mathbb{P} (|Z - \frac{7}{2}| \geq \frac{1}{2})$. 

Problem 2 (20 points)

Let $X$ be a discrete random variable chosen from the *geometric distribution*. That is,

$$
\mathbb{P}(X = k) = \begin{cases} 
(1 - p)^{k-1}p & \text{if } k \in 1, 2, 3, \ldots \\
0 & \text{otherwise}
\end{cases}
$$

Suppose $p = \frac{1}{3}$. That is, $X \sim \text{Geo} \left( \frac{1}{3} \right)$.

(a) Derive $\mathbb{E}[X]$ and $\mathbb{V}[X]$. Show your work.

(b) Use Markov’s inequality and Chebyshev’s inequality to bound $\mathbb{P}(X \geq 12)$.

(c) Compare your two results from part (b). For which values of $\alpha \geq 3$ does Markov’s inequality give a tighter bound on $\mathbb{P}(X \geq \alpha)$ than Chebyshev’s inequality?

(d) Akash just installed a new lightbulb. He knows that, in a given month, his lightbulb has a $\frac{1}{3}$ chance of fizzling out. Using your work from this problem, bound the probability that Akash’s lightbulb lasts at least a whole year.
Problem 3 (25 points)

We want to take blood tests for $N$ people. We can test them individually, but tests are expensive. Instead, we can pool the blood samples of $k$ people analyze them together. If this test is negative, this one test suffices for all $k$ people. If the test is positive, then each of the $k$ persons must be tested separately, and in all, $k + 1$ tests are required for the $k$ people. Assume that the probability $p$ that a test is positive is the same for all people and that these events are independent.

(a) Find the probability $p_{\text{pool}}$ that the test for a pooled sample of $k$ people will be positive.

(b) What is the expected value of the number of tests necessary under pooling? Assume that $N$ is divisible by $k$.

(c) For small $p$, using the union bound (see HW1 Problem 1.e.iii) as an approximation for $p_{\text{pool}}$, show that the value of $k$ which will minimize the expected number of tests with pooling is approximately $\frac{1}{\sqrt{p}}$. Since $p$ is small, you can assume

$$p_{\text{pool}} \approx \sum_{i=1}^{k} p = kp.$$  

(d) For fixed $k$ and $n$, give an inequality that tells us for which values of $p$ is pooling better than testing every individual. You can continue with the assumptions from part (c).
Problem 4 (30 points)

For constants \( c \) and \( n \), consider the discrete distribution function \( U_{c,n} \), which is uniform on \( n + 1 \) points in the range \( \{0, \frac{c}{n}, \frac{2c}{n}, \ldots, c\} \):

\[
\mathbb{P}(U_{c,n} = x) = \begin{cases} 
\frac{1}{n+1} & \text{if } x = \frac{ic}{n}, \ 0 \leq i \leq n \\
0 & \text{otherwise}
\end{cases}
\]

Let \( \bar{x} = (x_1, x_2, \ldots, x_m) \) be a sample of \( m \) values chosen independently from the distribution. We define:

<table>
<thead>
<tr>
<th>The “empirical expectation”</th>
<th>( \hat{\mathbb{E}}[X] = \frac{1}{m} \sum_{i=1}^{m} x_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>The “empirical second moment”</td>
<td>( \hat{\mathbb{E}}[X^2] = \frac{1}{m} \sum_{i=1}^{m} x_i^2 )</td>
</tr>
<tr>
<td>The “empirical variance”</td>
<td>( \hat{\mathbb{V}}[X] = \hat{\mathbb{E}}[X^2] - (\hat{\mathbb{E}}[X])^2 )</td>
</tr>
<tr>
<td>The “empirical standard deviation”</td>
<td>( \hat{\sigma}[X] = \sqrt{\hat{\mathbb{V}}[X]} )</td>
</tr>
</tbody>
</table>

(a) Repeat the following with \( c = 1 \) and \( c = 10 \), and sample sizes \( m = 4 \) and \( m = 1000 \):

Plot the 4 empirical statistics defined above as a function of \( n \) for \( n = 10, 20, \ldots, 100 \).

Note: You should have four plots, with four separate quantities graphed on each; (\( c = 1 \) and \( m = 4 \)), (\( c = 1 \) and \( m = 1000 \)), (\( c = 10 \) and \( m = 4 \)), etc.

(b) Compute analytically \( \mathbb{E}[X] \), \( \mathbb{E}[X^2] \), \( \mathbb{V}[X] \), and \( \sigma[X] \) for \( X \sim U_{c,n} \) as functions of \( n \).

Hint: You may use the series identity \( \sum_{i=0}^{k} i^2 = \frac{k(k+1)(2k+1)}{6} \).

(c) Compute \( \lim_{n \to \infty} \) of the above functions: \( \mathbb{E}[X] \), \( \mathbb{E}[X^2] \), \( \mathbb{V}[X] \), and \( \sigma[X] \).

(d) How do the analytical results compare to your plots?