1. Suppose you would like to take an SRS of size n from a list of N units, but do not know the population size N in advance. Consider the following procedure:

(a) Set $S_0 = \{1, 2, \ldots, n\}$, so that the initial sample for consideration consists of the first n units on the list.

(b) For $k = 1, 2, \ldots$, generate a random number u_k between 0 and 1. If $u_k > n/(n+k)$, then set S_k equal to S_{k-1}. If $u_k \le n/(n+k)$, then select one of the units in S_{k-1} at random, and replace it by unit $(n+k)$ to form S_k.

Show that S_{N-n} from this procedure is an SRS of size n. *Hint:* Use induction.

Solution:

Fix the sample size at n, and let $k = N - n$ be the difference in the size between the population to the sample. For clarity, denote $S_k : \Omega \to \{1, 2, \ldots, (n+k)\}^n$ a random variable whose realizations are denoted $S_k(\omega) = s_k$. We must show that S_k takes on each subset of $\{1, 2, \ldots, N = n+k\}$ of size n with equal probability. We proceed by induction on k from $k = 0$ to $k = N - n$.

In the base case, $k = 0$ implies $N = n$ and S_0 takes on only $s_0 = \{1, \ldots, n\}$ with probability 1, which is the only subset of $\{1, \ldots, N = n+0\}$ of size n. Hence it is trivially a random sample of all such subsets.

Now, assume the induction hypothesis, i.e. S_k produces an SRS for a population of size $n + k$ – equivalently,

$$P(S_k = s_k) = 1 \Big/ \binom{n+k}{n}$$

for each subset s_k of size n from $\{1, 2, \ldots, (n+k)\}$. Upon the realization $S_{k+1} = s_{k+1}$, either the sample remains fixed with $S_{k+1} = s_k$, in which case $u_{k+1} \in \left(\frac{n}{n+k+1}, 1\right)$ with probability $\left(1 - \frac{n}{n+k+1}\right) = \left(\frac{k+1}{n+k+1}\right)$, or S_{k+1} takes on s_k with a randomly selected element replaced with $(n+k+1)$, in which case $u_{k+1} \in \left(0, \frac{n}{n+k+1}\right]$ with probability $\left(\frac{k+1}{n+k+1}\right)$. If we assume that the selection of u_{k+1} is independent of the generation of the sample from the previous step, then in the first case

$$\begin{aligned} P(S_{k+1} = s_{k+1}) &= P(S_k = s_k) P\left(U_{k+1} \in \left(\frac{n}{n+k+1}, 1\right)\right) \\ &= \binom{n+k}{n}^{-1} \left(\frac{k+1}{n+k+1}\right) \\ &= \binom{n+k+1}{n}^{-1} \end{aligned}$$

In the second case, denote $s_k = \{a_1, a_2, \ldots, a_n\}$, then $s_{k+1} = s_k \setminus \{a_i\} \cup \{k+1+n\}$ where a_i is the realization of the random selection in the algorithm. For each event where $\{a_1, a_2, \ldots, a_n\} \setminus a_i \subset S_k$, there are $(n+k) - (n-1) = k+1$ possible choices that fix each $a_1, \ldots, a_n$ except for a_i. Since each of those events are disjoint and have equal probability, $P(S_{k+1} = s_{k+1} | u_k, a_i) = P(\{a_1, \ldots, a_n\} \setminus a_i \in S_k) = (k+1)\binom{n+k}{n}^{-1}$. Now, assuming the independence of realization of u_k and a_i and the previous selection, we have

$$\begin{aligned}
P(S_{k+1} = s_{k+1}) &= \left((k+1)\binom{n+k}{n}^{-1}\right) \cdot P\left(U_{k+1} \in \left(0, \frac{n}{n+k+1}\right]\right) \cdot P\Big(A_i = a_i\Big) \\
&= (k+1)\binom{n+k}{n}^{-1} \cdot \frac{n}{n+k+1} \cdot \left(\frac{1}{n}\right) \\
&= \binom{n+k}{n}^{-1} \frac{k+1}{n+k+1} \\
&= \binom{n+k+1}{n}^{-1}.
\end{aligned}$$

Note that the collection of possible realizations $S_{k+1} = s_{k+1}$ contains all realizations $S_k = s_k$ which is each subset of size n $\{1, \ldots, n+k\}$ by the induction hypothesis. Certainly, s_{k+1} is a subset of size n of $\{1, 2, \ldots, n+k+1\}$ as possibly only $(n+k+1)$ is added to s_k in the $(k+1)$th step, and if $\{a_1, \ldots, a_n\}$ is a subset of $\{1, \ldots, n+k+1\}$, then if $(n+k+1) = a_i$ for some a_i, then $\{a_1, \ldots, a_i = (n+k+1), \ldots, a_n\}$ is a possible realization in step two, otherwise $(n+k+1)$ is not a_i for any i, and the subset is realized in step 1. Hence each subset of size n from $\{1, 2, \ldots, k+n+1\}$ is realized.

$\square$