Summer of Math Exposition 4: Sampling

Ian Chen

August 7, 2025

Sampling is a fundamental part in statistical studies. Instead of collecting data from a large population, which is often infeasible at worse and expensive at best, statisticians collect a ***representative sample*** and infer conclusions about the population from the sample. Here, we will explore algorithms to efficiently sample from data streams.

We expect a prerequisite knowledge of Big-O and elementary probability. See §4 for a refresher.

In §1, we develop the problem and a simple solution. Then, releasing the assumption that we know the population size, we explore a simple type of resevoir sampling algorithm in §2, and refine it to optimality in §3.

1 Problem Definition

Let $D = (x_1, x_2, \ldots)$ be a data stream with weights $(w_1, w_2, \ldots)$. Let k be the number of samples we wish the collect, and N the size of the stream (if known). We wish to compute a sample $S(D) \subset D$, where $\mathbf{Pr}(x_i \in S(D)) \propto w_i$ and $\left|S(D)\right| = \min(n, \left|D\right|)$. That is, we want to sample without replacement. We are allowed to use the $random()$, generating a uniform real in $[0, 1)$, and the $randint$(a, b), generating a uniform integer in $[a, b]$, functions.

Here's an example:

There are a few properties of this approach that can be undesireable, which all stem from the fact that we need to store the entire data stream. This is because we make multiple passes of the data, first in finding the sum of all the weights, and a new pass when selecting an element (and removing). From here on, we explore resevoir sampling [5], which are single-pass sampling techniques.

2 Bottom-k Sampling

First, let us assume unit weights, namely $w_i = 1$.

For a survey [2].

2.1 Application

For the main algorithm of estimating reach sizes [1].

3 Optimal Resevoir Sampling

For optimality [5]. For the unweighted, [4], and the weighted, [3].

4 Prerequisites

References

[1] Edith Cohen. "Size-Estimation Framework with Applications to Transitive Closure and Reachability". In: *Journal of Computer and System Sciences* 55.3 (Dec. 1997), pp. 441–453. ISSN: 0022-0000. DOI: 10.1006/jcss.1997.1534. URL: http://dx.doi.org/10.1006/jcss.1997.1534.

[2] Edith Cohen and Haim Kaplan. "Summarizing data using bottom-k sketches". In: *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*. PODC07. ACM, Aug. 2007, pp. 225–234. DOI: 10.1145/1281100.1281133. URL: http://dx.doi.org/10.1145/1281100.1281133.

[3] Pavlos S. Efraimidis and Paul G. Spirakis. "Weighted random sampling with a reservoir". In: *Information Processing Letters* 97.5 (Mar. 2006), pp. 181–185. ISSN: 0020-0190. DOI: 10.1016/j.ipl.2005.11.003. URL: http://dx.doi.org/10.1016/j.ipl.2005.11.003.

[4] Kim-Hung Li. "Reservoir-sampling algorithms of time complexity O(n(1 + log(N/n)))". In: *ACM Trans. Math. Softw.* 20.4 (Dec. 1994), pp. 481–493. ISSN: 0098-3500. DOI: 10.1145/198429.198435. URL: https://doi.org/10.1145/198429.198435.

[5] Jeffrey S. Vitter. "Random sampling with a reservoir". In: *ACM Trans. Math. Softw.* 11.1 (Mar. 1985), pp. 37–57. ISSN: 0098-3500. DOI: 10.1145/3147.3165. URL: https://doi.org/10.1145/3147.3165.