RESERVOIR SAMPLING

LARRY, LI

Claim 1. Given an input stream with n elements $\{a_1, a_2, \cdots, a_n\}$. Reservoir sampling can choose $k \leq n$ elements with each of equal probability $\frac{k}{n}$

Proof. If $k = 1$, we can show that up to loop index i, each element in $\{a_1, \cdots, a_i\}$ is chosen with probability $\frac{1}{i}$. This can be proved by mathematical induction.

Base case: if $i = 1$, since a_1 is chosen, this verifies a_1 is chosen with $\Pr(chosen) = \frac{1}{i} = 1$.

Suppose the above claim holds for $i = m$ case.

For $i = m+1$ case, for all elements' indices in $\{1, \cdots, m\}$ suppose $a_j, j \in \{1, \cdots, m\}$ is chosen. In the next round, a_j will survive with the probability $\frac{m}{m+1}$. Hence a_j will be chosen with probability $\frac{1}{m} \times \frac{m}{m+1} = \frac{1}{m+1}$. As for element a_{m+1} will be chosen with probability $\frac{1}{m+1}$.

By the induction hypothesis, claim 1 is true for $k = 1$.

Let's get back to general k case:

Base case: if $i = k$, $a_j, j \in \{1, \cdots, k\}$ each is selected with probability $\frac{k}{k} = 1$

Assume this holds true for $i = m$ case, i.e., each element is chosen with probability $\frac{k}{m}$

For $i = m + 1$ case, without loss of generality, suppose a_j is one among the chosen k elements in the m-th round, then a_j will remain being chosen with probability $\frac{k}{m} \times \left(1 - \frac{1}{m+1}\right) = \frac{k}{m+1}$ in the $(m + 1)$-th round. As for element a_{m+1}, it will survive in the $(m + 1)$-th round with probability $\frac{k}{m+1}$.

By the induction hypothesis, claim 1 is true. □

Date: April 26, 2012.