Reservoir Sampling

Algorithm Steps:

1. For first K elements, put them into the sample set S.
2. For the ith element, put it into S, with probability $x = \frac{K}{i}$, where i is N in this case. Then randomly replace an element in S with probability $\frac{1}{K}$

Proof:

When N <= K, each element has probability of 1.
When N = K+1, there are 2 cases:

When current element is selected with probability $\frac{K}{N}$, then each old element remains in the sample set S, with probability $\frac{K}{N}\left(1 - \frac{1}{K}\right)$.

When current element isn't selected with probability $(1 - \frac{K}{N}) = \frac{N-K}{N}$, then each old element remains in the sample set S, with probability $\frac{N-K}{N} * 1$.

Therefore, each old element remains in the sample set S, with probability
$\frac{K}{N}\left(1 - \frac{1}{K}\right) + \frac{N-K}{N} = \boldsymbol{\frac{N-1}{N}} = \boldsymbol{\frac{K}{N}}$

For N = k+i (for all i):

P(nth element is selected) =$\frac{K}{K+i} = \frac{K}{N}$
For any previous element X,
P(X still in the sample set S) = P(X was in S last time)*P(X isn't replace by nth element)

1. P(X was in S last time)= $\frac{K}{K+i-1} = \frac{K}{N-1}$
2. P(X isn't replace by nth element)=1-P(X is replaced by nth element)=1-$(\frac{K}{N} * \frac{1}{K}) = \frac{N-1}{N}$

P(X still in the sample set S)= $\frac{K}{N-1} * \frac{N-1}{N} = \frac{K}{N}$. Therefore, all elements have P=$\frac{K}{N}$ in S.