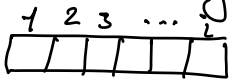
Reservoir Sampling: Unbiased Sampling from Streaming Data.

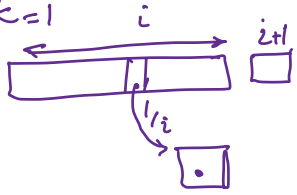
Given: Streaming data (at any time t, t instances have been observed)



Unbiased (uniform)

Objective: Maintain a Sample of Size k from the Stream

Suppose $k=1$




```
u = U_int[1, i+1]
if (u == 1)
    Sample ← A[i+1]
// otherwise, no change
```

Proof (by induction)

$P_{success}(A[j])$ = Prob. that $A[j]$ is the Sample

$\Rightarrow P_{success}(A[i+1]) = \frac{1}{i+1}$

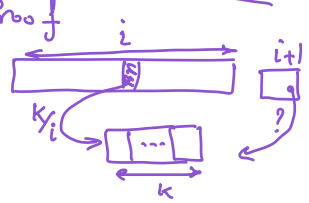
$\forall j \leq i \quad P_{success}(A[j])$ = Prob. that it was the Sample at iter i AND it Survived the Last Update

$$= \frac{1}{i} \times \left(1 - \frac{1}{i+1}\right) = \frac{1}{i} \times \frac{i}{i+1} = \frac{1}{i+1}$$

Extension to $K \geq 1$

```
for i = 1 to k
    S[i] = A[i]
for i > k:
    u = U_int[1, i]
    if (u ≤ k)
        S[u] = A[i]
```

Proof



$P_{success}(A[i+1]) = \frac{k}{i+1}$ ✓

$\forall j \leq i$

$P(A[j] \text{ replaced} \mid \text{in Sample}) = \frac{1}{i+1}$

$\Rightarrow P_{success}(A[j]) = \frac{k}{i} \times \left(1 - \frac{1}{i+1}\right)$

$= \frac{k}{i+1}$