Reservoir Sampling Proof

We need to prove the in a stream of n sample where we keep k samples, each sample is kept with equal probability. That is,

$$P(S_i \text{ is kept}) = \frac{k}{n} \qquad \forall i \in [1, n]$$

There are two cases to consider:

1. When $i \leq k$, where the code initializes the list of kept samples.
2. when $i > k$, where the code randomly keeps incoming samples, throwing out old samples in the process.

1. let $i \leq k$:

$$P(S_i) = 1 \cdot \prod_{j=k+1}^{n} \frac{j-1}{j} = \frac{(n-1)!\,/\,(k-1)!}{n!\,/\,k!} = \frac{(n-1)!\,k!}{n!\,(k-1)!} = \frac{k}{n}$$

probability we keep it at first

probability we keep it on the j^{th} ~~sample~~ sample

2. let $i > k$:

$$P(S_i) = \frac{k}{i} \cdot \prod_{j=i+1}^{n} \frac{j-1}{j} = \frac{k}{i} \cdot \frac{(n-1)!/(i-1)!}{n!/i!} = \frac{k}{i} \cdot \frac{(n-1)!\,i!}{n!\,(i-1)!} = \frac{k}{i} \cdot \frac{i}{n} = \frac{k}{n}$$

probability we keep the sample as it first appears

probability we keep the sample i ~~[illegible]~~ when we see the j^{th} sample