COMP 6730 Advanced Database Systems, Homework #0

Problem 1: (30 points)

We have a function F: {0,...,n-1} → {0,...,m-1}. We know that, for $0 \leq x, y \leq n-1$, $F((x + y) \bmod n) = (F(x) + F(y)) \bmod m$. The only way we have for evaluating F is to use a lookup table that stores the values of F. Unfortunately, an Evil Adversary has changed the value of 1/5 of the table entries when we were not looking.

Describe a simple randomized algorithm that, given an input z, outputs a value that equals $F(z)$ with probability at least 1/2. Your algorithm should work for every value of z, regardless of what values the Adversary changed. Your algorithm should use as few lookups and as little computation as possible.

Suppose I allow you to repeat your initial algorithm three times. What should you do in this case, and what is the probability that your enhanced algorithm returns the correct answer?

Solution:
The algorithm is simply as follows:
(1) Pick a value x uniformly at random from {0,...,n-1}.
(2) Let y = (z − x) mod n.
(3) Return (F(x) + F(y)) mod m.
Now we analyze the probability that the algorithm returns the correct result. Apparently, in the table there are n/2 pairs of (x, y) that satisfy (x + y) mod n = z because for each x there is only one y in the table that satisfies the condition and swapping x and y results in the same table lookups (and should be counted only once). Among these n/2 pairs, the adversary can at most touch n/5 pairs. Thus,
Pr(correct) ≥ 1 − (n/5) / (n/2) = 3/5.
If we repeat the algorithm three times, we simply return the "majority vote". That is, if at least two runs return the same value, we return that value; otherwise we randomly pick one value and return it.
Now, Pr(correct) ≥ Pr(at least two runs are correct) ≥ $(3/5)^3 + (3/5)^2 \times (2/5) \times 3 = 81/125$.

Note that we cannot count the probability of the event that only one run is correct and our random choice picks it. This is because the two false runs might "collude" and return the same wrong result, fooling our algorithm to return that wrong result, with no chance of returning the correct result. We cannot bound the probability of this event. Thus, only 81/125 as computed above is what we can guarantee. Finally, note that 81/125 > 3/5.

Problem 2: (30 points)

We have a standard six-sided die. Let X be the number of times that a 6 occurs over n throws of the die. Let p be the probability of the event $X \geqslant n / 4$. Compare the best upper bounds on p that you can obtain using Markov's inequality, Chebyshev's inequality, and Chernoff bounds.

Solution: Define random variables $X_i = \begin{cases} 1, & \text{if } i\text{'th throw shows 6} \\ 0, & \text{otherwise} \end{cases}$ $(1 \le i \le n)$. Furthermore, define random variable $X = \sum_{i=1}^{n} X_i$. Then $E(X_i) = 1/6$, and from the linearity of expectation, $E(X) = n/6$. Thus, Markov inequality gives:

$$\Pr(X \ge n/4) < \frac{E(X)}{n/4} = \frac{n/6}{n/4} = \frac{2}{3}. \quad (1)$$

In order to use Chebyshev inequality, we further need the fact that $\mathrm{Var}(X_i) = 1/6 \times 5/6 = 5/36$ and hence $\mathrm{Var}(X) = (5/36)n$ due to the independence of the n throws. Thus, Chebyshev's inequality gives

$$\Pr(X \ge n/4) = \Pr(X - n/6 \ge n/12) \le \Pr(|X - n/6| \ge n/12) < \frac{Var(X)}{(n/12)^2} = \frac{5n/36}{n^2/144} = \frac{20}{n}. \quad (2)$$

Since X is the sum of 0/1 independent random variables, we can apply Chernoff bounds:

$$\Pr(X \ge n/4) = \Pr(X \ge (1+1/2)\frac{n}{6} \le \exp(-\frac{1}{3} \cdot \frac{n}{6} \cdot \frac{1}{4}) = \exp(-\frac{n}{72}). \quad (3)$$

Problem 3: (25 points)

A monkey types on a 26-letter keyboard that has lowercase letters only. Each letter is chosen independently and uniformly at random from the alphabet. If the monkey types 1,000,000 letters, what is the expected number of times that sequence "proof" appears?

Solution: The word "proof" possibly starts from position 1, position 2, …, and position 1,000,000 − 4 = 999,996. Define 999,996 random variables $X_i = \begin{cases} 1, & \text{if "proof" occurs from position } i \\ 0, & \text{otherwise} \end{cases}$ $(1 \le i \le 999,996)$. Clearly, $E(X_i) = \Pr(X_i = 1) = 1/26^5$. Define another random variable $X = \sum_{i=1}^{999,996} X_i$. Then X is the number of times that sequence "proof" appears in the monkey's whole writing experience. From the linearity of expectation, $E(X) = \sum_{i=1}^{999,996} E(X_i) = \frac{999,996}{26^5}$. Note that the linearity of expectation holds even though the X_i's are correlated.

Problem 4: (15 points)

Suppose that we roll a standard fair die 100 times. Let X be the sum of the numbers that appear over the 100 rolls. Use Chebyshev's inequality to bound $\Pr(|X - 350| \geqslant 50)$.

Solution: Let X_i be the number for each roll. $E(X_i) = 7/2$. $E(X_i^2) = 1/6*(1+4+9+16+25+36) = 91/6$. Then, $\mathrm{Var}(X_i) = E(X_i^2) - E(X_i)^2 = 35/12$. From the linearity of expectation, $E(X) = 100*7/2 = 350$. As X_i's are independent, $\mathrm{Var}(X) = 100 * \mathrm{Var}(X_i) = 3500/12$. Now, from Chebyshev's inequality,

$\Pr(|X - 350| \ge 50) < \mathrm{Var}(X) / 2500 = 7/60$.