I'm currently investigating the use of *Bloom filters* to create scalable machine vision algorithms. To analyze how discerning an algorithm that is based on Bloom filters can be, I needed the following result[1].

Problem Suppose we have an array of bits $A = a_0, a_1, \ldots, a_{n-1}$ of length n and an ideal hash function h such that the distribution of $h(k)$ is uniform in the range $[0, n)$. Given an arbitrary sequence of keys $K = k_0, k_1, \ldots, k_{m-1}$ we set

$$a_i = \begin{cases} 1 & \text{if } h(k) = i \text{ for some } k \in K \\ 0 & \text{otherwise} \end{cases}$$

How many of the bits may be expected to be set after the m keys have been hashed? Treating A as a characteristic function over the first n natural numbers, we can rephrase this as: what is the expected size of the set $\{h(k_i) \mid 0 \leq i < m\}$?

Solution Let $E_n(x)$ be the expected number of bits set after hashing the first x keys. Then

$$\begin{aligned} E_n(0) &= 0 \\ E_n(x+1) &= P(h(k_{x+1}) \text{ is set}) \cdot E_n(x) + P(h(k_{x+1}) \text{ is not set}) \cdot (E_n(x) + 1) \\ &= \frac{E_n(x)}{n} \cdot E_n(x) + \frac{n - E_n(x)}{n} \cdot (1 + E_n(x)) \\ &= 1 + (\frac{n-1}{n}) \cdot E_n(x) \end{aligned}$$

Thus $E_n(x)$ may be expressed as the sum of a geometric series:

$$\begin{aligned} E_n(x) &= \sum_{0 \leq i < x} (\frac{n-1}{n})^i \\ &= \frac{1 - (\frac{n-1}{n})^x}{1 - (\frac{n-1}{n})} \\ &= n[1 - (\frac{n-1}{n})^x] \end{aligned}$$

Tom Gibara
April 2007

[1] I intend to write more about the algorithm here if it develops into something more substantial.