Algorithm 6.5.6: Reservoir Sampling

```
# A random sample of a population 𝒫 into
# indexed array 𝒮 supporting a pre-determined
# sample size n = |𝒮|, and requiring only
# sequential access to 𝒫.
# |𝒫| not required.

𝒮 ← first n elements of 𝒫
j ← n
while( elements exist in 𝒫 ) do
  e ← next element from 𝒫
  if( Bernoulli(n/(j+1)) is True ) then (
    i ← Equilikely(0, n-1)
    𝒮[i] ← e
  )
  increment j
)

# 𝒮 is an unbiased random sample of
# n elements from 𝒫.
```

Algorithm 6.5.6: Reservoir Sampling

People who are miserable on Monday have lots of company. It's the worst day of the week for millions, according to researchers at the University of Vermont Complex Systems Center ***who analyze Twitter messages for happiness sentiment.***[1]

(TWITTER PROBABLY RUNS RS **INSIDE** IT'S SAMPLING APIS.)

Sampling a collection of simulation objects or events for statistical analysis.

Efficient sampling of an SQL query when Fisher-Yates (partial) shuffling[2] requires too much memory.

[1] https://qz.com/1097506/monday-is-the-worst-...
[2] Algorithm 6.5.2

```
# A random sample of a population 𝒫 into
# indexed array 𝒮 supporting a pre-determined
# sample size n = |𝒮|, and requiring only
# sequential access to 𝒫.
# |𝒫| not required.

𝒮 ← first n elements of 𝒫
j ← n
while( elements exist in 𝒫 ) do
  e ← next element from 𝒫
  if( Bernoulli(n/(j+1)) is True ) then (
    i ← Equilikely(0, n−1)
    𝒮[i] ← e
  )
  increment j
)

# 𝒮 is an unbiased random sample of
# n elements from 𝒫.
```

Algorithm 6.5.6: Reservoir Sampling

Note that the text increments j before the call to *Bernoulli*, so its invocation is

$$Bernoulli\left(\frac{n}{j}\right)$$

Whereas the algorithm (for which the proof is constructed) increments j **after** the call to *Bernoulli* and uses

$$Bernoulli\left(\frac{n}{j+1}\right)$$

These algorithms are equivalent, there are no smoke and mirrors in this lecture.

```
# A random sample of a population P into
# indexed array S supporting a pre-determined
# sample size n = |S|, and requiring only
# sequential access to P.
# |P| not required.

S ← first n elements of P
j ← n
while( elements exist in P ) do
  e ← next element from P
  if( Bernoulli(n/(j+1)) is True ) then (
    i ← Equilikely(0, n-1)
    S[i] ← e
  )
  increment j
)

# S is an unbiased random sample of
# n elements from P.
```

Algorithm 6.5.6: Reservoir Sampling

Proof by Induction

```
S ← first n elements of P
j ← n
while( elements exist in P ) do
  e ← next element from P
  if( Bernoulli(n/(j+1)) is True ) then (
    i ← Equilikely(0, n − 1)
    S[i] ← e
  )
  increment j
)
```

1. Show RS (reservoir sampling) algorithm is true for some fixed $|\mathcal{S}| = n = |\mathcal{P}| - 1$
2. Assume RS algorithm is true for some sample size $|\mathcal{S}| = n < j < |\mathcal{P}|$
3. Given (2), show the RS algorithm is true for sample size $|\mathcal{S}| = n < j + 1 \leq |\mathcal{P}|$

where

$\mathcal{S}$	Sample of population generated by algorithm
$n = \|\mathcal{S}\|$	Sample size desired (**guaranteed by RS algorithm**), assumed $n < \|\mathcal{P}\|$
$\mathcal{P}$	A population from which to sample, **of unknown size, without random access**
j	Ordinal $(1, 2, 3, \ldots)$ counter of elements from $\mathcal{P}$

Notational Oddities

Conventional "n choose k": choose k elements from a set of n

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

But we are using n as our sample size $|\mathcal{S}|$, and our "marker" for progression through $\mathcal{P}$ is j, so we say **"n items from a set of size j"**:

$$\binom{j}{n} = \frac{j!}{n!(j-n)!} \qquad \text{and the probability of each set occurring is} \qquad P_{j,n} = \frac{1}{\binom{j}{n}} = \frac{n!(j-n)!}{j!}$$

Step 1 (base case): show RS true for $|\mathcal{S}| = n = |\mathcal{P}| - 1$

(This was a question on a previous LGA...)

1. Place the first n elements of $\mathcal{P}$ into a list or array $\mathcal{S}$, there is now one element remaining in $\mathcal{P}$, call it a_m.
2. Flip a $\frac{n}{n+1}$ biased coin... $u \leftarrow Random()$
3. ...if it is heads ($u < \frac{n}{n+1}$), we pick an element of $\mathcal{S}$ in an unbiased manner (each has probability $\frac{1}{n}$) and replace it with a_m; the probability of each possible $\mathcal{S}$ containing a_m is the conjunction (AND) of these events

$$\Pr(\text{each } \mathcal{S} \text{ containing } a_m) = \frac{n}{n+1}\left(\frac{1}{n}\right) = \frac{1}{n+1}$$

4. ...if it is tails ($u > \frac{n}{n+1}$), we discard a_m and $\mathcal{S}$ is the sole sample of $\mathcal{P}$ without a_m.

$$\Pr(\mathcal{S} \text{ does not contain } a_m) = 1 - \frac{n}{n+1} = \frac{n+1-n}{n+1} = \frac{1}{n+1}$$

When $|\mathcal{S}| = n = |\mathcal{P}| - 1$, the RS algorithm is an unbiased construction of $\mathcal{S}$ from $\mathcal{P}$.

Step 2: assume RS true for an arbitrary $|\mathcal{S}| = n < j < |\mathcal{P}|$

Let $\mathcal{P}' \subset \mathcal{P}, |\mathcal{P}'| = j > n$.

By definition, if $\mathcal{S}$ is a random sample from $\mathcal{P}$, the probability of any particular $\mathcal{S}$ is

$$\Pr(\mathcal{S}) = \frac{1}{\binom{j}{n}} = \frac{n!(j-n)!}{j!} = P_{j,n}$$

Step 3: given (2), show RS true for $|\mathcal{S}| = n < j+1 \leq |\mathcal{P}|$

RS considers the adding $j+1^{\text{th}}$ element of $\mathcal{P}$ to $\mathcal{S}$ with probability $\frac{n}{j+1}$, there are two outcomes:

a. The $j+1^{\text{th}}$ element of $\mathcal{P}$ is discarded and never considered again.

b. The $j+1^{\text{th}}$ element of $\mathcal{P}$ **displaces** a pre-existing element of $\mathcal{S}$, the displaced element is chosen in an unbiased manner with probability $\frac{1}{n}$.

We want to show that the probability of all possible samples $\mathcal{S}$ after the $j+1^{\text{th}}$ element is processed is $P_{j+1,n}$.

Step 3, case a: the $j+1^{\text{th}}$ element of $\mathcal{P}$ is discarded

The probability of (any particular) $\mathcal{S}$ **not containing** the $j+1^{\text{th}}$ element is

$$P_{j,n}^{\triangle} \cdot \Pr(\text{ element discarded })$$

△ From step 2, the probability of choosing any possible $\mathcal{S}$ from the first j elements of $\mathcal{P}$.

$$P_{j,n} \cdot \Pr\left(u > \frac{n}{j+1}\right) = P_{j,n}\left(1 - \frac{n}{j+1}\right) = \underbrace{\left(\frac{n!(j-n)!}{j!}\right)}_{P_{j,n}} \cdot \left(\frac{j+1-n}{j+1}\right)$$

By definition, $j!(j+1) = (j+1)!$, likewise $(j-n)!(j+1-n) = (j+1-n)!$

$$\Pr\left(\text{ any particular } \mathcal{S} \text{ not containing } j+1^{\text{th}} \text{ element}\right) = \frac{n!(j+1-n)!}{(j+1)!} = P_{j+1,n}$$

Step 3, case b: the $j+1^{\text{th}}$ element of $\mathcal{P}$ accepted into $\mathcal{S}$

The probability of (each possible) $\mathcal{S}$ **containing** the $j+1^{\text{th}}$ element is

$$P_{j,n} \cdot \Pr(\text{ element accepted }) \cdot \left(\frac{1}{n}\right)^{\square}$$

- $\square$ Unbiased probability of choosing the displaced element

$$P_{j,n} \cdot \Pr\left(u < \frac{n}{j+1}\right) \cdot \frac{1}{n} \quad = \quad \frac{n!(j-n)!}{j!} \cdot \left(\frac{n}{j+1}\right) \cdot \frac{1}{n} \quad = \quad \frac{n!(j-n)!}{j!} \cdot \left(\frac{\cancel{n}}{j+1}\right) \cdot \frac{1}{\cancel{n}}$$

Step 3, case b: the $j+1^{\text{th}}$ element of $\mathcal{P}$ accepted into $\mathcal{S}$

The probability of (each possible) $\mathcal{S}$ **containing** the $j+1^{\text{th}}$ element is

$$P_{j,n} \cdot \Pr(\text{ element accepted }) \cdot \left(\frac{1}{n}\right)^{\square}$$

- $\square$ Unbiased probability of choosing the displaced element

$$P_{j,n} \cdot \Pr\left(u < \frac{n}{j+1}\right) \cdot \frac{1}{n} = \frac{n!(j-n)!}{j!} \cdot \left(\frac{n}{j+1}\right) \cdot \frac{1}{n} = \frac{n!(j-n)!}{j!} \cdot \left(\frac{\cancel{n}}{j+1}\right) \cdot \frac{1}{\cancel{n}}$$

$$= \frac{n!(j-n)!}{j!} \cdot \left(\frac{1}{j+1}\right) \neq \frac{n!(j+1-n)!}{(j+1)!} = P_{j+1,n}$$

Wait a tick, that's not right... What did we miss?

Step 3, case b: the $j+1^{\text{th}}$ element of $\mathcal{P}$ accepted into $\mathcal{S}$

$$P_{j,n} \cdot \Pr\left(u < \frac{n}{j+1}\right) \cdot \frac{1}{n} = \frac{n!(j-n)!}{j!} \cdot \left(\frac{n}{j+1}\right) \cdot \frac{1}{n} = \frac{n!(j-n)!}{j!} \cdot \left(\frac{\cancel{n}}{j+1}\right) \cdot \frac{1}{\cancel{n}}$$

$$= \frac{n!(j-n)!}{j!} \cdot \left(\frac{1}{j+1}\right) \neq \frac{n!(j+1-n)!}{(j+1)!} = P_{j+1,n}$$

Wait a tick, that's not right... What did we miss?

- **What are we missing?** There is more than one "pathway" to each possible sample $\mathcal{S}$ that includes the $j+1^{\text{th}}$ element.
  The $\frac{1}{n}$ term is wrong...

Step 3, case b: the $j+1^{\text{th}}$ element of $\mathcal{P}$ accepted into $\mathcal{S}$

$$P_{j,n} \cdot \Pr\left(u < \frac{n}{j+1}\right) \cdot \frac{1}{n} = \frac{n!(j-n)!}{j!} \cdot \left(\frac{n}{j+1}\right) \cdot \frac{1}{n} = \frac{n!(j-n)!}{j!} \cdot \left(\frac{\cancel{n}}{j+1}\right) \cdot \frac{1}{\cancel{n}}$$

$$= \frac{n!(j-n)!}{j!} \cdot \left(\frac{1}{j+1}\right) \neq \frac{n!(j+1-n)!}{(j+1)!} = P_{j+1,n}$$

Wait a tick, that's not right... What did we miss?

- **What are we missing?** There is more than one "pathway" to each possible sample $\mathcal{S}$ that includes the $j+1^{\text{th}}$ element.
  The $\frac{1}{n}$ term is wrong...
- Historical tidbit: when computers were really slow with limited memory, we used reservoir sampling **without** a formal proof of its correctness...

...the $\frac{1}{n}$ term is wrong, Let's do some Counting!

$|\mathcal{S} = \{C, E, F\}| = n = 3$ $\quad |\mathcal{P}'| = j = 8$ $\quad \mathcal{P}' \subset \mathcal{P}$ $\quad$ considering $j+1 = 9$

j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	...
$\mathcal{P}$	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**L**	**M**	**Q**	...
$\mathcal{S}$	**A**	**B**	**C**		**E**	**F**						
Discarded				**D**	**A**	**B**	**G**	**H**				

...the $\frac{1}{n}$ term is wrong, Let's do some Counting!

$|\mathcal{S} = \{C, E, F\}| = n = 3$ $\quad |\mathcal{P}'| = j = 8$ $\quad \mathcal{P}' \subset \mathcal{P}$ $\quad$ considering $j+1 = 9$

j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	10	11	...
$\mathcal{P}$	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**L**	M	Q	...
$\mathcal{S}$	**A**	**B**	**C**		**E**	**F**						
Discarded				**D**	**A**	**B**	**G**	**H**				

What is the probability of having $\mathcal{S} = \{\mathrm{C}, \mathrm{E}, \mathrm{L}\}$ after considering $j+1$?

…the $\frac{1}{n}$ term is wrong, Let's do some Counting!

$|\mathcal{S} = \{C,E,F\}| = n = 3$ $\quad |\mathcal{P}'| = j = 8$ $\quad \mathcal{P}' \subset \mathcal{P}$ $\quad$ considering $j+1 = 9$

j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	…
$\mathcal{P}$	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**L**	**M**	**Q**	…
$\mathcal{S}$	**A**	**B**	**C**		**E**	**F**						
Discarded				**D**	**A**	**B**	**G**	**H**				

Let $\alpha = \Pr(j+1^{\text{th}} \text{ element accepted})$, **notationally** $\Pr(\text{ACE}) = \Pr(\mathcal{S} = \{A,C,E\})$

$$\begin{aligned}\Pr(\text{CEL}) \;=\; & \Pr(\text{ACE})^{\nabla}\alpha\frac{1}{n}^{\text{pick A}} + \Pr(\text{BCE})^{\nabla}\alpha\frac{1}{n}^{\text{pick B}} + \Pr(\text{CDE})^{\nabla}\alpha\frac{1}{n}^{\text{pick D}} + \\ & + \Pr(\text{CEF})^{\nabla}\alpha\frac{1}{n}^{\text{pick F}} + \Pr(\text{CEG})^{\nabla}\alpha\frac{1}{n}^{\text{pick G}} + \Pr(\text{CEH})^{\nabla}\alpha\frac{1}{n}^{\text{pick H}}\end{aligned}$$

∇ Probability of this particular $\mathcal{S}$ **before** $j+1$ element considered.

. . . the $\frac{1}{n}$ term is wrong, Let's do some Counting!

$|\mathcal{S} = \{C, E, F\}| = n = 3 \qquad |\mathcal{P}'| = j = 8 \qquad \mathcal{P}' \subset \mathcal{P}$ considering $j+1 = 9$

j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	10	11	. . .
$\mathcal{P}$	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**L**	M	Q	. . .
$\mathcal{S}$	**A**	**B**	**C**		**E**	**F**						
Discarded				**D**	**A**	**B**	**G**	**H**				

$$\Pr(\text{CEL}) = \Pr(\text{ACE})\,\alpha\frac{1}{n}^{\text{pick A}} + \Pr(\text{BCE})\,\alpha\frac{1}{n}^{\text{pick B}} + \Pr(\text{CDE})\,\alpha\frac{1}{n}^{\text{pick D}} + \\ + \Pr(\text{CEF})\,\alpha\frac{1}{n}^{\text{pick F}} + \Pr(\text{CEG})\,\alpha\frac{1}{n}^{\text{pick G}} + \Pr(\text{CEH})\,\alpha\frac{1}{n}^{\text{pick H}}$$

	Particular Case	
How many addends?	$6 = 8 - 2$	
$\Pr(\textbf{pre-existing } \mathcal{S})^{\nabla}$**?**	$P_{8,3}$ $\triangle$	

…the $\frac{1}{n}$ term is wrong, Let's do some Counting!

$|\mathcal{S} = \{C,E,F\}| = n = 3$ $\quad |\mathcal{P}'| = j = 8$ $\quad \mathcal{P}' \subset \mathcal{P}$ $\quad$ considering $j+1 = 9$

j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	…
$\mathcal{P}$	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**L**	**M**	**Q**	…
$\mathcal{S}$	**A**	**B**	**C**		**E**	**F**						
Discarded				**D**	**A**	**B**	**G**	**H**				

$$\Pr(\text{CEL}) = \Pr(\text{ACE})\,\alpha\frac{1}{n}^{\text{pick A}} + \Pr(\text{BCE})\,\alpha\frac{1}{n}^{\text{pick B}} + \Pr(\text{CDE})\,\alpha\frac{1}{n}^{\text{pick D}} + \\ + \Pr(\text{CEF})\,\alpha\frac{1}{n}^{\text{pick F}} + \Pr(\text{CEG})\,\alpha\frac{1}{n}^{\text{pick G}} + \Pr(\text{CEH})\,\alpha\frac{1}{n}^{\text{pick H}}$$

	Particular Case	**General Case**
How many addends?	$6 = 8-2$	$j-(n-1)$
$\Pr(\textbf{pre-existing } \mathcal{S})^{\nabla}$**?**	$P_{8,3}$ $\triangle$	$P_{j,n}$ $\triangle$

. . . the $\frac{1}{n}$ term is wrong, Let's do some Counting!

$$\Pr(\text{CEL}) = \Pr(\text{ACE})\,\alpha\frac{1}{n}^{\text{pick A}} + \Pr(\text{BCE})\,\alpha\frac{1}{n}^{\text{pick B}} + \Pr(\text{DCE})\,\alpha\frac{1}{n}^{\text{pick D}} + \\ + \Pr(\text{CEF})\,\alpha\frac{1}{n}^{\text{pick F}} + \Pr(\text{CEG})\,\alpha\frac{1}{n}^{\text{pick G}} + \Pr(\text{CEH})\,\alpha\frac{1}{n}^{\text{pick H}}$$

	Particular Case	**General Case**
How many addends?	$6 = 8-2$	$j-(n-1)$
$\Pr(\textbf{pre-existing } \mathcal{S})^{\nabla}$**?**	$P_{8,3}$	$P_{j,n}$

$$\Pr(\text{CEL}) = \sum_{i=1}^{j-(n-1)} P_{j,n}\alpha\frac{1}{n} = P_{j,n}\alpha\left[\sum_{i=1}^{j-(n-1)}\frac{1}{n}\right]$$

$$= P_{j,n}\cdot \underbrace{\Pr(j+1^{\text{th}}\text{ element accepted})}_{\alpha} \underbrace{\left(\frac{j-(n-1)}{n}\right)}_{\text{corrected } \frac{1}{n} \text{ term}}$$

. . . and the RHS is generalized

Step 3, case b: the $j+1^{\text{th}}$ element of $\mathcal{P}$ accepted into $\mathcal{S}$ (CORRECTED)

The probability of each possible $\mathcal{S}$ containing the $j+1^{\text{th}}$ element is

$$P_{j,n} \cdot \Pr(j+1^{\text{th}} \text{ element accepted }) \cdot \underbrace{\left(\frac{j-(n-1)}{n}\right)}_{\text{corrected } \frac{1}{n} \text{ term}}$$

$$= \quad P_{j,n} \cdot \Pr\left(u < \frac{n}{j+1}\right) \cdot \left(\frac{j+1-n}{n}\right) \quad = \quad \left(\frac{n!(j-n)!}{j!}\right)\left(\frac{\cancel{n}}{j+1}\right)\left(\frac{j+1-n}{\cancel{n}}\right)$$

By definition, $j!(j+1) = (j+1)!$, likewise $(j-n)!(j+1-n) = (j+1-n)!$

$$= \quad \frac{n!(j+1-n)!}{(j+1)!} \quad = \quad P_{j+1,n}$$

Tying a Pretty Bow on Top

Let a_m be the $j+1^{\text{th}}$ element of $\mathcal{P}$.

1 RS works when $|\mathcal{S}| = n = |\mathcal{P}| - 1$

2&3 If a_m is discarded, the probability of having each possible sample set **not containing** a_m is $P_{j+1,n}$.
If a_m is accepted into $\mathcal{S}$, the probability of having each possible sample set **containing** a_m is $P_{j+1,n}$.

$\Rightarrow$ After the $j+1^{\text{th}}$ element of $\mathcal{P}$ is considered for inclusion of $\mathcal{S}$, **all possible sample sets have an equal probability of occurring, namely** $P_{j+1,n}$.

There are $\binom{j+1}{n}$ such sets, so the probabilities add to one:

$$\sum_{i=1}^{\binom{j+1}{n}} P_{j+1,n} = \underbrace{\binom{j+1}{n} P_{j+1,n} = \binom{j+1}{n} \frac{1}{\binom{j+1}{n}}}_{\text{Uses definition of } P_{j,n}} = 1 \qquad \square$$