1. (a) Prove that the item returned by \textsc{GetOneSample}(S) is chosen uniformly at random from \(S\).

\textbf{Solution:} Let \(P(i, n)\) denote the probability that \textsc{GetOneSample}(S) returns the \(i\)th item in a given stream \(S\) of length \(n\). We consider the behavior of the last iteration of the algorithm. With probability \(1/n\), \textsc{GetOneSample} returns the last item in the stream, and with probability \((n-1)/n\), \textsc{GetOneSample} returns the recursively computed sample of the first \(n-1\) elements of \(S\). Thus, we have the following recurrence for all positive integers \(i\) and \(n\):

\[
P(i, n) = \begin{cases} 
0 & \text{if } i > n \\
1/n & \text{if } i = n \\
\frac{n-1}{n}P(i, n-1) & \text{if } i < n
\end{cases}
\]

The recurrence includes the implicit base case \(P(i, 0) = 0\) for all \(i\). The induction hypothesis implies that \(P(i, n-1) = 1/(n-1)\) for all \(i < n\). It follows immediately that \(P(i, n) = 1/n\) for all \(i \leq n\), as required. ■

\textbf{Solution:} Fix an arbitrary positive integer \(i \leq n\). \textsc{GetOneSample}(S) returns the \(i\)th item in the stream if and only if it accepts the \(i\)th item and rejects the \(j\)th item for all \(j > i\). Thus, the probability of returning the \(i\)th item is exactly

\[\frac{1}{i} \prod_{j=i+1}^{n} \frac{j-1}{j}.
\]

Each integer between \(i+1\) and \(n-1\) appears once as a numerator and once as a denominator in the product, the product equals \(i/n\). We conclude that \textsc{GetOneSample}(S) returns the \(i\)th item with probability \(1/n\). ■

\textbf{Rubric:} These are not the only solutions.

(b) Describe and analyze an algorithm that returns a subset of \(k\) distinct items chosen uniformly at random from a data stream of length at least \(k\). The integer \(k\) is given as part of the input to your algorithm. Prove that your algorithm is correct.

\textbf{Solution (O(k) time per element):} We chain together \(k\) instances of \textsc{GetOneSample}, where the items rejected by each instance are treated as the input stream for the next instance. Folding these \(k\) instances together gives us the following algorithm, which runs in \(O(kn)\) time:

\textsc{GetSamples}(S, k):
\[
\ell \leftarrow 0 \\
\text{while } S \text{ is not done} \\
\quad x \leftarrow \text{next item in } S \\
\quad \text{for } j \leftarrow 1 \text{ to } \min(k, \ell) \\
\quad \quad \text{if } \text{RANDOM}(\ell - j + 1) = 1 \\
\quad \quad \quad \text{swap } x \leftrightarrow \text{sample}[j] \\
\text{return } \text{sample}[1..k]
\]

Correctness follows inductively from the correctness of \textsc{GetOneSample} as follows. The analysis in part (a) implies that \text{sample}[1] is equally likely to be any element of \(S\).
Let $S'[1..n-1]$ be the sequence of items rejected by $\text{sample}[1]$; at each iteration $\ell$, the item $S'[\ell-1]$ is either $S[\ell]$ (with probability $1 - 1/\ell$) or the previous value of $\text{sample}[1]$ (with probability $1/\ell$). The output array $\text{sample}[2..k-1]$ has the same distribution as the output array from $\text{GetSamples}(S',k-1)$. Thus, the inductive hypothesis implies that $\text{sample}[2..k]$ contain $k-1$ distinct items chosen uniformly at random from $S \setminus \{\text{sample}[1]\}$. We conclude that $\text{sample}[1..k]$ contain $k$ distinct items chosen uniformly at random from $S$, as required.

\textbf{Rubric:} This solution is worth at most 3 points, because it’s slower than necessary.

\textbf{Solution (O(1) time per element):} Recall the Fisher-Yates shuffle algorithm from Homework 3.

\begin{footnotesize}
\begin{verbatim}
FisherYates(S[1..n]):
    for \(\ell \leftarrow 1\) to \(n\)
        \(j \leftarrow \text{Random(}\ell\text{)}\)
        swap $S[j] \leftrightarrow S[\ell]$
    return $S$
\end{verbatim}
\end{footnotesize}

The Homework 3 solutions include a proof that $\text{FisherYates}$ produces each permutation of the input array with equal probability. In particular, each subset of $k$ elements of $S$ has the same probability of appearing in the prefix $S[1..k]$ of the output array.

Now we modify the Fisher-Yates algorithm in two different ways. First, let the input be given in a stream rather than an explicit array; in particular, we do not know the value of $n$ in advance. Second, we maintain only the first $k$ elements of the output array.

\begin{footnotesize}
\begin{verbatim}
FYSample(S,k):
    for \(\ell \leftarrow 1\) to \(k\)
        \(\text{sample}[\ell] \leftarrow \text{next item in } S\)
        swap $\text{sample}[\ell] \leftrightarrow \text{sample}[\text{Random}(\ell)]$
    for \(\ell \leftarrow k+1\) to \(\infty\)
        if $S$ is done
            return $\text{sample}[1..k]$
        \(j \leftarrow \text{Random}(\ell)\)
        if $j \leq k$
            \(\text{sample}[j] \leftarrow \text{next item in } S\)
\end{verbatim}
\end{footnotesize}

The $k$ elements produced by this modified algorithm have exactly the same distribution as the first $k$ elements of the output of Fisher-Yates. Thus, $\text{FYSample}$ chooses a $k$-element subset of the stream uniformly at random, as required. The algorithm runs in $O(n)$ time, using only $O(k)$ space.\footnote{Knuth’s \textit{The Art of Computer Programming} (Volume II) attributes this sampling algorithm to Alan Waterman some time before 1969.}

In fact, the algorithm remains correct if we remove the swap in the initial for-loop (in gray); this simplification changes the probabilities of the output permutation, but not of the output set. Thus, we can simplify the algorithm as follows, using $\text{sample}[k+1]$ as a trash can.
FYSAMPLE3(S, k):
    for ℓ ← 1 to ∞
        sample[min(k + 1, RANDOM(ℓ))] ← next item in S
        if S is done
            return sample[1..k]

Rubric: Yes, "See the Homework 3 solutions." is a proof. These are not the only correct solutions.
2. (a) Prove that $\Pr[\tilde{n} > (1 + \epsilon)n] \leq 1/(1 + \epsilon)$. \hfill [Hint: Markov’s inequality]

Solution: Let $x_1, x_2, \ldots, x_n$ denote the distinct items in $S$. Let $X$ denote the number of items $x \in S$ such that $h(x) \leq m/((1 + \epsilon)n)$.

$$\Pr[\tilde{n} \geq (1 + \epsilon)n] = \Pr[\frac{m}{\tilde{h}} \geq (1 + \epsilon)n] \quad \text{[definition of } \tilde{n}]$$

$$= \Pr[\tilde{h} \leq \frac{m}{(1 + \epsilon)n}] \quad \text{[algebra]}$$

$$= \Pr[h(x) \leq \frac{m}{(1 + \epsilon)n} \text{ for some } x \in S] \quad \text{[definition of } \tilde{h}]$$

$$= \Pr[X \geq 1] \quad \text{[definition of } X]$$

$$\leq \mathbb{E}[X] \quad \text{[Markov’s inequality]}$$

$$\leq \sum_{i=1}^{n} \Pr[h(x_i) \leq \frac{m}{(1 + \epsilon)n}] \quad \text{[linearity of expectation]}$$

$$= \sum_{i=1}^{n} \frac{1}{(1 + \epsilon)n} \quad \text{[uniformity of } h]$$

$$= \frac{1}{1 + \epsilon}$$

(We can also skip the two lines in red, using the union bound instead of Markov’s inequality.)

Rubric: This solution is considerably more detailed than necessary for full credit. A perfect solution must explicitly invoke the fact that $h$ is uniform (or more formally, that $h$ is drawn from a uniform family of hash functions).

(b) Prove that $\Pr[\tilde{n} < (1 - \epsilon)n] \leq 1 - \epsilon$. \hfill [Hint: Chebyshev’s inequality]

Solution: Let $Y$ denote the number of items $x$ such that $h(x) < m/((1 - \epsilon)n)$, and observe that $m/\tilde{h} < (1 - \epsilon)n$ if and only if $Y = 0$. Because the hash function $h$ is uniform, linearity of expectation implies $\mathbb{E}[Y] < 1/(1 - \epsilon)$. Because $h$ is pairwise independent, $Y$ is the sum of pairwise independent indicator variables. Thus, Chebyshev’s inequality implies $\Pr[Y \leq 0] < 1/\mathbb{E}[Y] = 1 - \epsilon$.

Rubric: A perfect solution must explicitly invoke the fact that $h$ is both uniform—since otherwise we can’t compute $\mathbb{E}[Y]$—and pairwise independent—since otherwise we can’t apply this form of Chebyshev’s inequality.

(c) Estimate the smallest value of $k$ (as a function of the accuracy parameter $\epsilon$) such that $\Pr[|\tilde{n}_k - n| > \epsilon n] \leq 1/4$.

Solution: Let $X$ denote the number of items $x$ such that $h(x) \leq km/((1 + \epsilon)n)$. Linearity of expectation and the uniformity of $h$ imply that $\mathbb{E}[X] = k/(1 + \epsilon)$. Because the hash function $h$ is 2-uniform, $X$ is the sum of pairwise independent indicator variables, so Chebyshev’s inequality implies

$$\Pr[\tilde{n}_k > n + \epsilon n] \leq \Pr[X \geq k] = \Pr[X \geq (1 + \epsilon)\mathbb{E}[X]] \leq \frac{1 + \epsilon}{\epsilon^2 k}.$$
Similarly, let $Y$ denote the number of items $x$ such that $h(x) \leq km/((1 - \varepsilon)n)$. Linearity of expectation and the uniformity of $h$ imply that $E[Y] = k/(1 - \varepsilon)$. Again, $Y$ is the sum of pairwise-independent indicator variables, so Chebyshev’s inequality implies

$$
Pr[\bar{n}_k < n - \varepsilon n] \leq Pr[Y < k] \leq Pr[Y \leq (1 - \varepsilon)E[Y]] \leq \frac{1 - \varepsilon}{\varepsilon^2 k}.
$$

The union bound now implies $Pr[|\bar{n}_k - n| > \varepsilon n] \leq 2/(\varepsilon^2 k)$. Setting $k = 8/\varepsilon^2 = O(1/\varepsilon^2)$ makes this upper bound less than $1/4$. ■

**Rubric:** A perfect solution must explicitly invoke the fact that $h$ is 2-uniform.

(d) Estimate the smallest value of $d$ (as a function of the confidence parameter $\delta$) such that $Pr[|\bar{N} - n| > \varepsilon n] \leq \delta$.

**Solution:** For each index $1 \leq i \leq d$, let $Z_i = [|\bar{n}_{k,i} - n| > \varepsilon n]$, and let $Z = \sum_{i=1}^{d} Z_i$.

If $|\bar{N} - n| > \varepsilon n$, then we must have $Z \geq d/2$—either at least half of the estimates $\bar{n}_{k,i}$ are too large, or at least half of the estimates $\bar{n}_{k,i}$ are too small. It follows that

$$
Pr[|\bar{N} - n| > \varepsilon n] \leq Pr[Z \geq d/2].
$$

The indicator variables $Z_i$ are mutually independent, and we have $Pr[Z_i = 1] \leq 1/4$ for all $i$ by part (c). Let $W_1, W_2, \ldots, W_d$ be mutually independent indicator variables where $Pr[W_i = 1] = 1/4$ for all $i$, and let $W = \sum_{i=1}^{d} W_i$. We immediately have

$$
Pr[Z \geq d/2] \leq Pr[W \geq d/2];
$$

intuitively, in any sequence of $d$ independent coin flips, if we increase the probability that each coin comes up heads, we also increase the probability of getting at least $d/2$ heads.

Finally, we apply Chernoff bounds with $\mu = E[W] = d/4$ and $\Delta = 1$: \footnote{Sorry, $\delta$ was already taken.}

$$
Pr[W \geq d/2] = Pr[W \geq (1 + \Delta)\mu] \leq \exp(-\Delta^2\mu/3) = \exp(-d/12).
$$

We conclude that if $d = 12\ln(1/\delta) = \Theta(\log(1/\delta))$, then $Pr[|\bar{N} - n| > \varepsilon n] < \delta$. ■

**Rubric:** Notice that this solution does not depend on the solution for part (c) at all. A perfect solution must explicitly invoke the fact that the estimates $\bar{n}_{i,k}$ are mutually independent. This is more detail than necessary for full credit; in particular, no penalty for implicitly assuming that $E[X] = d/4$. 

---

\footnote{Sorry, $\delta$ was already taken.}