1 Random hats

a)

Using the hint we define as $X_{i,j}$ the indicator random variable which depicts the event that the i-th person and the j-th person have exchange their hats. Generalizing this thought it is true that random variable X is the following:

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{i,j}$$

Given the fact that X_i, j is a indicator random variable and using the linearity of expected values, the expected value of X, $E[X]$ is the following:

$$\begin{aligned} E[X] &= E[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{i,j}] \implies \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[X_{i,j}] \implies \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Pr(X_{i,j}) \end{aligned}$$

Since each person get a uniformly random hat over n choice then the i-th person will get a hat (j-th's person hat) over n possible choices, while the j-th person receive a hat (i-th's person hat) over n-1 possible choices. Hence, the probability $Pr(X_{i,j}$ is the following:

$$Pr(X_{i,j}) = \frac{1}{n} \cdot \frac{1}{n-1}$$

Combining the above two equations, the expected value of X, $E[X]$ is:

$$\begin{aligned} E[X] &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Pr(X_{i,j}) \implies \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{n} \cdot \frac{1}{n-1} \implies \\ &= \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 1 \implies \\ &= \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^{n-1} (n-i) \implies \\ &= \frac{1}{n} \cdot \frac{1}{n-1} \frac{(n-1) \cdot n}{2} \implies \\ &= \frac{1}{2} \end{aligned}$$

b)

The definition of variance from the lecture is:

$$\begin{aligned}
Var(X) &= E[X^2] - E[X]^2 \implies \\
&= E[X^2] - \frac{1}{4} \implies \\
&= E[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{i,j}^2] - \frac{1}{4} \implies \\
&= E[(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{i,j}) \cdot (\sum_{y=1}^{n-1} \sum_{z=y+1}^{n} X_{y,z})] - \frac{1}{4} \implies \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{y=1}^{n-1} \sum_{z=y+1}^{n} E[X_{i,j} \cdot X_{y,z})] - \frac{1}{4}
\end{aligned}$$

Given the assumption that $X_{i,j}$ and $X_{y,z}$ are not independent then we have 2 use cases:

- $i = y$ and $j = z$: In this case indicator random variables $X_{i,j}$ and $X_{y,z}$ describes the exchange of hats between the same two people. So $X_{i,j} \cdot X_{y,z} = X_{i,j}^2$ and this is still an indicator random variable. Hence, as we used in sub question (a) we know that $E[X_{i,j}^2] = E[X_{i,j}] = \frac{1}{n \cdot (n-1)}$
- Each person i, j, y, z is a different one : In this case, we need at the same time $X_{i,j}$ and $X_{y,z}$ to be equal to one. This implies that uniformly the i-th person choose over n hats, the j-th person choose over n-1 hats, the y-th person choose over n-2 hats and the last one z-th person choose over n-3 hats. So the expected values is $E[X_{i,j}^2] = E[X_{i,j}] = \frac{1}{n \cdot (n-1) \cdot (n-2) \cdot (n-3)}$.

So the initial equation for variance will be computed as follows:

$$\begin{aligned}
Var(X) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{y=1}^{n-1} \sum_{z=y+1}^{n} E[X_{i,j} \cdot X_{y,z})] - \frac{1}{4} \implies \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[X_{i,j}^2] + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{y=1}^{n-1} \sum_{z=y+1}^{n} E[X_{i,j} \cdot X_{y,z})] - \frac{1}{4} \implies \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{n \cdot (n-1)} + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{y=1}^{n-1} \sum_{z=y+1}^{n} \frac{1}{n \cdot (n-1) \cdot (n-2) \cdot (n-3)} - \frac{1}{4} \implies \\
&= \frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 1 + \frac{1}{n \cdot (n-1) \cdot (n-2) \cdot (n-3)} \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{y=1}^{n-1} \sum_{z=y+1}^{n} 1 - \frac{1}{4} \implies \\
&= \frac{1}{2} + \frac{1}{4} - \frac{1}{4} \implies \\
&= \frac{1}{2}
\end{aligned}$$

2 Random counter, part 2

a)

Starting with the given inequality:

$$
\begin{aligned}
(1-\epsilon)\cdot m \le \tilde{m} \le (1+\epsilon)\cdot m &\implies \\
m - m\cdot\epsilon \le \tilde{m} \le m + m\cdot\epsilon &\implies \\
-m\cdot\epsilon \le \tilde{m} - m \le m\cdot\epsilon &\implies \\
|\tilde{m} - m| \le m\cdot\epsilon &
\end{aligned}
$$

Given that $\tilde{m} = 2^X - 1$ and combining our results from homework 3 we can easily imply that:

$$
\begin{aligned}
\tilde{m} &= 2^X - 1 \implies \\
E[\tilde{m}] &= E[2^X - 1] \implies \\
E[\tilde{m}] &= E[2^X] - 1 \implies \\
E[\tilde{m}] &= m + 1 - 1 \implies \\
E[\tilde{m}] &= m
\end{aligned}
\qquad\qquad
\begin{aligned}
\tilde{m}^2 &= (2^X - 1)^2 \implies \\
E[\tilde{m}^2] &= E[(2^X - 1)^2] \implies \\
E[\tilde{m}^2] &= E[2^{2X} + 1 - 2\cdot 2^X] \implies \\
E[\tilde{m}^2] &= E[2^{2X}] + 1 - 2\cdot E[2^X]] \implies \\
E[\tilde{m}^2] &= \frac{3m^2}{2} - \frac{m}{2}.
\end{aligned}
$$

Assuming that we need to calculate the probability that $\tilde{m}$ is NOT an approximation of m then from initial inequality the Chebyshev's inequality will be the following:

$$
\begin{aligned}
Pr(|\tilde{m} - m| > m\cdot\epsilon) &\le \frac{Var(\tilde{m})}{(m\cdot\epsilon)^2} \implies \\
Pr(|\tilde{m} - m| > m\cdot\epsilon) &\le \frac{E[\tilde{m}^2] - E[\tilde{m}]^2}{(m\cdot\epsilon)^2} \implies \\
Pr(|\tilde{m} - m| > m\cdot\epsilon) &\le \frac{\frac{3}{2}\cdot m^2 - \frac{m}{2} - m^2}{(m\cdot\epsilon)^2} \implies \\
Pr(|\tilde{m} - m| > m\cdot\epsilon) &\le \frac{1}{2}\cdot\frac{m(m-1)}{(m\cdot\epsilon)^2} \implies \\
Pr(|\tilde{m} - m| > m\cdot\epsilon) &\le \frac{1}{2}\cdot\frac{(m-1)}{m\cdot\epsilon^2} \implies \\
Pr(|\tilde{m} - m| > m\cdot\epsilon) &\le \frac{1}{2\cdot\epsilon^2}
\end{aligned}
$$

b)

Given $\tilde{m} = \frac{1}{t}\cdot\sum_{j=1}^{t} 2^{X_j} - 1$ then the expected value of $\tilde{m}$ is the following:

$$
\begin{aligned}
E[\tilde{m}] = E[\frac{1}{t}\cdot\sum_{j=1}^{t} 2^{X_j} - 1] &\implies \\
= \frac{1}{t}\cdot\sum_{j=1}^{t} E[2^{X_j}] - 1 &\implies
\end{aligned}
$$

$$= \frac{1}{t} \cdot \sum_{j=1}^{t} m + 1 - 1 \implies$$
$$= \frac{1}{t} \cdot \sum_{j=1}^{t} m + 1 - 1 \implies$$
$$E[\tilde{m}] = m$$

And the variance of of $\tilde{m}$ based on independence of random counters is the following:

$$Var(\tilde{m}) = Var(\frac{1}{t} \cdot \sum_{j=1}^{t} 2^{X_j} - 1) \implies$$
$$= \frac{1}{t^2} \cdot \sum_{j=1}^{t} Var(2^{X_j} - 1) \implies$$
$$= \frac{1}{t^2} \cdot \sum_{j=1}^{t} Var(2^{X_j}) \implies$$
$$= \frac{1}{t^2} \cdot \sum_{j=1}^{t} E[2^{2 \cdot X_j}] - E[2^{X_j}]^2 \implies$$
$$= \frac{1}{t^2} \cdot \sum_{j=1}^{t} \frac{1}{2} \cdot m(m-1) \implies$$
$$Var(\tilde{m}) = \frac{1}{2 \cdot t} \cdot m(m-1)$$

c)

Plugging the results from sub question (b) to Chebyshev's inequality of sub question (a):

$$Pr(|\tilde{m} - m| > m \cdot \epsilon) \leq \frac{Var(\tilde{m})}{(m \cdot \epsilon)^2} \implies$$
$$Pr(|\tilde{m} - m| > m \cdot \epsilon) \leq \frac{1}{2 \cdot t} \cdot \frac{(m-1)}{m \cdot \epsilon^2} \implies$$
$$Pr(|\tilde{m} - m| > m \cdot \epsilon) \leq \frac{1}{2 \cdot t \cdot \epsilon^2}$$

Ensuring that $\tilde{m}$ is an approximation of m with probability at least 99% is equal to ensuring that $\tilde{m}$ is NOT an approximation of m with probability at most 1%. So:

$$Pr(|\tilde{m} - m| > m \cdot \epsilon) \leq \frac{1}{2 \cdot t \cdot \epsilon^2} \leq 0.01 \implies$$
$$t \geq \frac{1}{2 \cdot 0.01 \cdot \epsilon^2} \implies$$
$$t \geq \frac{50}{\epsilon^2}$$

3 Generalization of Randomized Median Algorithm

a)

Since it is not needed to repeat the analysis we will just specify the changes in the required lines. First and foremost, we should note that the change in the initial randomized median algorithm is that we do not care about the median number but in general the k-th smallest element, so the random set R will be surrounded around not the median $\frac{n^{\frac{3}{4}}}{2}$ but around the k-th smallest element among n elements, so $\frac{k}{n} \cdot n^{\frac{3}{4}}$. Moreover, as we want in the initial algorithm the median to be between l_d and l_u now we want the k-th element to follow the same rule. Hence the changes will be the following:

- Line 2 : Let d be the $(\lfloor\frac{k}{n} \cdot n^{\frac{3}{4}}\rfloor - \sqrt{n})$th element in the sorted set R.
- Line 3 : Let u be the $(\lfloor\frac{k}{n} \cdot n^{\frac{3}{4}}\rfloor + \sqrt{n})$th element in the sorted set R.
- Line 6 : If $l_d > k$ or $l_u > n - k$ then FAIL.
- Line 8 :Output the $(k - l_d + 1)$th element in the sorted order of C.

b)

Following each line of the generalization of randomized median algorithm we will calculate the running time of the modified algorithm.

- Line 1: Sampling from S a random set R containing $n^{\frac{3}{4}}$ elements has a cost of $O(n^{\frac{3}{4}})$
- Line 2: Sorting the random set R demands $n^{\frac{3}{4}} \cdot log(n^{\frac{3}{4}})$ calculations so the cost is $O(n^{\frac{3}{4}} \cdot log(n^{\frac{3}{4}}))$
- Line 5 : By comparing every element in S to d and u, computing the set C and the numbers l_d and l_u in the worst case needs n comparisons, so the cost is $O(n)$.
- Line 7 : Sorting C set with $|C| \approx \frac{n}{log(n))}$ elements needs $\frac{n}{log(n))} \cdot log(\frac{n}{log(n)})$ calculations so the cost is $O(\frac{n}{log(n))} \cdot log(\frac{n}{log(n)}))$

Assuming that $O(n^{\frac{1}{4}}) > O(log(n))$ then the modified algorithm has a linear running time $O(n)$.

c)

The algorithm describes three 'bad' events E_1, E_2 and E_3 such that if they do not happen the algorithm does not fail. In our case, the difference is that we care about the k-th smallest element, which we define as s_k and not the median m. Moreover, we should point out that in both cases E_3 remains the same as the number of elements in C set is still the same. However, the definition of $E_{3.1}$ and $E_{3,2}$ has been changed as follows:

- $E_1 : Y_1 = |r \in R|r < s_k| < k \cdot n^{\frac{3}{4}} - \sqrt{n}$

- $E_2 : Y_1 = |r \in R | r > s_k| < \frac{n-k}{n} \cdot n^{\frac{3}{4}} - \sqrt{n}$
- $E_{3,1}$: At least $4 \cdot \frac{n-k}{n} \cdot n^{\frac{3}{4}}$ elements of C are greater than s_k
- $E_{3,1}$: At least $4 \cdot \frac{k}{n} \cdot n^{\frac{3}{4}}$ elements of C are smaller than s_k

d)

We define the random variable X_i as:

$$X_i = \begin{cases} 1 & \text{if the i-th sample is less than or equal to the} s_k \\ 0 & otherwise \end{cases}$$

The X_i are independent since the sampling is done with replacement. Because there are k elements that are smaller or equal to s_k, the propability that a randomly chosen element of S is less than or equal to the s_k can be written as:

$$Pr(X_i = 1) = \frac{k}{n}$$

The event is equivalent to: $Y_i = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$
Since Y_i is the sum of Bernoulli trials, it is true that $Y \sim Bin(n^{\frac{3}{4}}, \frac{k}{n})$.So:

$$E[Y] = n^{\frac{3}{4}} \cdot \frac{k}{n} = \frac{k}{n^{\frac{1}{4}}}$$

$$Var[Y] = n^{\frac{3}{4}} \cdot \frac{k}{n} \cdot \frac{n-k}{n} = \frac{k(n-k)}{n^{\frac{5}{4}}} < \frac{n^{\frac{3}{4}}}{4}$$

Applying Chebyshev's inequality then yields:
$Pr(E_1) = Pr[Y_1 < \frac{k}{n} \cdot n^{\frac{3}{4}} - \sqrt{n}] \leq Pr[|Y_1 - E[Y_1]| > \sqrt{n}] \leq \frac{Var(Y_i)}{n} \leq \frac{1}{4 \cdot n^{\frac{1}{4}}}$
The Z_i are independent since the sampling is done with replacement. Because there are at least $4 \cdot \frac{n-k}{n} \cdot n^{\frac{3}{4}}$ elements of C are greater than s_k then the order of u in the sorted order of S was at least $k + \frac{n-k}{n} \cdot n^{\frac{3}{4}}$ and thus the set R has at least $\frac{n-k}{n} \cdot \frac{3}{4} - \sqrt{n}$ samples among the largest $n - k + \frac{n-k}{n} \cdot n^{\frac{3}{4}}$ elements in S. Similarly, for $Pr(E_{3,1})$: We define the random variable Z_i as:

$$Z_i = \begin{cases} 1 & \text{if the i-th sample is among the} n - k + \frac{n-k}{n} \cdot n^{\frac{3}{4}} \text{largest elements in S} \\ 0 & otherwise \end{cases}$$

So we have $Pr(Z_1) = \frac{n-k+\frac{n-k}{n} \cdot n^{\frac{3}{4}}}{n} = \frac{n-k}{n} \cdot (1 - 4 \cdot \frac{4}{n^{\frac{1}{4}}})$ with $Z \sim Bin(n^{\frac{3}{4}}, \frac{n-k}{n} \cdot (1 - 4 \cdot \frac{4}{n^{\frac{1}{4}}}))$.So mean and variance are:

$$E[Z] = n^{\frac{3}{4}} \cdot \frac{n-k}{n} \cdot (1 - 4 \cdot \frac{4}{n^{\frac{1}{4}}}) = n^{\frac{3}{4}} \cdot \frac{n-k}{n} - 4\frac{n-k}{n} \cdot \sqrt{n}$$

$$Var[Z] = n^{\frac{3}{4}} \cdot \frac{n-k}{n} \cdot (1 - 4 \cdot \frac{4}{n^{\frac{1}{4}}}) \cdot (1 - \frac{n-k}{n} \cdot (1 - 4 \cdot \frac{4}{n^{\frac{1}{4}}}) < \frac{n^{\frac{3}{4}}}{4}$$

Applying Chebyshev's inequality then yields:
$Pr(E_{3,1}) = Pr[Z < \frac{n-k}{n} \cdot n^{\frac{3}{4}} - \sqrt{n}] \leq Pr[|Z - E[Z]| \geq \frac{3n-4k}{\sqrt{n}}] \leq \frac{Var(Z)}{\frac{3n-4k}{\sqrt{n}}} \leq \frac{n^{\frac{7}{4}}}{4 \cdot (3n-4k)^2)}$