Sketch Equations

Lee Rhodes
Yahoo! Inc., 701 First Ave., Sunnyvale, CA 94089, USA

December 7, 2017

Abstract

The math behind the sketch algorithms for unique counting has been well described in papers by Giroire[2], Bar Yossef[1] and many others. The presentation of these concepts in the theoretical research literature is often abstract with details deferred to other papers in order to save space. This makes acquiring intuitive understanding of these mathematical concepts a challenge if one is not familiar with this scientific discipline and its mathematical conventions. The objective here is to develop the important mathematical concepts so that individuals with a background of first-year college calculus can follow them.

1 Introduction

The Theta Sketch Framework encompasses many possible sketch algorithms only a few of which have been implemented in the Sketch Library. In this paper we will discuss the mathematics of the Bernoulli Sampling and KMV algorithms in some detail. This should provide sufficient understanding of how these kinds of algorithms work. For the analysis of the Alpha algorithm we will defer to the above TSF paper.

2 Hypothetical Sketch Produced by Bernoulli Sampling

2.1 Fixed Theta Sampling

Suppose we have a stream A of n items $a_1, a_2, \ldots, a_n$ and an arbitrary, fixed sampling probability $1 > \theta > 0$.

The traditional Bernoulli variable, b_i, is defined as a random, independent, weighted coin flip for each item a_i:

$$b_i = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1-\theta. \end{cases}$$

Equivalently, we could compute a uniform hash on the interval [0,1] for each a_i. The Bernoulli variable becomes:

$$b_i = \begin{cases} 1 & h(a_i) < \theta \\ 0 & h(a_i) \geq \theta. \end{cases}$$

We will use this latter definition as it more closely aligns with what we actually do.

Our sketch consists of two elements, a set S of hash values $h(a_i)$ selected by the above Bernoulli sampling process, and a predefined value of θ. Note that we don't actually implement this algorithm. It is impractical as we do not know θ upfront. Suspend disbelief for a few moments and pretend that we did. The payoff will be that the mathematics is relatively straightforward.

From the Bernoulli Distribution the expected value, mean and variance are

$$E[b_i = 1] = \mu = \theta$$
$$\sigma^2(b_i) = \theta(1-\theta)$$

A stream of n Bernoulli Trials defines a sample set S of size $|S|$:

$$|S| = \sum_{i \in n} b_i \quad \text{where } |S| \text{ is a random variable.}$$

The expected value of $|S|$ is

$$E\left[|S|\right] = E\left[\sum_{i \in n} b_i\right] = \sum_{i \in n} E[b_i] = n\theta.$$

Because the samples are independent, the variance is

$$\sigma^2\left(|S|\right) = \sum_{i \in n} \sigma^2(b_i) = n\theta(1-\theta).$$

The estimate of n, is simply

$$\hat{n} = \frac{|S|}{\theta}. \tag{2.1}$$

To establish unbiasedness we compute the expected value of $\hat{n}$

$$E[\hat{n}] = E\left[\frac{1}{\theta}\sum_{i \in n} b_i\right] = \frac{1}{\theta}\sum_{i \in n} E(b_i) = \frac{1}{\theta} n\theta = n.$$

To understand the error, we compute the variance of $\hat{n}$,

$$\sigma^2(\hat{n}) = \sigma^2\left(\frac{|S|}{\theta}\right) = \frac{1}{\theta^2}\sigma^2(|S|) = \frac{1}{\theta^2}(n\theta(1-\theta)) = \frac{n}{\theta} - n.$$

To compute the Relative Standard Error, we divide by n^2 and take the square root

$$\text{RSE}(\hat{n}) = \sqrt{\frac{\sigma^2(\hat{n})}{n^2}} = \sqrt{\frac{1}{n\theta} - \frac{1}{n}} = \sqrt{\frac{1}{E\left[|S|\right]} - \frac{1}{n}} < \frac{1}{\sqrt{E\left[|S|\right]}}. \tag{2.2}$$

2.2 Fixed k Sampling

The above derivation assumed a fixed θ sampling where the size of the sample set,$|S|$, is a random variable. It is not hard to imagine turning this around so that the resulting size of the sample set is bounded by a constant k, and then θ becomes a variable that must be constantly adjusted by the sketch algorithm as new items arrive so that $k = n\theta$. (We haven't discussed how to do that!)

Equations for the estimate (2.1) and the RSE (2.2) become

$$\hat{n} = \frac{k}{\theta} \tag{2.3}$$

$$\text{RSE}(\hat{n}) = \sqrt{\frac{1}{k} - \frac{1}{n}} < \frac{1}{\sqrt{k}} \tag{2.4}$$

2.3 Subsets of Fixed k Sampling

Suppose we were to choose, by set operations or other means, a subset, S_{sub} of the k samples in the sketch (S, θ) to represent a subpopulation of the original n. The estimate $\widehat{n_{sub}} = |S_{sub}|/\theta$ and the RSE$(\widehat{n_{sub}}) = 1/\sqrt{|S_{sub}|}$.

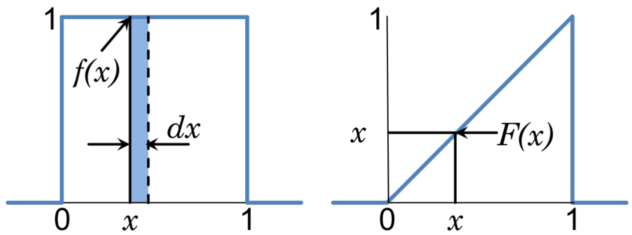

Figure 1: Uniform density (left) and distribution (right)

3 KMV Equations

We do implement a variant of the KMV sketch in the Sketch Library called the QuickSelect sketch. The subtle differences between the conventional definition of the KMV sketch and the QuickSelect sketch is summarized at the end. This derivation is similar to that of Giroire[2], but is more direct and includes rudimentary steps that Giroire omits.

3.1 Preliminaries

3.1.1 Uniform Probability Density and Distribution

One of the fundamental concepts of sketches is that the raw input stream of unique values is transformed into a stream of unique hash values that have a uniform random distribution. This is accomplished internally by a hash function that has good avalanche and bit-independence properties.

We begin by defining a continuous uniform random variable X on the interval [0,1]:

The Prabability Density Function (PDF) (Figure 1, left):

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

$$\int_0^1 f(x)dx = 1$$

$$f(x_0) = P(x = x_0)$$

The Cumulative Distribution Function (CDF) (Figure 1, right):

$$\begin{aligned} F(x_0) &= P(x < x_0) \\ &= \int_0^{x_0} f(x)dx = x_0 \end{aligned} \tag{3.2}$$

3.1.2 Expected Value of $g(x)$ Given Density $f(x)$

Given a random variable X with a density function $f(x)$, and another function of X, $g(x)$, the expected value of $g(x)$ is given by the inner product of f and g. See Wikipedia Expected Value and Appendix A for a discrete example.

$$E[g(x)] = \int_0^1 g(x)f(x)dx \tag{3.3}$$

3.1.3 Euler Beta Function

The Euler Beta function is a special function that has different forms that can be very useful depending on the context. It is particularly useful in solving the integrals that occur in Order Statistics by converting the integrals into Gamma or Factorial expressions.

For $\mathbb{R}(a), \mathbb{R}(b) > 0$

$$\mathbf{B}(a,b) = \int_{t=0}^{1} t^{a-1}(1-t)^{b-1}dt \tag{3.4}$$
$$= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

For integers $a, b > 0$

$$\Gamma(a) = (a-1)!$$
$$\mathbf{B}(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!} \tag{3.5}$$

3.1.4 The k^{th} Order Statistic, part 1

We start with a set of n labeled random variables $X_1, \ldots, X_n$ in the interval $[0,1]$ that have a density $f(x)$ and a distribution $F(x)$. If we take one instance of all the $X's$, we can order them and identify them by their order $X_{(1)}, \ldots, X_{(n)}$, which is independent of the labels. Our goal is to find the density function and expected value of the k^{th} minimum value (KMV), $M_{(k)}$. This analysis only assumes that the underlying probability density of the $X's$ is a real analytic function. At the end of the analysis we simplify to the uniform random density case.

The density of $M_{(k)}$ (Figure 2)

$$f_{(k)}(x)dx = P(M_{(k)} \in dx)$$
$$= P(\text{exactly one of } X's \in dx, \text{exactly } k-1 \text{ of } X's < x)$$

There are n $X's$, each with the same $f(x)$.

$$= P(X_1 \in dx, \text{exactly } k-1 \text{ of the other } X's < x) +$$
$$P(X_2 \in dx, \text{exactly } k-1 \text{ of the other } X's < x) +$$
$$\cdots +$$
$$P(X_n \in dx, \text{exactly } k-1 \text{ of the other } X's < x).$$

$$f_{(k)}(x)dx = nP(X_1 \in dx, \text{exactly } k-1 \text{ of the other } X's < x) \quad \text{choice of } X_1 \text{ is arbitrary}$$

From probability theory.

$$P(X_1 \in dx) = f(x)dx$$
$$P(\text{at least } (k-1)X's < x) = (F(x))^{k-1}$$
$$P(\text{at least } (n-k)X's > x) = (1-F(x))^{n-k}$$

Note that there are $\binom{n-k}{k-1}$ combinations of $(k-1)$ $X's < x$.

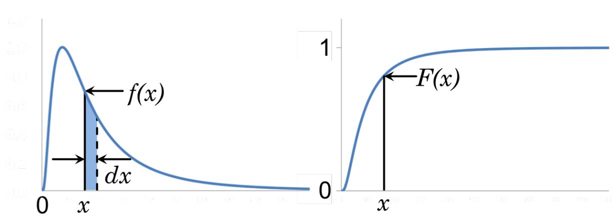

Figure 2: Some arbitrary density (left) and distribution (right)

To force exactly $(k-1)$ $X's < x$ we partition the probability space into three parts: $X \in dx, X < x, X > x + dx$.

$$f_{(k)}(x)dx = n \quad f(x)dx \quad \binom{n-1}{k-1}(F(x))^{k-1} \quad (1-F(x))^{n-k}$$

Let's simplify the above by assuming the uniform random probability density instead of an arbitrary density. Recall that $f(x_0) = 1$ and $F(x) = x_0$ from 3.1 and 3.2.

$$\begin{aligned} f_{(k)}(x)dx &= n\binom{n-1}{k-1}x^{k-1}(1-x)^{n-k}dx \\ &= k\binom{n}{k}x^{k-1}(1-x)^{n-k}dx \end{aligned} \tag{3.6}$$

The Expected Value of $M_{(k)}$ becomes

$$\begin{aligned} E[M_{(k)}] &= \int_0^1 (x)\left[k\binom{n}{k}x^{k-1}(1-x)^{n-k}\right]dx \\ &= k\binom{n}{k}\int_0^1 x^k(1-x)^{n-k}dx \end{aligned} \tag{3.7}$$

3.1.5 The k^{th} Order Statistic, part 2: Using the Beta Function

The integral of 3.7 can be recognized as a form of the Beta integral from 3.4.

Let $a = k+1, b = n-k+1, a+b = n+2$

$$\mathbf{B}(k+1, n-k+1) = \int_{t=0}^1 t^k(1-t)^{n-k}dt,$$

and the Beta factorial form from 3.5

$$= \frac{k!(n-k)!}{(n+1)!}.$$

This means that 3.7 can be written

$$\begin{aligned} E[M_{(k)}] &= k\binom{n}{k}\mathbf{B}(k+1, n-k+1) \\ &= k\binom{n}{k}\frac{k!\,(n-k)!}{(n+1)!} \\ &= \frac{k\,n!}{k!(n-k)!}\,\frac{k!\,(n-k)!}{(n+1)!} \\ E[M_{(k)}] &= \frac{k}{n+1} \end{aligned} \tag{3.8}$$

3.2 The Expected Value of the Inverse of $M_{(k)}$

In order to estimate n we need to derive $E[\frac{1}{M_{(k)}}]$. From 3.3 and 3.6 we have

$$\begin{aligned} E\left[\frac{1}{M_{(k)}}\right] &= \int_0^1 \left(\frac{1}{x}\right)\left[k\binom{n}{k}x^{k-1}(1-x)^{n-k}\right]dx \\ &= k\binom{n}{k}\int_{x=0}^1 x^{k-2}(1-x)^{n-1}dx. \end{aligned} \tag{3.9}$$

Again using the Beta integral and factorial forms from 3.4 and 3.5 for the integral in 3.9:

Let $a = k-1, b = n-k+1, a+b = n$

$$\mathbf{B}(k-1, n-k+1) = \int_{t=0}^{1} t^{k-2}(1-t)^{n-k}dt$$
$$= \frac{(k-2)!(n-k)!}{(n-1)!} \tag{3.10}$$

Substituting 3.10 into 3.9:

$$E\left[\frac{1}{M_{(k)}}\right] = k\binom{n}{k}\mathbf{B}(k-1,\ n-k+1)$$
$$= \frac{k\ n!}{(n-k)!\ k!}\frac{(k-2)!\ (n-k)!}{(n-1)!}$$
$$E\left[\frac{1}{M_{(k)}}\right] = \frac{n}{k-1} \tag{3.11}$$

We don't know n. What we want is $\hat{n}$, an asymptotically unbiased estimate of n.
Solving 3.11 for n it becomes the estimate, $\hat{n}$

$$\hat{n} = (k-1)\left(\frac{1}{M_{(k)}}\right) = \frac{k-1}{M_{(k)}} \tag{3.12}$$

$$E[\hat{n}] = (k-1)E\left[\frac{1}{M_{(k)}}\right] = (k-1)\frac{n}{k-1} = n. \tag{3.13}$$

This proves that our estimate of n is indeed unbiased.

3.3 The Variance of $\hat{n}$

From the expected value of $\hat{n}$ from 3.13 we have:

$$E[\hat{n}] = n$$

The variance of $\hat{n}$

$$\sigma^2[\hat{n}] = E[\hat{n}^2] - E[\hat{n}]^2 = E[\hat{n}^2] - n^2$$

Evaluating the term, $E[\hat{n}^2]$ by squaring 3.13:

$$E[\hat{n}^2] = (k-1)^2 E\left[\left(\frac{1}{M_{(k)}}\right)^2\right]$$

Evaluating the term, $E\left[\left(\frac{1}{M_{(k)}}\right)^2\right]$

$$E\left[\left(\frac{1}{M_{(k)}}\right)^2\right] = \int_0^1 \left(\frac{1}{x^2}\right)\left[k\binom{n}{k}x^{k-1}(1-x)^{n-k}\right]dx$$
$$= k\binom{n}{k}\int_0^1 x^{(k-2)-1}(1-x)^{(n-k+1)-1}dx$$

Again using the Beta integral and factorial forms from 3.4 and 3.5:
Let $a = k-2, b = n-k+1, a+b = n-1$

$$= k\binom{n}{k}\mathbf{B}(k-2, n-k+1)$$
$$= \frac{k\ n!}{(n-k)!\ k!}\frac{(k-3)!\ (n-k)!}{(n-2)!}$$
$$E\left[\left(\frac{1}{M_{(k)}}\right)^2\right] = \frac{n\ (n-1)}{(k-1)\ (k-2)}$$

Returning to the evaluation of $E[\hat{n}^2]$:

$$\begin{aligned} E[\hat{n}^2] &= (k-1)^2 \frac{n\,(n-1)}{(k-1)\,(k-2)} \\ &= \frac{n\,(n-1)\,(k-1)}{(k-2)} \end{aligned}$$

Returning to the evaluation of $\sigma^2[\hat{n}]$

$$\begin{aligned} \sigma^2[\hat{n}] &= \frac{n\,(n-1)\,(k-1)}{(k-2)} - n^2 \\ &= \frac{n\,(n-1)\,(k-1) - (k-2)\,n^2}{k-2} \\ \sigma^2[\hat{n}] &= \frac{n^2 - n(k-1)}{k-2} < \frac{n^2}{k-2} \end{aligned}$$

Normalizing the variance by n^2 and taking the square root results in the Relative Standard Error (RSE):

$$RSE[\hat{n}] = \sqrt{\frac{\sigma^2}{n^2}} = \sqrt{\frac{1}{k-2} - \frac{k-1}{n(k-2)}} < \frac{1}{\sqrt{k-2}} \tag{3.14}$$

$$RSE[\hat{n}]_{n\to\infty} = \frac{1}{\sqrt{k-2}} \tag{3.15}$$

This proves that the RSE is always less than a constant!

3.4 Equation Differences Between KMV and QuickSelect

Much of the research literature on KMV sketches defines a cache of size k that holds the k^{th} minimum value ($M_{(k)}$) and $k-1$ hash values less than $M_{(k)}$. The Theta Sketch Framework (TSF), however, is more flexible and different from the standard KMV definition. In the TSF, the label k is used as a user configured parameter that defines the maximum RSE for the sketch. Instead of $M_{(k)}$ being a member of the cache array of hash values, TSF sketches have a separate register called θ and define different Theta Choosing Functions (TCF) algorithms for computing θ for different sketch families. See Theta Sketch Framework.

This requires minor changes to the above equations for the QuickSelect family.

Ref / Equation	KMV	QuickSelect
TCF(θ) $=$	$M_{(k)}$	$M_{(k+1)}$
3.8 $E[\theta] =$	$\frac{k}{n+1}$	$\frac{k+1}{n+1}$
3.11 $E\left[\frac{1}{\theta}\right] =$	$\frac{n}{k-1}$	$\frac{n}{k}$
3.12 $\hat{n} =$	$\frac{k-1}{\theta}$	$\frac{k}{\theta}$
3.14 $RSE(\hat{n}) \leq$	$\frac{1}{\sqrt{k-2}}$	$\frac{1}{\sqrt{k-1}}$

References

[1] Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. *Randomization and Approximation Techniques In Computer Science*, pages 110. Springer, 2002.

[2] F. Giroire. Order statistics and estimating cardinalities of massive data sets. *Discrete Applied Mathematics*, 157-2:406–427, 2009.

[3] Anirban Dasgupta, Kevin Lang, Lee Rhodes, Justin Thaler. A Framework for Estimating Stream Expression Cardinalities, Theta Sketch Framework

A Discrete Example of $E[g(x)]$, Equation 3.3

Assume our random variable X is the result of rolling a single, fair 6-sided die. It's density and distribution are shown in Figure 3.

$$\begin{aligned} x &= \text{The face value of rolling the die once} \\ x_i &= \text{One of the specific (labeled) face values: } 1, 2, 3, 4, 5, 6 \\ f(x) &= \text{The density function of } x \\ &= \frac{1}{6} \\ g(x) &= \text{A function that produces a value given } x \\ &= x \end{aligned}$$

The expected value (average) of rolling the die many times

$$\begin{aligned} E[X] &= \sum_{i=1}^{6} g(x_i) f(x_i) \\ &= \frac{1}{6} \sum_{i=1}^{6} g(x_i) \\ &= \frac{1}{6}(1+2+3+4+5+6) \\ &= 3.5 \end{aligned}$$

Let's compute the expected value of the inverse of X

$$\begin{aligned} g_2(x) &= \frac{1}{x} \\ E\left[\frac{1}{X}\right] &= \sum_{i=1}^{6} g_2(x_i) f(x_i) \\ &= \frac{1}{6} \sum_{i=1}^{6} g_2(x_i) \\ &= \frac{1}{6}\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}\right) \\ &= 0.408\overline{3} \end{aligned}$$

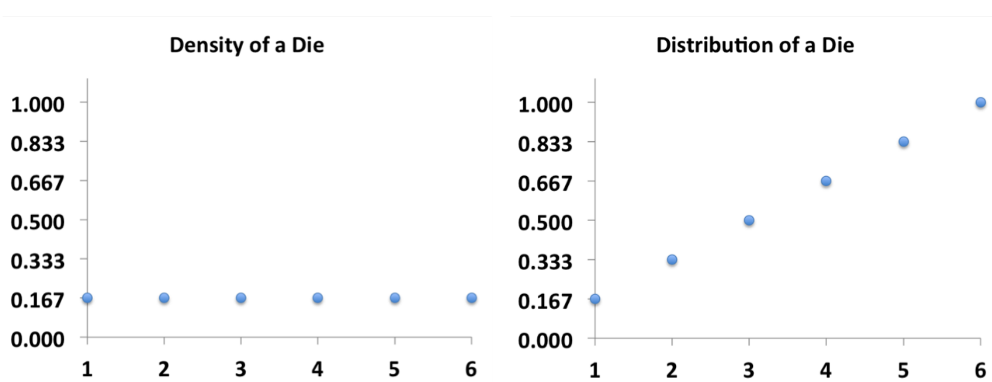

Figure 3: Density and distribution of a single die