Dictionary random Sampling

- want to sample m items
- Dictionary implemented via Hashing

1. Set $D = \emptyset$.
2. while ($|D| < m$) do
3. $p = rand(1, n)$
4. if $p \notin D$ => insert p in D
5. end while

- in D i put the indexes, sort,
=> I get the samples

if $m \leq \frac{n}{2}$

Cost
1) UNSUCCESSFUL → $p \in D$ already => $\mathbb{P}(p \in D) = \frac{|D|}{n} < \frac{m}{n} \leq \frac{\frac{n}{2}}{n} = \frac{1}{2}$
2) SUCCESSFUL $p \notin D$

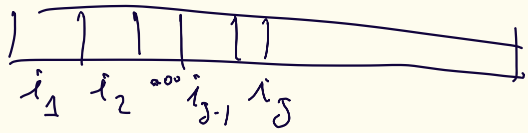
$\Rightarrow \frac{m}{n} \leq \frac{1}{2}$

Can say $m < \frac{1}{2} n$ cuz i turn down the problem assuming $m > \frac{1}{2} n$
=> In the D I put the position I don't wanna sample

Streaming model, n - known => Scanning & select

1. $s = 0$
2. for ($j = 1$; ($j \leq n$) && ($s < m$)) do
3. $p = rand(0, 1)$ // Real number
4. if $\left(p < \frac{m - s}{n - j + 1}\right)$ then
5. select S[j];
6. s++;
7. end if
8. end for

case $m=1$



$$\mathbb{P}(\text{pick } i_j) = \mathbb{P}(\text{pick } i_j) \cdot \mathbb{P}(\text{not pick other items})$$

$\forall j$

other items → $i_1, i_2, \dots i_{j-1}$

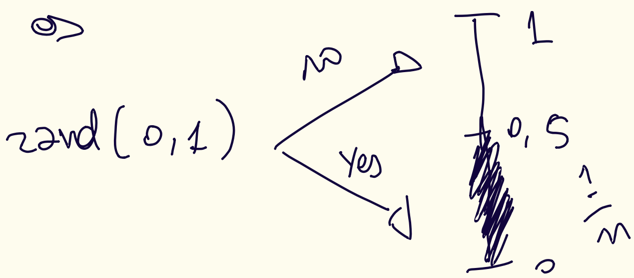
(NOT) (pick other items)

$$= \boxed{\frac{1}{n-j+1} \cdot \left(1 - \frac{j-1}{n}\right)} =$$

$$= \frac{1}{n-j+1} \cdot \frac{n-(j-1)}{n} = \frac{1}{\cancel{(n-j+1)}} \cdot \frac{\cancel{(n-j+1)}}{n} =$$

$$= \frac{1}{n} \rightarrow \text{probability pick } i_j$$

• generate the probability



case $m \geq 1$ (extracting m-items)

$$\mathbb{P}(\text{pick an item } i_j) = \frac{m-s}{n-j+1}$$

s → items already sampled

$n-j+1$ → how ~~many~~ much items are left to sample

current item

$$\frac{\text{how many picked}}{\text{how many left}} = \frac{m}{n} \quad \text{when } m=1 \Rightarrow P = \frac{1}{n}$$

at the beginning $i_1 \Rightarrow s=0;\ j=1 \Rightarrow \frac{m-0}{n-1+1} = \frac{m}{n}$

RESERVOIR SAMPLING : streaming model , n unknown

1. Initialize $R[1, m] = S[1, m]$ // for i in $\emptyset..m$ {
$R[i] = S[i]$ }

2. for each next item $S[j]$ do // $j > m$ $\Rightarrow$ for j in $m..$ S.len()

3. $h = \text{rand}(1, j)$

4. if $(h \leq m)$

5. set $R[h] = S[j]$

6. end if

8. end for

9. return array R.

es $m = 2$

S :

7	15	1	0	8	3	4
1	2	3	4	5	6	7

R :

7	15
1	2

if $h \leq \underset{2}{m}$ $\Rightarrow$ set $R[h] = S[j]$

$j=3$

for j in $3..7$

$h = \text{rand}(1, 3) = 1$ $1 \leq 2 \Rightarrow R[1] = S[3] \Rightarrow R =$

1	15
1	2

$j = 4$

$h = \text{rand}(1, 4) = 3$ $3 \leq 2$ no

$j = 5$

$h = \text{rand}(1, 5) = 2$ $2 \leq 2$ yes $R[2] = S[5]$ $\Rightarrow R =$

1	8
1	2

$j = 6$

$h = \text{rand}(1, 6) = 5$ $5 \leq 2$ no

$j = 7$

$h = \text{rand}(1, 7) = 1$ $1 \leq 2$ yes $R[1] = S[7]$

R =

4	8
1	2

$m \leq n$

1) $\boxed{\text{Passo BASE}}$ $\boxed{n = m}$ $\Rightarrow$ $\mathbb{P}$ should be $\boxed{\frac{m}{n}} = \frac{n}{n} = 1$

$\Rightarrow$ I'm picking all the items $\iff \mathbb{P} = 1$

2) $\boxed{\text{Passo induttivo}}$: $\boxed{n-1}$

- every item i_j with $j = 1, 2, \dots, n-1$ is picked with $\mathbb{P} = \frac{m}{n-1}$

$\Rightarrow$ PROOF THAT every item $j = 1, 2, \dots, n$ is picked with $\mathbb{P} = \frac{m}{n}$

AN ITEM

$\Rightarrow \mathbb{P}(i_n \text{ is picked}) = \frac{m}{n}$

$\mathbb{P}(i_j \text{ is picked}, \underline{j < n}) =$

$= \mathbb{P}(i_j \text{ is in } R \text{ at step } n-1) \overset{\text{AND}}{\wedge} (i_n \ \underline{\text{is not picked}} \overset{\text{OR}}{\vee}$

$i_n \text{ is picked} \overset{\text{AND}}{\wedge} i_j \text{ not kicked out})$

$= \mathbb{P}(i_j \in R \text{ at step } n-1) \cdot [\mathbb{P}(i_n \text{ is not picked}) + \mathbb{P}(i_n$ is picked $\wedge$ doesn't overwrite $i_j))$

$$\frac{m}{n-1} \cdot \left[\left(1 - \frac{m}{n}\right) + \frac{\cancel{m}}{n} \cdot \frac{m-1}{\cancel{m}}\right] =$$

$$= \frac{m}{n-1} \cdot \left(\frac{n - \cancel{m}}{n} + \frac{\cancel{m} - 1}{n}\right) = \frac{m}{\cancel{n-1}} \cdot \frac{\cancel{n-1}}{n} = \frac{m}{n}$$