Homework 3 - Solutions

Problem 1

Let $X_n$ be the random variable for the number of 6’s after $n$ dice rolls. Thus, we know that $X_n \sim \text{Binomial}(n, \frac{1}{6})$.

a) Markov: $P(X_n > \frac{n}{4}) < \frac{E[X_n]}{\frac{n}{4}} = \frac{2}{3}$

b) Chebyshev: $P(|X_n - E[X_n]| > \frac{n}{12}) < \frac{\text{Var}(X_n)}{(\frac{n}{12})^2} = \frac{n^2}{6} \cdot 12 = \frac{20}{n}$

c) Chernoff: $P(X_n > \frac{n}{4}) < \frac{E[e^{tX_n}]}{e^{t(\frac{n}{4})}} = \frac{\left(\frac{1}{6} e^t + \frac{5}{6}\right)^n}{e^{nt/4}} \leq \frac{e^{\frac{5}{6}(e^t-1)}}{e^{nt/4}}$

So choosing, $t^* = \ln \frac{3}{2}$ gives us a bound of $\approx e^{-0.018n}$. Using the Chernoff bound from the book is also acceptable.

Problem 2

Let $X = \frac{1}{N} \sum_{i=1}^{N} X_i$. So, we have:

$$P(|X - p| \geq \epsilon p) = P(|NX - NP| \geq \epsilon NP) = P(NX \geq (1 + \epsilon)NP) + P(NX \leq (1 - \epsilon)NP)$$

$$\leq e^{-NP\epsilon^2/3} + e^{-NP\epsilon^2/2}$$

The last inequality follows by the Chernoff bounds for the sum of IID Poisson trials (note that $NX$ is distributed as the sum of $N$ Poisson trials with probability of $p$ of outputting 1). Letting $N \geq 6 \ln \frac{1}{\delta} \frac{1}{5p\epsilon^2}$, we get that $P(|X - p| \geq \epsilon p) \leq \delta$.

So for $\epsilon = 0.1$, $\delta = 0.05$, and $0.2 < p < 0.8$, we need to query $\lceil N > 1797 \rceil$ people. In words, if we assume $0.2 < p < 0.8$, then taking an opinion poll of at least 1797 people gives us 95% confidence that our sample estimate for the actual fraction of people who want the president to be impeached is within 10% of the true fraction.

Problem 3

a) Let our estimate be $Y_t = \frac{1}{t} \sum_{i=1}^{t} X_i$. Take $t = \frac{\epsilon^2}{\epsilon^2 + \delta}$. Note that $E[Y_t] = E[\frac{\sum_{i=1}^{t} X_i}{t}] = \frac{tE[X]}{t} = E[X]$ and $\text{Var}(Y_t) = \text{Var}(\frac{\sum_{i=1}^{t} X_i}{t}) = \frac{1}{t^2} \sum_{i=1}^{t} \text{Var}(X_i) = \frac{\text{Var}(X)}{t}$

Thus, we get $P(|Y_t - E[X]| > \epsilon E[X]) = P(|Y_t - E[Y_t]| > \epsilon E[Y_t]) \leq \frac{\text{Var}(Y_t)}{\epsilon^2 E[Y_t]^2} = \frac{\text{Var}(X)\epsilon^2 \delta}{\epsilon^2 E[X]^2} = \delta$
b) Using part a) and letting $\delta = \frac{1}{4}$, we get that $\frac{4n^2}{e^2}$ samples are enough for a weak estimate.

c) Let $Y_i$ be the $i^{th}$ weak estimate and $Z_i = 1$ if $|Y_i - E[X]| \leq \epsilon E[X]$ and 0 otherwise, and let $Y$ be the median of $N$ weak estimates. Note that $E[Z_i] = P(Z_i = 1) \geq \frac{3}{4}$ by the definition of a weak estimate. Note that $|Y - E[X]| \geq \epsilon E[X]$ (which is the event that $Y$ is not within $\epsilon E[X]$ of $E[X]$) implies that $\sum_{i=1}^{N} Z_i \leq \frac{N}{2}$ (if the median if out of range, then at least half of the weak estimates used to produce that median must also be out of range).

Let us take $N = 24 \ln \frac{1}{\delta}$ weak estimates. By Chernoff bound for sum of IID Poisson trials

$$P(|Y - E[X]| \geq \epsilon E[X]) \leq P(\sum_{i=1}^{N} Z_i \leq \frac{N}{2}) = P(\sum_{i=1}^{N} Z_i \leq (1 - \frac{1}{3}) \frac{3N}{4}) \leq e^{-\frac{3N}{4} \frac{1}{2} \frac{1}{2} \frac{1}{2}} = e^{-\frac{N}{4}} \leq \delta$$

Thus, $O\left(\frac{3}{e^2} \ln \frac{1}{\delta}\right)$ samples are sufficient.

Problem 4

a) Consider a path from root to leaf. Let it contain $k$ good nodes. The number of elements left in the leaf is at most $n \left(\frac{2}{3}\right)^k$ which must be at least 1. Thus, we have $k \leq (\log_2 \frac{3}{2})^{-1} \log_2 n < 2 \log_2 n$. So, there can be no more than $\left\lceil \frac{2 \log_2 n}{k} \right\rceil$ good nodes on any path from root to leaf.

b) Consider a path from root to leaf and its first $c' \log_2 n$ nodes. Let $X_i$ be an indicator random variable for whether or not the node $i$ deep is a good node. Given the result of part a, we can also reason that once we have seen $c \log_2 n$ good nodes, the path must have reached a leaf. We know that $E[X_i] = \frac{1}{2}$. Setting $c = 2$ (from part a), and $c' = 24$, we get that

$$P\left(\sum_{i=1}^{c' \log_2 n} X_i \leq c \log_2 n\right) = P\left(\sum_{i=1}^{c' \log_2 n} X_i \leq (1 - \frac{3c}{c'}) (\frac{c' \log_2 n}{3})\right)$$

$$\leq e^{-\frac{c' \log_2 n (1 - \frac{3c}{c'})^2 \frac{1}{2}}{2}} = e^{-8 \log_2 n (\frac{1}{2})^2 \frac{1}{2}} \leq \frac{1}{n^2}$$

Thus, with probability greater than $1 - \frac{1}{n^2}$ there are at most $\left\lfloor \frac{24 \log_2 n}{k} \right\rfloor$ nodes on any path from root to leaf.

c) Let $E_i$ be the event that the $i^{th}$ path is longer than $24 \log_2 n$ nodes. Since there is a leaf for every entry of the original list, there are $n$ paths from root to leaf. We know by part b), $P(E_i) \leq \frac{1}{n^2}$. By the union bound, $P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i) \leq n \frac{1}{n^2} = \frac{1}{n}$. Thus, with probability greater than $1 - \frac{1}{n}$, there are no paths from root of leaf of length greater than $24 \log_2 n$.

d) The number of comparisons needed to place an element at the correct index is the length of the path from the root to the element’s leaf in this quicksort graph. There are $n$ leaves, and with probability greater than $1 - \frac{1}{n}$, all paths to these leaves have length at most $24 \log_2 n$, so the amount of comparisons is at most $n \cdot 24 \log_2 n \in O(n \log_2 n)$ with probability greater than $1 - \frac{1}{n}$.