1 Reservoir Sampling.

```
function   ReserviorSample (stream, k) {
    sample = [ ]     /* empty list for the output sample */
    i = 0            /* index used to go through the stream */
    while i < k :
        Sample [i] = Stream [i]
        i += 1
    while stream [i] is not empty :
        p = random (0, i)
        if p < k :
            sample [p] = stream [i]
        i += 1
    return sample
}
```

Prove for Correctness by Induction:

At the beginning, my algorithm copies first k elements from stream to the sample, which creates the basis for proof by induction.

Base case: the algorithm trivially works for $k=1$.

Induction assumption: for a stream with k elements, all elements are chosen with the same final probability $\frac{1}{k}$.

Inductive step: to show for a stream with $k+1$ elements, all elements have same probability of $\frac{1}{k+1}$ to be sampled.

From my second loop, the probability to choose the next element is $\frac{1}{k+1}$, and all other elements can stay with prob $\frac{1}{k}$ by assumption.

So, the current reservoir element stay with prob $1-\frac{1}{k+1} = \frac{k}{k+1}$.

Therefore, all previous elements have final prob of $\frac{1}{k} \cdot \frac{k}{k+1} = \frac{1}{k+1}$ to be the reservoir element after this iteration. Thus, all elements still have same prob of being selected as the reservoir element.

2 Median trick

Theorem: Let X be an unbiased estimator of a quantity Q. Let $\{X_{ij}\}_{i\in[t],j\in[k]}$ be a collection of independent RVs with X_{ij} distributed identically to X, where

$$t = O\left(\log\frac{1}{\delta}\right),\ k = O\left(\frac{Var[X]}{\epsilon^2 E[X]^2}\right)$$

Let $Z = \underline{median_{i\in[t]}\frac{1}{k}\sum_{j=1}^{k}X_{ij}}$. Then, $\mathbf{Pr}(|Z-Q|\geq\epsilon Q)\leq\delta$.

median of means for each row.

Proof sketch: Chebyshev and Chernoff. (Homework problem)

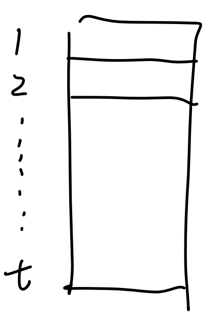
$$t = c_1\log\frac{1}{\delta},\quad k = C_2\frac{Var(X)}{\varepsilon^2 E(X)^2}$$

Chebyshev: $P(|Z-E(Z)|\geq k)\leq\frac{Var(Z)}{k^2}$.

Chernoff: $P(|Z-E(Z)|\geq\delta E(Z))\leq 2e^{-\frac{\delta^2}{3}E(Z)}$

for each row, $P(|\bar{X}-E(X)|>kQ)\leq\frac{Var(X)}{k\varepsilon^2 Q^2}$

Since $k = C_2\frac{Var(X)}{\varepsilon^2 E(X)^2}$, $P(|\bar{X}-E(X)|>kQ)\leq\frac{1}{C_2}$



$\leq\frac{1}{C_2}$ Let S represent the estimated median for all rows

$\leq\frac{1}{C_2}$ let $R_i = \begin{cases}1 & \text{if row } i \text{ has mean greater than median}\\ 0 & \text{o.w.}\end{cases}$

$\vdots$

$$S = \sum_{i=1}^{t}R_i$$

$$E(S) = np = t\cdot\frac{1}{C_2} = \frac{t}{C_2}$$

$$\begin{aligned}\Pr(|Z-Q|\geq\varepsilon Q) &= P\left(S\geq\frac{t}{2}\right)\\ &= P\left(|S-E(S)|\geq\frac{t}{2}-E(S)\right)\\ &= P\left(|S-E(S)|\geq\frac{t}{2}-\frac{t}{C_2}\right)\end{aligned}$$

$$\frac{t}{2}-\frac{t}{C_2} = \frac{(C_2-2)t}{2C_2} = \boxed{\frac{C_2-2}{2}}\cdot\frac{t}{C_2}$$

Then, use Chernoff bound: $P(|S-E(S)|\geq\underline{\delta E(S)})$

$$0 < \delta = \frac{c_2 - 2}{2} < 1$$

$$0 < c_2 - 2 < 2$$

$$2 < c_2 < 4$$, which force δ to be in bound $(0,1)$

Using Chernoff bound, $P(|S - E(S)| \geq \delta E(S)) \leq 2e^{-\frac{\delta^2}{3}E(S)}$

where $\delta = \frac{c_2 - 2}{2}$

$$2e^{-\frac{\delta^2}{3}E(S)} = 2e^{-\frac{\delta^2}{3}\cdot\frac{t}{c_2}} = 2e^{-\frac{(c_2-2)^2}{3\cdot 4}\cdot\frac{c_1}{c_2}\log(\frac{1}{\delta})}$$

$$= 2e^{\frac{c_1(c_2-2)^2}{12c_2}\log\delta}$$

$$= 2\delta^{\frac{c_1(c_2-2)^2}{12c_2}}$$

Let $c_2 = 3$, $= 2\delta^{\frac{c_1}{36}}$

Let $c_1 = 36$, $= 2\delta > \delta$

Therefore, by chernoff bound and chebyshev bounds used before

$$\Pr(|Z - \theta| \geq \varepsilon\theta) = P(|S - E(S)| \geq \delta E(S)) \leq 2e^{-\frac{\delta^2}{3}E(S)} = \delta$$

$$\Rightarrow \Pr(|Z - \theta| \geq \varepsilon\theta) \leq \delta$$

3. Variance of Morris Counter

Prove $Var[Z] = \frac{m(m-1)}{2}$

Properties of Morris algorithm

1. The expectation of the variable $Z=2^{X_m}$ satisfies the following:

$$E[Z]=m+1$$

Corollary: Morris algorithm outputs an unbiased estimator of m.

1. The variance of Z is equal to **$Var[Z]=m(m-1)/2$**

Observation: No improvement in terms of concentration as m grows since $Var(Z)/E(Z)^2$ is constant.

$$Var(Z) = E(Z^2) - (E(Z))^2$$
$$= E(2^{2X_m}) - (m+1)^2$$

To find $E(2^{2X_m})$, use definition of Expectation:

$$E(2^{2X_m}) = \sum_{i=1}^{\infty} 2^{2i} P(X_m = i)$$

By Morris algorithm,

$$P(X_m=i) = \frac{1}{2^{i-1}} P(X_{m-1}=i-1) + (1-\frac{1}{2^i}) P(X_{m-1}=i)$$

$$\therefore E(2^{2X_m}) = \sum_{i=1}^{\infty} 2^{i+1} P(X_{m-1}=i-1) + \sum_{i=1}^{\infty} 2^{2i} P(X_{m-1}=i) - \sum_{i=1}^{\infty} 2^{i} P(X_{m-1}=i)$$
$$= 2^2 E(2^{X_{m-1}}) + E(2^{2X_{m-1}}) - E(2^{X_{m-1}})$$
$$= 3E(2^{X_{m-1}}) + E(2^{2X_{m-1}})$$
$$= 3(m-1+1) + E(2^{2X_{m-1}})$$
$$= 3m + E(2^{2X_{m-1}})$$

$$\Rightarrow E(2^{2X_m}) = 3m + E(2^{2X_{m-1}})$$

for $m=0$, $E(2^{2X_0}) = 1$

$m=1$, $E(2^{2X_1}) = 3 + E(2^{2X_0})$

$m=2$, $E(2^{2X_2}) = 6 + 3 + E(2^{2X_0})$

$\vdots$

$$\Rightarrow E(2^{2X_m}) = 1 + \sum_{i=1}^{m} 3i = 1 + \frac{3}{2}m(m+1)$$

Therefore, plug in the value,

$$Var(Z) = E(2^{2X_m}) - (m+1)^2$$
$$= 1 + \frac{3}{2}m(m+1) - (m+1)^2$$

$$= \cancel{1} + \frac{3}{2}m^2 + \frac{3}{2}m - m^2 - 2m - \cancel{1}$$
$$= \frac{1}{2}m^2 - \frac{1}{2}m$$
$$= \frac{m(m-1)}{2}$$

4. Uniform RVs.

Let V_k be the k-th smallest hashed value

$$P(V_k \le x) = Pr(\text{at least } k \text{ samples in } [0, x])$$
$$= \sum_{l=k}^{n} \binom{n}{l} \cdot x^{l} (1-x)^{n-l} = \Delta$$

(a)
$$\frac{d\Delta}{dx} = \sum_{l=k}^{n} \binom{n}{l} \cdot \left(l \cdot x^{l-1} \cdot (1-x)^{n-l} - (n-l)(1-x)^{n-l-1} \cdot x^{l} \right)$$
$$= \sum_{l=k}^{n} \binom{n}{l} \cdot l \cdot x^{l-1} (1-x)^{n-l} - \sum_{l=k}^{n-1} \binom{n}{l} (n-l)(1-x)^{n-l-1} x^{l}$$
$$= \sum_{l=k}^{n} n \cdot \binom{n-1}{l-1} x^{l-1} (1-x)^{n-l} - \sum_{l=k}^{n-1} n \cdot \binom{n-l}{l} (1-x)^{n-l-1} \cdot x^{l}$$
$$= \underbrace{n \cdot \binom{n-1}{k-1} \cdot x^{k-1} \cdot (1-x)^{(n-1)-(k-1)}}_{\text{pdf of beta distribution.}}$$
$$= \frac{n!}{(k-1)!(n-k)!} \cdot x^{k-1} \cdot (1-x)^{(n-1)-(k-1)}$$

(b) Therefore, $E_{V_k \sim Beta(\alpha, \beta)}[V_k] = \frac{\alpha}{\alpha+\beta}$, where $\alpha = k$, $\beta = n-k+1$

$$E(V_k) = \frac{k}{k+n-k+1} = \frac{k}{n+1}$$