The Adaptive sampling revisited

Matthew Drescher * Guy Louchard † Yvik Swan ‡

November 23, 2019

Abstract

The problem of estimating the number n of distinct keys of a large collection of N data is well known in computer science. A classical algorithm is the adaptive sampling (AS). The cardinality n can be estimated by $R \cdot 2^D$, where R is the final bucket (cache) size and D is the final depth at the end of the process. Several new interesting questions can be asked about AS (some of them were suggested by P.Flajolet and popularized by J.Lumbroso). The distribution of $W = \log(R2^D/n)$ is known, we rederive this distribution in a simpler way. We provide new results on the moments of D and W. We also analyze the final cache size R distribution. We consider colored keys: assume that among the n distinct keys, n_C do have color C. We show how to estimate $p = \frac{n_C}{n}$. We also study colored keys with some multiplicity given by some distribution function. We want to estimate mean and variance of this distribution. Finally, we consider the case where neither colors nor multiplicities are known. There we want to estimate the related parameters. An appendix is devoted to the case where the hashing function provides bits with probability different from 1/2.

Keywords: Adaptive sampling, moments, periodic components, hashing functions, cache, colored keys, key multiplicity, Stein method, urn model, asymmetric adaptive sampling

2010 Mathematics Subject Classification: 68R05, 68W40.

1 Introduction

The problem of estimating the number n of distinct keys of a large collection of N data is well known in computer science. It arises in query optimization of database systems. It has many practical applications. For example consider a stream (or log) of login events for a large popular website. We would like to know how many unique users login per month. A naive approach would be to insert each login credential into a set which is stored in memory. The cardinality of the set will of course be equal to the number of unique logins. However, if the number of distinct logins makes the cardinality of the set too large to fit into memory, this simple method will not work. While strategies involving more machines and/or writing to disk exist, see the paper by Rajaraman and Ullman [19], the estimation technique we study here is an alternative requiring no additional infrastructure.

A classical algorithm is the adaptive sampling (AS), the original idea of which is due to Wegman in [23]. The mean and variance of AS are considered by Flajolet in [7] . Let us summarize the principal features of AS. Elements of the given set of N data are hashed into binary keys. These keys are infinitely long bit streams such that each bit has probability 1/2 of being 0 or 1. A uniformity assumption is made on the hashing function .

The algorithm keeps a bucket (or cache) B of at most b distinct keys. The depth of sampling, d which is defined below, is also saved. We start with $d = 0$ and throw only distinct keys into B. When

*Université Libre de Bruxelles, Département d'Mathématique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium, email: knavely@gmail.com

†Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium, email: louchard@ulb.ac.be

‡Université de Liège, Département de Mathématique, zone polytech 1, 12, allée de la découverte, Bât. B37 pkg 33a, B-4000 Liège, Belgium, email: yswan@ulg.ac.be

B is full, depth d is increased by 1, the bucket is scanned, and only keys starting with 0 are kept. (If the bucket is still full, we wait until a new key starting with 0 appears. Then d is again increased by 1 and we keep only keys starting with 00). The scanning on the set is resumed and only distinct keys starting with 0 are considered. More generally, at depth d, only distinct keys starting with 0^d are taken into account. When we have exhausted the set of N data, n can be estimated by $R2^D$, where R is the random final bucket (cache) size and D is the final depth at the end of the process (total execution number of the process). R is the number of (all distinct) keys in the final cache and is immediately computed. We can summarize the algorithm with the following pseudo code

Algorithm 1

Parameter: bucket (or cache) B of at most b distinct keys.
Input: a stream $S = (s_1, s_2, \ldots, s_N)$
Output: the final bucket size R and the final depth D
Initialization: $B := \emptyset$ and $d := 0$
for all $x \in S$ **do**
  if $h(x) = 0^d \ldots$ **then**
    if $x \notin B$ **then**
      $B := B \cup x$
    end if;
  end if;
  if $|B| > b$ (overflow of cache) **then**
    $d := d + 1$
    filter (B) (remove keys which hash value doesn't match $0^d \ldots$)
  end if;
end for;
$D := d$
return R, D;

AS has some advantages in terms of processing time and of conceptual simplicity. As mentioned in [7] (see also the paper by Astrahan et al. [3]), AS outperforms standard sorting methods by a factor of about 8. In terms of storage consumptions, using 100 words of memory will provide for a typical accuracy of 12%. This is to be contrasted again with sorting, where the auxiliary memory required has to be at least as large as the file itself. Also an exact algorithm using a hash table will need a huge auxiliary memory. Finally AS is an unbiased estimator of cardinalities of large files that necessitates minimal auxiliary storage and processes data in a single pass.

In a paper by Gibbons [9] we are introduced to the Distinct Sampling approach for distinct value queries and reports over streams with known error bounds. This approach is based on adaptive selection of a sample during one single pass through the data and is very similar conceptually to AS. This sample is then used to estimate key queries such as "count distinct" or *how many distinct values satisfy a given predicate?*, and "Event Reports" or pre-scheduled, hard coded queries.

In fact [9] shows experimental results which are more than 5 times more accurate than previous work and 2-4 orders of magnitude faster. This work is currently being considered for improved implementation of the widely used open source Postgres SQL database. See [22].

Several new interesting questions can be asked about AS (some of them were suggested by P. Flajolet and popularized by J. Lumbroso). The distribution of $W = \log(R2^D/n)$ is known (see [14]), but in Sec.3, we rederive this distribution in a simpler way. In Sec.4 we provide new results on the moments of D and W. The final cache size R distribution is analyzed in Sec.5. Colored keys are considered in Sec.6: assume that we have a set of colors and that each key has some color. Assume also that among the n distinct keys, n_C do have color C and that n_C is large such that $\frac{n_C}{n} = p = \Theta(1)$. We show how to estimate p. We consider keys with some multiplicity in Sec.7: assume that, to each key κ_i, we attach a counter giving its *observed* multiplicity μ_i. Also we assume that the multiplicities

of color C keys are given by iid random variables (RV), with distribution function F_C, mean μ_C, variance σ_C^2 (functions of C). We show how to estimate μ_C and σ_C^2 . Sec.8 deals with the case where neither colors nor multiplicities are known. We want to estimate keys color, their multiplicities and their number. An appendix is devoted to the case where the hashing function provides bits with probability different from 1/2.

2 Preliminaries.

Let us first give the main notations we will use throughout the paper. Other particular notations will be provided where it is needed.

$$
\begin{aligned}
N &:= \text{ } \textit{total} \text{ number of keys, } N \text{ large,}\\
n &:= \text{number of } \textit{distinct} \text{ keys, } n \text{ large,}\\
\sim &:= \text{ asymptotic to, for large } n,\\
b &:= \text{ cache size, } b \text{ fixed, independent of } n,\\
\overset{b}{\sim} &:= \text{ asymptotic to, for large } n \text{ and } b,\\
R &:= \text{ number of keys in the cache, at the end of the process,}\\
D &:= \text{ depth of the cache, at the end of the process,}\\
Z &:= \frac{R2^D}{n},\\
\lg &:= \log_2,\\
W &:= \lg(Z),
\end{aligned}
$$

Flajolet gives the exact distribution in [7]

$$p(r,d) := \mathbb{P}(R=r, D=d) = \binom{n}{r}\Big(\frac{1}{2^d}\Big)^r\Big(1-\frac{1}{2^d}\Big)^{n-r}\Big[1-\sum_{k=0}^{b-r}\binom{n-r}{k}\Big(\frac{1}{2^d}\Big)^k\Big(1-\frac{1}{2^d}\Big)^{n-r-k}\Big], \quad (1)$$

$$p(.,d) := \mathbb{P}(D=d) = \sum_{r=0}^{b} p(r,d),$$

$$p(r,.) := \mathbb{P}(R=r) = \sum_{d} p(r,d),$$

$$P(r,d) := \mathbb{P}(R=r, D\le d).$$

The sample of R elements at the end of the execution is random as the hashed keys are $i.i.d$ random variables: AS produces random samples.

We can now see Adaptive Sampling as an urn model, where balls (keys), are thrown into urn $D=d$ with probability $1/2^d$. We recall the main properties of such a model.

- Asymptotic independence. We have asymptotic independence of urns, for all events related to urn d (d large) containing $\mathcal{O}(1)$ balls. This is proved, by Poissonization-De-Poissonization, in [10], [17] and [16]. This technique can be biefly described as follows. First we construct the corresponding generating function. Then we Poissonize (see, for instance, the paper by Jacquet and Szpankowski [11] for a general survey): instead of using a *fixed* number of balls, we use N balls, where N is a Poisson random variable. It follows that the urns become *independent* and the number of balls in urn ℓ is a Poisson random variable. We turn then to complex variables, and with Cauchy's integral theorem, we De-Poissonize the generating function, using [11, Thm.10.3 and Cor.10.17]. The error term is $\mathcal{O}(n^{-\gamma})$ where γ is a positive constant.

- Asymptotic distributions. We obtain asymptotic distributions of the interesting random variables as follows. The number of balls in each urn is asymptotically Poisson-distributed with parameter $n/2^d$ in urn d containing $\mathcal{O}(1)$ balls (this is the classical asymptotic for the Binomial distribution). This means that the asymptotic number ℓ of balls in urn d is given by

$$\exp\left(-n/2^d\right)\frac{\left(n/2^d\right)^\ell}{\ell!},$$

and with $\eta = d - \lg n, L := \ln 2$, this is equivalent to a Poisson distribution with parameter $e^{-L\eta}$. The asymptotic distributions are related to Gumbel distribution functions (given by $\exp\left(-e^{-x}\right)$) or convergent series of such. The error term is $\mathcal{O}(n^{-1})$.

- Uniform Integrability. We have uniform integrability for the moments of our random variables. To show that the limiting moments are equivalent to the moments of the limiting distributions, we need a suitable rate of convergence. This is related to a uniform integrability condition (see Loève's book [13, Section 11.4]). For Adaptive Sampling, the rate of convergence is analyzed in detail in [15]. The error term is $\mathcal{O}(n^{-\gamma})$.

- Mellin transform. Asymptotic expressions for the moments are obtained by Mellin transforms (for a good reference to Mellin transforms, see the paper by Flajolet et al. [8]). The error term is $\mathcal{O}(n^{-\gamma})$. We proceed as follows (see [15] for detailed proofs): from the asymptotic properties of the urns, we have obtained the asymptotic distributions of our random variables of interest. Next we compute the Laplace transform $\phi(\alpha)$ of these distributions, from which we can derive the dominant part of probabilities and moments as well as the (tiny) periodic part in the form of a Fourier series. This connection will be detailed in the next sections.

- Fast decrease property. The gamma function $\Gamma(s)$ decreases exponentially in the direction $i\infty$:

$$|\Gamma(\sigma + \mathbf{i}t)| \sim \sqrt{2\pi}|t|^{\sigma-1/2}e^{-\pi|t|/2}.$$

Also, we this property is true for all other functions we encounter. So inverting the Mellin transforms is easily justified.

- Early approximations. If we compare the approach in this paper with other ones that appeared previously, then we can notice the following. Traditionally, one would stay with exact enumerations as long as possible, and only at a late stage move to asymptotics. Doing this, one would, in terms of asymptotics, carry many unimportant contributions around, which makes the computations quite heavy, especially when it comes to higher moments. Here, however, approximations are carried out as early as possible, and this allows for streamlined (and often automatic) computations of the higher moments.

We set $\eta = d - \lg n$, (1) leads to

$$p(r,d) \sim f(r,\eta) = \exp(-2^{-\eta})\frac{2^{-r\eta}}{r!}\left[1 - \exp(-2^{-\eta})\sum_{k=0}^{b-r}\frac{2^{-k\eta}}{k!}\right], \tag{2}$$

and similar functions for $P(r,d)$. Asymptotically, the distribution will be a periodic function of the fractional part of $\lg n$. The distribution $P(r,d)$ does not converge in the weak sense, it does however converge along subsequences n_m for which the fractional part of $\lg n_m$ is constant. This type of convergence is not uncommon in the Analysis of Algorithms. Many examples are given in [15].

From (2), we compute the Laplace transform, with $\tilde{\alpha} := \alpha/L$:

$$\phi(r,\alpha) = \int_{-\infty}^{\infty} e^{\alpha\eta} f(r,\eta)d\eta = \frac{\Gamma(r-\tilde{\alpha})}{Lr!} - \sum_{k=0}^{b-r}\frac{\Gamma(r+k-\tilde{\alpha})}{Lr!k!2^{r+k-\tilde{\alpha}}}.$$

The k-th moments of Z are already given in [14] and [15]. As shown in [14], we must have $k \leq b$. For the sake of completeness, we repeat them here, with $\chi_l := \frac{2l\pi\mathbf{i}}{L}$, $\left\{ {k \atop i} \right\}$ denoting the Stirling number of the second kind, and $\mathbb{V}(X)$ denoting the Variance of the random variable X:

$$\begin{aligned}
&\mathbb{E}[Z^k] \sim m_{1,k} + w_{1,k},\\
&m_{1,k} = 1 + \frac{(b-k)!}{L}\sum_{i=1}^{k-1}\left\{ {k \atop i} \right\}\frac{2^{k-i}-1}{(k-i)(b-i)!},\\
&w_{1,k} = \sum_{l\neq 0}\frac{1}{L}\sum_{j=1}^{k-1}\left\{ {k \atop j} \right\}\left[(1-2^{k-j}\right]\Gamma(j-k+\chi_l)\binom{b-k+\chi_l}{b-j}e^{-2l\pi\mathbf{i}\lg n},\\
&m_{1,1} = 1, w_{1,1} = 0, m_{1,2} = 1 + \frac{1}{(b-1)L}, \mathbb{V}(Z) \sim \frac{1}{(b-1)L}.
\end{aligned} \tag{3}$$

$w_{i,j}$ will always denote a periodic function of $\lg n$ of small amplitude. Note that, in [7], Flajolet already computed $m_{1,1}, m_{1,2}, w_{1,1}$ and $w_{1,2}$.

3 Asymptotic distribution of $W = D - \lg n + \lg R$

W corresponds to the bit size of Z and has some independent interest. Let us recover this distribution from (2). In the sequel, we will denote $\mathbb{E}(Y|A)P(A)$ by $\mathbb{E}(Y;A)$, with Y either a Boolean event or a random variable. We have the following theorem

Theorem 3.1 *The asymptotic distribution of $W = D - \lg n + \lg R$, with $R > 0$ is given by*

$$\mathbb{P}(W \leq \alpha, R > 0) \sim \sum_{r=1}^{b}\sum_{l\geq 0}\exp(-2^{-\varphi})\frac{2^{-r\varphi}}{r!}\left[1 - \exp(-2^{-\varphi})\sum_{k=0}^{b-r}\frac{2^{-k\varphi}}{k!}\right],$$

where

$$\begin{aligned}
\{x\} &:= \textit{ fractional part of } x,\\
\varphi &:= \lfloor\{\lg n\} - \lg r + \alpha\rfloor - \{\lg n\} - \ell.
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{P}(W \leq \alpha, R > 0) &= \mathbb{P}[D \leq \lg n - \lg R + \alpha, R > 0]\\
&= \mathbb{P}[D \leq \lfloor\lg n\rfloor + \lfloor\{\lg n\} - \lg R + \alpha\rfloor, R > 0]\\
&\sim \sum_{r=1}^{b}\sum_{\ell\geq 0}\exp(-2^{-(\eta-\ell)})\frac{2^{-r(\eta-\ell)}}{r!}\left[1 - \exp(-2^{-(\eta-\ell)})\sum_{k=0}^{b-r}\frac{2^{-k(\eta-\ell)}}{k!}\right],
\end{aligned}$$

with

$$\eta = \lfloor\{\lg n\} - \lg r + \alpha\rfloor - \{\lg n\},$$

or

$$\mathbb{P}(W \leq \alpha, R > 0) \sim \sum_{r=1}^{b}\sum_{l\geq 0}\exp(-2^{-\varphi})\frac{2^{-r\varphi}}{r!}\left[1 - \exp(-2^{-\varphi})\sum_{k=0}^{b-r}\frac{2^{-k\varphi}}{k!}\right],$$

with

$$\varphi := \lfloor\{\lg n\} - \lg r + \alpha\rfloor - \{\lg n\} - \ell.$$

This is exactly Theorem 4.1 in [14] that we obtain here in a simpler way. ■

4 Moments of $D - \lg n$ and W

Recall that D is the final depth (number of times the cache is filtered = number of times the cache overflows). Two interesting parameters are given by the moments of $D-\lg n$ and W. Their asymptotic behaviour is given as follows, with $\psi(x)$ denoting the digamma function (the logarithmic derivative of $\Gamma(x)$)

Theorem 4.1 *The moments of $D - \lg n$ and W are asymptotically given by*

$$\mathbb{E}[(D - \lg n)^k; R = r] \sim \tilde{m}_{k,r} + \tilde{w}_{k,r},$$

where

$$\tilde{m}_{k,r} := \phi^{(k)}(r, 0),$$
$$w_{k,r} = \sum_{l\neq 0} \phi^{(k)}(r, \alpha)\Big|_{\alpha=-L\chi_l} e^{-2l\pi \mathbf{i} \lg n}.$$

For instance

$$\tilde{m}_{1,r} = -\frac{\psi(r)}{L^2 r} + \sum_{k=0}^{b-r} \frac{(\psi(r+k) - L)2^{-(r+k)}\Gamma(r+k)}{L^2\Gamma(r+1)\Gamma(k+1)}, \quad r > 0,$$
$$\tilde{m}_{1,0} = \frac{1}{2} + \frac{\gamma}{L} + \sum_{k=1}^{b} \frac{(\psi(k) - L)2^{-k}}{kL^2},$$
$$\tilde{w}_{1,r} = \sum_{l\neq 0} \left[-\frac{\psi(r+\chi_l)\Gamma(r+\chi_l)}{L^2\Gamma(r+1)} + \sum_{k=0}^{b-r} \frac{(\psi(r+k+\chi_l) - L)2^{-(r+k)}\Gamma(r+k+\chi_l)}{L^2\Gamma(r+1)\Gamma(k+1)} \right] e^{-2l\pi \mathbf{i} \lg n}, \quad r > 0,$$
$$\tilde{w}_{1,0} = \sum_{l\neq 0} \left[-\frac{\psi(\chi_l)\Gamma(\chi_l)}{L^2} + \sum_{k=0}^{b} \frac{\Gamma(k+\chi_l)}{L^2 k! 2^k} (\psi(k+\chi_l) - L) \right].$$

$$\mathbb{E}(W; R > 0) \sim \sum_{r=1}^{b} \tilde{m}_{1,r} + \sum_{r=1}^{b} p(r, .) \lg r + \sum_{r=1}^{b} \tilde{w}_{1,r},$$
$$\mathbb{E}(W^2; R > 0) \sim \sum_{r=1}^{b} \tilde{m}_{2,r} + 2\sum_{r=1}^{b} \tilde{m}_{1,r} \lg r + \sum_{r=1}^{b} p(r, .)(\lg r)^2 + \sum_{r=1}^{b} \tilde{w}_{2,r} + 2\sum_{r=1}^{b} \tilde{w}_{1,r} \lg r.$$

Proof. Using the techniques developed in [15], we obtain the dominant (constant) part of the moments of D as follows:

$$\mathbb{E}[(D - \lg n)^k; R = r] \sim \tilde{m}_{k,r} + w_{k,r},$$

where the non-periodic component is given by

$$\tilde{m}_{k,r} := \phi^{(k)}(r, 0),$$

and the corresponding periodic term is given by

$$w_{k,r} = \sum_{l\neq 0} \phi^{(k)}(r, \alpha)\Big|_{\alpha=-L\chi_l} e^{-2l\pi \mathbf{i} \lg n}.$$

This was already computed in [15], but with some errors. The first corrected values are now provided.

As $W = D - \lg n + \lg R$, the rest of the proof is immediate ■

It will be useful to obtain an asymptotic for the expectation of $D - \lg n$ (non-periodic component) for large b. This is computed as follows. First of all, we rewrite $\sum_{r=1}^{b} \tilde{m}_{1,r}$ as

$$\sum_{r=1}^{b} \tilde{m}_{1,r} = -\sum_{r=1}^{b} \frac{\psi(r)}{L^2 r} + \sum_{u=1}^{b} \left[\sum_{r=1}^{u} \frac{1}{\Gamma(r+1)\Gamma(u-r+1)} \right] \frac{(\psi(u) - L)2^{-u}\Gamma(u)}{L^2}.$$

Now it is clear that the main contribution of the second term is related to large u. So we set $r = \frac{u}{2} + v$. This gives, by Stirling,

$$\Gamma(r+1) \sim e^{-(u/2+v)} e^{v+v^2/u} \left(\frac{u}{2}\right)^{u/2+v} \sqrt{\pi u},$$

and

$$\Gamma(r+1)\Gamma(u-r+1) \sim e^{-u} e^{2v^2/u} \left(\frac{u}{2}\right)^{u} \pi u.$$

By Euler-Maclaurin, we have

$$\sum_{r=1}^{u} \frac{1}{\Gamma(r+1)\Gamma(u-r+1)} \sim 2 \sum_{v=0}^{u/2} \frac{e^u}{\left(\frac{u}{2}\right)^u \pi u} e^{-2v^2/u} \sim \int_0^\infty 2e^{-2v^2/u} dv \frac{e^u}{\left(\frac{u}{2}\right)^u \pi u} = \frac{e^u}{\left(\frac{u}{2}\right)^u \sqrt{2\pi u}},$$

and, finally,

$$\sum_{r=0}^{b} \tilde{m}_{1,r} \overset{b}{\sim} \frac{1}{2} + \frac{\gamma}{L} + \sum_{u=1}^{b} \left[-\frac{\psi(u)}{L^2 u} + \frac{(\psi(u) - L)2^{-u}}{uL^2} + \frac{(\psi(u) - L)2^{-u}\Gamma(u)}{L^2} \frac{e^u}{\left(\frac{u}{2}\right)^u \sqrt{2\pi u}} \right],$$

and, to first order,

$$\mathbb{E}(D - \lg n) \sim \sum_{r=0}^{b} \tilde{m}_{1,r} \overset{b}{\sim} \frac{1}{2} + \frac{\gamma}{L} - \sum_{u=1}^{b} \frac{1}{Lu} \sim -\lg b + \mathcal{O}(1) \tag{4}$$

The expected total time cost of the algorithm, $\mathbf{C}_n$, is given by

$$\mathbb{E}[\mathbf{C}_n] = n\mathcal{O}(\lg b) + \mathbb{E}[D]\mathcal{O}(b) = n\mathcal{O}(\lg b) + \lg n \mathcal{O}(b),$$

where $\mathcal{O}(\lg b)$ is the update cost of the cache for each key (we assume an efficient implementation of the cache, for instance a binary search tree) and $\mathcal{O}(b)$ is the update cost of the cache at each process execution.

5 Distribution of R

The asymptotic moments and distribution of the cache size R are given as follows

Theorem 5.1 *The non-periodic components of the asymptotic moments and distribution of R are given by*

$$\begin{aligned}
\mathbb{E}(R) &\sim \frac{b}{2L}, \\
\mathbb{E}(R^2) &\sim \frac{b(3b+1)}{8L}, \\
\mathbb{V}(R) &\sim \frac{b(3Lb - 2b + L)}{8L^2}, \\
\mathbb{P}(R = r) = p(r, .) &\sim \frac{1}{L} \left[\frac{1}{r} - \sum_{k=0}^{b-r} \frac{\Gamma(r+k)}{r!k!2^{r+k}} \right], r \geq 1,
\end{aligned}$$

Similarly, the periodic components are given by

$$w_1(R) = \sum_{r=0}^{b} r \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \lg n},$$
$$w_2(R) = \sum_{r=0}^{b} r^2 \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \lg n},$$
$$w_0(r) = \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \lg n}$$

Proof. We have

$$\mathbb{P}(R = r) = p(r, .) \sim \phi(r, 0) = \frac{1}{L}\left[\frac{1}{r} - \sum_{k=0}^{b-r} \frac{\Gamma(r+k)}{r!k!2^{r+k}}\right], r \geq 1,$$

and $p(0, .) = 1 - \sum_1^b p(r, .)$, with

$$\sum_1^b p(r, .) \sim \frac{1}{L}\left[H_b - \sum_{r=1}^{b} \sum_{k=0}^{b-r} \frac{\Gamma(r+k)}{r!k!2^{r+k}}\right]$$
$$= \frac{1}{L}\left[H_b - \sum_{u=1}^{b} \frac{(u-1)!}{2^u} \sum_{r=1}^{u} \frac{1}{r!(u-r)!}\right],$$

Where H_n denotes the $n-th$ harmonic number. This quantity was already obtained in [14] after some complicated algebra! This leads to

$$p(0, .) \sim 1 - \sum_{u=1}^{b} \frac{1}{u2^u L},$$

which is also the probability of $Z = 0$. This is also easily obtained from $\lim_{r\to 0} \phi(r, 0)$. Figure 1 gives $p(r, .)$ for $b = 50$

The moments of R are computed as follows.

$$\mathbb{E}(R) = \sum_{r=1}^{b} r p(r, .) \sim \frac{1}{L}\left[b - \sum_{u=1}^{b} \frac{(u-1)!}{2^u} \sum_{r=1}^{u} \frac{r}{r!(u-r)!}\right]$$
$$= \frac{1}{L}\left[b - \sum_{u=1}^{b} \frac{1}{2^u}[2^{u-1}]\right]$$
$$= \frac{b}{2L}.$$

More generally, the generating function of $p(r, .)$ is given by

$$\sum_{r=1}^{b} z^r p(r, .) \sim \frac{1}{L}\left[\sum_{r=1}^{b} \frac{z^r}{r} - \sum_{u=1}^{b} \frac{(u-1)!}{2^u} \sum_{r=1}^{u} \frac{z^r}{r!(u-r)!}\right]$$
$$= \frac{1}{L}\left[\sum_{r=1}^{b} \frac{z^r}{r} - \sum_{u=1}^{b} \frac{1}{u2^u}[(1+z)^u - 1]\right].$$

This leads to

$$\mathbb{E}(R^2) \sim \sum_{r=1}^{b} r^2 p(r, .) = \frac{b(3b+1)}{8L},$$

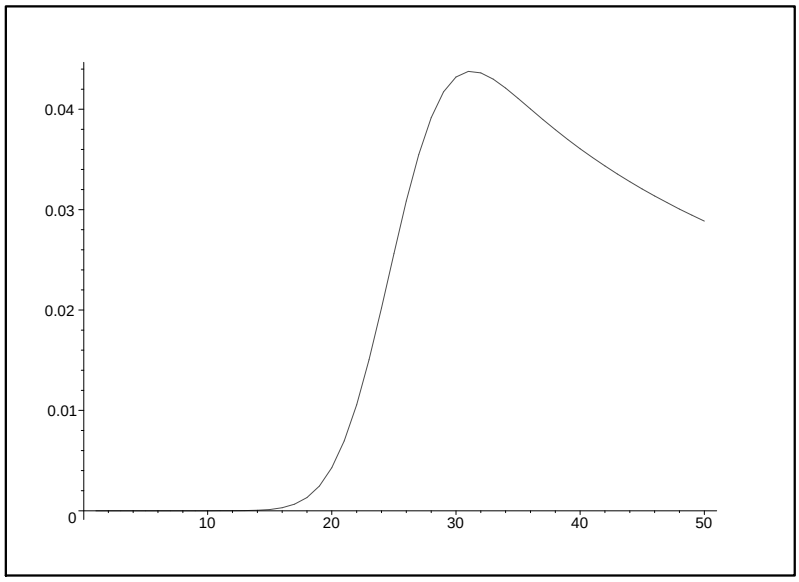

Figure 1: $p(r,.)$ for $b = 50$

$$\mathbb{V}(R) \sim \frac{b(3Lb - 2b + L)}{8L^2}.$$

Similarly, the periodic components are given by

$$w_1(R) = \sum_{r=0}^{b} r \sum_{l \neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \lg n},$$

$$w_2(R) = \sum_{r=0}^{b} r^2 \sum_{l \neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \lg n}.$$

■

6 Colors

Seasonal, or temporal context has become increasingly important in data mining [6]. For example its often important to be able to group events by the time which they occur or to understand event periodicity. We can represent different temporal contexts with colors. This motivates our analysis of colored keys.

Assume that we have a set of colors and that each key has some color. Let us give a simple example. We might be interested in knowing, for instance, the proportion p of elements whose multiplicity is below some constant M. So we say that such elements are of color "white", while the rest are of color "black". We attach to each *distinct* key κ_ℓ a counter ν_ℓ giving the number of times (multiplicity) this key appears in the sample. At the end of AS, we have in the cache U_W white keys. AS produces random samples and it can gather exact counts of the frequency of the sampled elements (since a sampled element enters the sample in its very first occurrence and if an element is kicked out from the sample, it will never be sampled again).

This leads to the unbiased estimates $\tilde{n}_W = 2^D U_W, \tilde{n} = 2^D R, \tilde{p}_W = U_W/R$.

We note that the observed multiplicities will be used with more detail in Sec. 7, 8.

Another example is as follows. A situation naturally maps to real life situations when data is "strongly seasonal". Such is the case when noisy "seasonal" differences in data is to be naturally expected, and therefore to be ignored when performing analytics of a data stream. For example, even though the number of viewers on two consecutive baseball games might show the second game receiving roughly half as many views as the first, it makes little sense to conclude that the baseball team is becoming less popular if the first was a night game and the second a weekend day game. See the report by Wong et al. [24] for an example of outlier detection where seasonality must be ignored.

Assume now that among the n distinct keys, n_C do have color C and that n_C is large such that $\frac{n_C}{n} = p = \Theta(1), q := 1 - p$. In the cache, the R keys (we assume $R > 0$) contain U keys with color C with probability distribution

$$\mathbb{P}(U = u | R = r) = \frac{\binom{n_C}{u}\binom{n-n_C}{r-u}}{\binom{n}{r}},$$

and, if $r = o(n)$, this is asymptotically given by the conditioned binomial distribution $Bin(r, p)$. We want to estimate p. We are interested in the distribution of the statistic $\tilde{p} = U/R$. We have

Theorem 6.1 *The asymptotic moments of the statistic $\tilde{p} = U/R$ are given by*

$$\begin{aligned} \mathbb{E}\left(\frac{U}{R}; R > 0\right) &\sim p, \\ \mathbb{E}\left(\left(\frac{U}{R}\right)^2; R > 0\right) &\sim p^2 + pq\mathbb{E}\left(\frac{1}{R}; R > 0\right), \\ \mathbb{V}\left(\frac{U}{R}; R > 0\right) &\sim pq\mathbb{E}\left(\frac{1}{R}; R > 0\right). \end{aligned} \tag{5}$$

Proof. We have

$$\mathbb{P}\left[\left(\frac{U}{R}\right) \leq \alpha; R > 0\right] = \mathbb{P}(U \leq \alpha R; R > 0] \sim \sum_{r=1}^{b} p(r, .) \sum_{u=0}^{\lfloor \alpha r \rfloor} \binom{r}{u} p^u q^{r-u}. \tag{6}$$

Now, conditioned on R, we have

$$\mathbb{E}(U|R) \sim Rp, \mathbb{E}(U^2|R) \sim Rpq + R^2p^2.$$

So, conditioned on R,

$$\begin{aligned} \mathbb{E}\left(\frac{U}{R}\right) &\sim p, \\ \mathbb{E}\left(\left(\frac{U}{R}\right)^2\right) &\sim p^2 + \frac{pq}{R}, \end{aligned}$$

and, unconditionning leads to the theorem. ■

Intuitively, if the cache size b is large, we should have an asymptotic Gaussian distribution for U/R. Actually, the fit is quite good, even for $b = 30$ (and $p = 0.2$).

This is proved in the next subsections.

6.1 The distribution of U/R for large b .

Let R be a (possibly degenerate) random variable taking values on the (strict) positive integers. Conditionning on R, let $U \sim Bin(R,p)$ for some known $0<p<1$, and set $Y = U/R$. It appears that, as R grows large, the distribution of Y becomes asymptotically Gaussian. This claim can be made precise as follows.

Theorem 6.2 *Let $V \sim \mathcal{N}(0,1)$ and write $Y^* = \sqrt{R}\frac{Y-p}{\sqrt{pq}}$. Then there exists an absolute constant $\kappa \in \mathbb{R}$ such that*

$$d_{\mathcal{W}}(Y^*, V) \leq \kappa \mathbb{E}\left\{\frac{1}{\sqrt{R}}\right\}$$

for $d_{\mathcal{W}}(Y^*,V)$ the Wasserstein distance between the law of Y^* and that of V; moreover this constant is such that

$$\kappa \leq \frac{q^2-p^2}{\sqrt{pq}} + 4\left[\frac{p^3+q^3}{pq} - 1\right]^{1/2}.$$

Proof. We will prove this theorem using the Stein methodology which, for $h \in \mathcal{H}$, ($\mathcal{H}$ is a nice class of test functions), suggests to write

$$\mathbb{E}h(Y^*) - \mathbb{E}h(V) = \mathbb{E}\left(Y^* f(Y^*) - f'(Y^*)\right)$$

with $f := f_h$ such that

$$xf(x) - f'(x) = h(x) - \mathbb{E}h(V). \tag{7}$$

(this is known as the *Stein equation* for the Gaussian distribution) so that

$$d_{\mathcal{W}}(Y^*,V) = \sup_{h\in\mathcal{H}} \mathbb{E}\left|h(Y^*) - h(V)\right| \leq \sup_{f_h} \left|\mathbb{E}\left(Y^* f(Y^*) - f'(Y^*)\right)\right|. \tag{8}$$

The reason why (8) is interesting is that properties of the solutions f of (7) are well-known – see, e.g., [4, Lemma 2.3] and [5, Lemma 2.3] – and quite good so that they can be used with quite some efficiency to tackle the rhs of (8). In the present configuration we know that f_h is continuous and bounded on $\mathbb{R}$, with

$$\|f_h'\| \leq \min\left(2\|h - \mathbb{E}h(V)\|, 4\|h'\|\right) \tag{9}$$

and

$$\|f_h''\| \leq 2\|h'\|. \tag{10}$$

In particular, if H is the class of Lipschitz-1 functions with $\|h'\| \leq 1$ (this class generates the Wasserstein distance) then

$$\|f_h'\| \leq 4 \text{ and } \|f_h''\| \leq 2.$$

These will suffice to our purpose.

Our proof follows closely the standard one for independent summands (see, e.g., [20, Section 3]). First we remark that, given $R \geq 1$, we can write Y^* as

$$Y^* = \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i$$

where, taking X_i i.i.d. $Bin(1,p)$, we let $\xi_i = (X_i - p)/\sqrt{pq}$ (which are centered and have variance 1). Next, for $r \geq 1$ and $1 \leq i \leq r$, define

$$Y_i^{*r} = Y^* - \frac{1}{\sqrt{r}}\xi_i = \frac{1}{\sqrt{r}} \sum_{j\neq i} \xi_j.$$

Next take f solution of (7) with h some Lipschitz-1 function. Then note that $\mathbb{E}\left\{\xi_i f\left(Y_i^{*r}\right)\right\} = 0$ for all $1 \leq i \leq r$. We abuse notations and, given R, write $Y_i^{*R} = Y_i^*$. Then

$$\begin{aligned}
\mathbb{E}\left\{Y^* f(Y^*) \,|\, R\right\} &= \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i f(Y^*) \,\middle|\, R \right\} \\
&= \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i \left(f(Y^*) - f(Y_i^*)\right) \,\middle|\, R \right\} \\
&= \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i \left(f(Y^*) - f(Y_i^*) - (Y^* - Y_i^*) f'(Y^*)\right) \,\middle|\, R \right\} \\
&\quad + \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i (Y^* - Y_i^*) f'(Y^*) \,\middle|\, R \right\}
\end{aligned}$$

so that

$$\begin{aligned}
\left|\mathbb{E}\left\{Y^* f(Y^*) - f'(Y^*) \,|\, R\right\}\right| &\leq \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \left|\xi_i \left(f(Y^*) - f(Y_i^*) - (Y^* - Y_i^*) f'(Y^*)\right)\right| \,\middle|\, R \right\} \\
&\quad + \left| \mathbb{E}\left\{ f'(Y^*) \left(1 - \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i (Y^* - Y_i^*)\right) \,\middle|\, R \right\} \right| \\
&=: |\chi_1(R)| + |\chi_2(R)|.
\end{aligned}$$

Recall that $Y^* - Y_i^* = \frac{1}{\sqrt{R}} \xi_i$. Then (by Taylor expansion) we can easily deal with the first term to obtain

$$|\chi_1(R)| = \frac{\|f''\|}{2} \frac{1}{\sqrt{R}} \mathbb{E}\left|\xi_1\right|^3.$$

Taking expectations with respect to R and using (10) we conclude

$$E|\chi_1(R)| \leq \mathbb{E}\left|\xi_1\right|^3 E\left(\frac{1}{\sqrt{R}}\right). \tag{11}$$

For the second term note how

$$\begin{aligned}
|\chi_2(R)| &= \left| \mathbb{E}\left\{ f'(Y^*) \left(1 - \frac{1}{R} \sum_{i=1}^{R} \xi_i^2\right) \,|\, R \right\} \right| \\
&= \left| \mathbb{E}\left\{ \frac{f'(Y^*)}{R} \sum_{i=1}^{R} \left(1 - \xi_i^2\right) \,\middle|\, R \right\} \right| \\
&\leq \frac{\|f'\|}{R} \mathbb{E}\left\{ \left| \sum_{i=1}^{R} \left(1 - \xi_i^2\right) \right| \,\middle|\, R \right\}.
\end{aligned}$$

Since $\mathbb{E}\left\{\sum_{i=1}^{R} \left(1 - \xi_i^2\right) \,|\, R\right\} = 0$ we can pursue to obtain

$$\begin{aligned}
|\chi_2(R)| &\leq \frac{\|f'\|}{R} \sqrt{\mathbb{V}\left(\sum_{i=1}^{R} \left(1 - \xi_i^2\right) \,\middle|\, R \right)} \\
&= \frac{\|f'\|}{\sqrt{R}} \sqrt{\mathbb{V}\left(\xi_1^2\right)}
\end{aligned}$$

where we used (conditional) independence of the ξ_i. Taking expectations with respect to R and using (9) we deduce (recall $\mathbb{V}\left(\xi_1^2\right) = \mathbb{E}\xi_1^4 - 1$)

$$\mathbb{E}|\chi_2(R)| \leq 4\sqrt{\mathbb{E}\xi_1^4 - 1}\mathbb{E}\left(\frac{1}{\sqrt{R}}\right). \tag{12}$$

Combining (11) and (12) we can conclude

$$d_{\mathcal{W}}(Y^*, V) \leq \left(\mathbb{E}|\xi_1^3| + 4\sqrt{\mathbb{E}\xi_1^4 - 1}\right)\mathbb{E}\left(\frac{1}{\sqrt{R}}\right).$$

The claim follows. ■

So we need the moments of $1/R$ for large b (we limit ourselves to the dominant term).

6.2 Moments of $1/R, R > 0$ for large b

We have the following property

Theorem 6.3 *The asymptotic moments of $1/R, R > 0$ for large b, with $R > 0$ are given by*

$$\mathbb{E}\left(\frac{1}{R^\alpha}; R > 0\right) \overset{b}{\sim} \frac{1}{L\alpha b^\alpha}(2^\alpha - 1), \alpha > 0.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{R}; R > 0\right) &= \sum_{r=1}^{b} p(r, .)/r \sim \frac{1}{L}\left[\sum_{r=1}^{b}\frac{1}{r^2} - \sum_{u=1}^{b}\frac{(u-1)!}{2^u}\sum_{r=1}^{u}\frac{1}{rr!(u-r)!}\right] \\
&= \frac{1}{L}\left[H_b^{(2)} - \sum_{u=1}^{b}\frac{1}{2^u}{}_2F_2[[1, -u+1]; [2, 2]; -1]\right] \\
&= \frac{1}{L}\left[-\psi(1, b+1) + \frac{\pi^2}{6} - \sum_{u=1}^{\infty}\frac{1}{2^u}{}_2F_2[[1, -u+1], [2, 2], -1]\right] \\
&+ \frac{1}{L}\sum_{u=b+1}^{\infty}\frac{1}{2^u}{}_2F_2[[1, -u+1], [2, 2], -1].
\end{aligned}$$

where $\psi(n, x)$ is the nth polygamma function, that is the nth derivative of the digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$ and ${}_2F_2$ is the hypergeometric function.

But[1]

$$\begin{aligned}
&\sum_{u=1}^{\infty}\frac{1}{2^u}{}_2F_2[[1, -u+1], [2, 2], -1] \\
&= \sum_{r=1}^{\infty}\sum_{u=r}^{\infty}\frac{(u-1)!}{2^u rr!(u-r)!} \\
&= \sum_{r=1}^{\infty}\frac{1}{r^2}\sum_{u=r}^{\infty}\frac{(u-1)!}{2^u(r-1)!(u-r)!} \\
&= \sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{u=v+1}^{\infty}\frac{(u-1)!}{2^u v!(u-v-1)!} \\
&= \sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{w=v}^{\infty}\frac{w!}{2^{w+1}v!(w-v)!}
\end{aligned}$$

[1]We are indebted to H.Prodinger for this identity

$$
\begin{aligned}
&= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{w=v}^{\infty}\binom{w}{v}2^{-w}\\
&= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{s=0}^{\infty}\binom{s+v}{v}2^{-(s+v)}\\
&= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}2^{-v}\sum_{s=0}^{\infty}\binom{-v-1}{s}(-2)^{-s}\\
&= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}2^{-v}\left(1-\frac{1}{2}\right)^{-(v+1)}\\
&= \sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\\
&= \zeta(2) = \frac{\pi^2}{6}. \qquad (13)
\end{aligned}
$$

Now

$$\psi(1,b+1) \overset{b}{\sim} \frac{1}{b} + \mathcal{O}\left(\frac{1}{b^2}\right),$$

and

$$\sum_{u=b+1}^{\infty}\frac{1}{2^u}{}_2F_2[[1,-u+1],[2,2],-1] = T_1 + T_2,$$

where

$$
\begin{aligned}
T_1 &= \sum_{r=1}^{b+1}\sum_{u=b+1}^{\infty}\frac{(u-1)!}{2^u r r!(u-r)!}\\
&= \frac{1}{2}\sum_{v=0}^{b}\frac{1}{(v+1)^2}\sum_{w=b}^{\infty}\binom{w}{v}2^{-w},\\
T_2 &= \sum_{r=b+1}^{\infty}\sum_{u=r}^{\infty}\frac{(u-1)!}{2^u r r!(u-r)!}\\
&= \frac{1}{2}\sum_{v=b}^{\infty}\frac{1}{(v+1)^2}\sum_{w=v}^{\infty}\binom{w}{v}2^{-w}\\
&= \frac{1}{2}\sum_{v=b}^{\infty}\frac{1}{(v+1)^2}2.
\end{aligned}
$$

In order to compute T_1, we now turn to the asymptotics of $\binom{w}{v}2^{-w}$ for large w. We obtain, by Stirling and setting $w = 2v+\alpha$,

$$
\begin{aligned}
\binom{w}{v}2^{-w} &\sim \frac{e^{-w}w^w\sqrt{2\pi w}}{e^{-(w-v)}(w-v)^{w-v}\sqrt{2\pi(w-v)}2^w e^{-v}v^v\sqrt{2\pi v}}\\
&= \frac{e^{-(2v+\alpha)}(2v+\alpha)^{2v+\alpha}\sqrt{2\pi(2v+\alpha)}}{e^{-(v+\alpha)}(v+\alpha)^{v+\alpha}\sqrt{2\pi(v+\alpha)}2^{2v+\alpha}e^{-v}v^v\sqrt{2\pi v}}\\
&\sim \frac{e^{-v}(2v)^{2v+\alpha}\left(1+\frac{\alpha}{2v}\right)^{2v+\alpha}\sqrt{2}}{v^{v+\alpha}\left(1+\frac{\alpha}{v}\right)^{v+\alpha}2^{2v+\alpha}e^{-v}v^v\sqrt{2\pi v}}\\
&\sim \frac{e^{\alpha+\frac{\alpha^2}{4v}}\sqrt{2}}{e^{\alpha+\frac{\alpha^2}{2v}}\sqrt{2\pi v}}
\end{aligned}
$$

$$\sim \frac{e^{-\frac{\alpha^2}{4v}}}{\sqrt{\pi v}}$$
$$= 2\frac{e^{-\frac{\alpha^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}},$$

with $\sigma^2 = 2v$. This is a Gaussian function, centered at $2v$ with variance $\sigma^2 = 2v$. So, by Euler-Maclaurin, replacing sums by integrals, we obtain

- if $b/2 < v \leq b$,
$$\sum_{w=b}^{\infty} \binom{w}{v} 2^{-w} \overset{b}{\sim} 2,$$
- if $0 \leq v < b/2$,
$$\sum_{w=b}^{\infty} \binom{w}{v} 2^{-w} \text{ is exponentially negligible ,}$$
- if $v \geq b$,
$$\sum_{w=b}^{\infty} \binom{w}{v} 2^{-w} = 2 \text{ by (13),}$$
but this will not be used in the sequel,

and finally

$$T_1 + T_2 \overset{b}{\sim} \frac{1}{2}\left[2\sum_{v=b/2}^{b} \frac{1}{(v+1)^2} + 2\sum_{v=b}^{\infty} \frac{1}{(v+1)^2}\right] \overset{b}{\sim} \frac{2}{b}.$$

This leads to

$$\mathbb{E}\left(\frac{1}{R}; R > 0\right) \overset{b}{\sim} \frac{1}{L}\left[\frac{2}{b} - \frac{1}{b}\right] = \frac{1}{Lb}. \tag{14}$$

In the neighbourhood of $v = b/2$, only part of the Gaussian is integrated. But if we choose an interval $\Delta := [b/2 - b^{5/8}, b/2 + b^{5/8}]$, $(b^{5/8} \gg \sigma)$, this contributes to

$$\mathcal{O}\left(\int_{\Delta} \frac{1}{v^2} dv\right) = \mathcal{O}(b^{-5/8}) = o(1/b).$$

Similarly, we derive (we omit the details)

$$\mathbb{E}\left(\frac{1}{R^2}; R > 0\right) \overset{b}{\sim} \frac{3}{2Lb^2},$$
$$\mathbb{V}\left(\frac{1}{R^2}; R > 0\right) \overset{b}{\sim} \frac{1}{b^2}\left[\frac{3}{2L} - \frac{1}{L^2}\right],$$
$$\mathbb{E}\left(\frac{1}{R^{1/2}}; R > 0\right) \overset{b}{\sim} 2(\sqrt{2} - 1)/(L\sqrt{b}).$$

More generally,

$$\mathbb{E}\left(\frac{1}{R^{\alpha}}; R > 0\right) \overset{b}{\sim} \frac{1}{L\alpha b^{\alpha}}(2^{\alpha} - 1), \alpha > 0$$

■

Now we obtain, by (6) and Thm 6.2 the following Thm

Theorem 6.4 *The limiting distribution of U/R for large b is Gaussian.*

Note that, by (5) and (14), we obtain

$$\mathbb{E}\left(\left(\frac{U}{R}\right)^2; R>0\right) \overset{b}{\sim} p^2+\frac{pq}{Lb},$$
$$\mathbb{V}\left(\frac{U}{R}; R>0\right) \overset{b}{\sim} \frac{pq}{Lb}. \tag{15}$$

This provides a confidence interval for p. With a confidence level of 5% for instance, we have

$$\left[\frac{U}{R}-2\sqrt{\frac{pq}{Lb}} \leq p \leq \frac{U}{R}+2\sqrt{\frac{pq}{Lb}}\right],$$

and, as we can estimate p by $\frac{U}{R}$, this leads to

$$\left[\frac{U}{R}-2\sqrt{\frac{\frac{U}{R}\left(1-\frac{U}{R}\right)}{Lb}} \leq p \leq \frac{U}{R}+2\sqrt{\frac{\frac{U}{R}\left(1-\frac{U}{R}\right)}{Lb}}\right].$$

6.3 Several Colors

If we are interested in the joint distribution of the statistic $U_1/R, \dots U_k/R$, which correspond to k different colors among the present colors, we have an asymptotic conditional multinomial distribution. For instance, for $k=2$, this leads to

$$\binom{r}{u_1, u_2, r-u_1-u_2} p_1^{u_1} p_2^{u_2} (1-p_1-p_2)^{r-u_1-u_2},$$

with mean rp_1, rp_2. So

$$\mathbb{E}\left(\frac{U_1}{R}; R>0\right)=p_1, \mathbb{E}\left(\frac{U_2}{R}; R>0\right)=p_2,$$

and we obtain similarly, conditioned on R

$$\mathbb{E}(U_1U_2)=R(R-1)p_1p_2,$$
$$\mathbb{E}\left(\left(\frac{U_1}{R}\right)\left(\frac{U_2}{R}\right)\right)=p_1p_2-\frac{p_1p_2}{R},$$

and, unconditionning,

$$\mathbb{E}\left(\left(\frac{U_1}{R}\right)\left(\frac{U_2}{R}\right); R>0\right) \overset{b}{\sim} p_1p_2-\frac{p_1p_2}{Lb},$$

or

$$Cov\left(\frac{U_1}{R}, \frac{U_2}{R}; R>0\right) \overset{b}{\sim} -\frac{p_1p_2}{Lb}.$$

7 Multiplicities of colored keys

Counting the distinct number of keys is import for mining search engine data, see the paper by Kane et al [12]. It is important for search engines to correctly identify seasonal (colors) queries and make sure that their results are temporally appropriate. See the conference paper by Shokouhi, [21]. For example, queries with the key "Wimbledon" would take on a different color (season) throughout the year. In the winter we might return general information about the tennis tournament. In the spring perhaps logistics, travel and start date become more important. After the tournament starts the results of the matches become more relevant. During the championship match the local TV channel where to watch is the most relevant. One might be interested in estimating the sizes of these different seasons. How many tennis related queries do we expect to occur during the Wimbledon final?

Assume that the multiplicities of color C keys are given by iid random variables (RV), with distribution function F_C, mean μ_C, variance σ_C^2 (functions of C), all unknown. We want to estimate μ_C and σ_C^2. Of course, we can estimate μ_C by $\tilde{\mu}_C = N_C/n_C$ where N_C is the total number of observed color C keys and n_C is the number of *distinct* color C keys among the n *distinct* keys. n_C is classically estimated by $2^D U$ (recall that U is the number of color C keys among the R distinct keys in the cache). But the classical AS algorithm is not efficient enough to provide an estimate for σ_C^2. We proceed as follows: to each color C key κ_i, we attach a counter giving its *observed* multiplicity μ_i. From Section 6 (see(15)), we can estimate $p := n_C/n$ by $\tilde{p} = (U/R; R > 0)$. We have

$$\mathbb{E}(\tilde{p}; R > 0) \overset{b}{\sim} p,$$
$$\mathbb{V}(\tilde{p}; R > 0) \overset{b}{\sim} \frac{pq}{Lb}.$$

Also, we can estimate mean μ_C and variance σ_C^2 by $\tilde{\mu}_C$ and $\tilde{\sigma}_C^2$ as given by (the observed multiplicities are extracted in the cache at the end of AS)

$$\tilde{\mu}_C := \frac{V}{U}, \quad V := \sum_1^U \mu_i,$$
$$\tilde{\sigma}_C^2 := \frac{\sum_1^U (\mu_i - \tilde{\mu}_C)^2}{U}.$$

Next we estimate n by $\tilde{n} = R2^D$ (see Sec. 2). We have, *conditioned* on U,

Theorem 7.1 *The moments of $\tilde{\mu} = V/U$ are given by*

$$\mathbb{E}(\tilde{\mu}) = \mu,$$
$$\mathbb{E}(\tilde{\mu}^2) = \mu^2 + \sigma^2 \mathbb{E}\left(\frac{1}{U}\right),$$
$$\mathbb{V}(\tilde{\mu}) = \sigma^2 \mathbb{E}\left(\frac{1}{U}\right).$$

Proof. We only need

$$\mathbb{E}\left[\frac{V^2}{U^2}\middle| U\right] = \left[\frac{U\sigma^2 + U^2\mu^2}{U^2}\middle| U\right]$$

. ■

Now we estimate n_C by $\tilde{n}_C = \tilde{n}\tilde{p} = 2^D U$. But if we have two independent RV, X, Y, with mean and variance respectively $m_X, m_Y, \sigma_X^2, \sigma_Y^2$, it is easy to see that

$$\mathbb{E}(XY) = m_X m_Y, \tag{16}$$
$$\mathbb{V}(XY) = \sigma_X^2 m_Y^2 + \sigma_Y^2 m_X^2 + \sigma_X^2 \sigma_Y^2.$$

Here, our RV are not independent, but we can check that (16) is correct. The relation for the variances gives us a useful approximation. For instance

$$\mathbb{E}(\tilde{n}_C) \overset{b}{\sim} np = n_C, \text{ and, using (3),}$$
$$\mathbb{V}(\tilde{n}_C) \sim \mathbb{V}(\tilde{n})p^2 + \mathbb{V}(\tilde{p})n^2 + \mathbb{V}(\tilde{n})\mathbb{V}(\tilde{p}) \overset{b}{\sim} \frac{n^2 p(Lb - L + pL + 1 - p)}{L^2(b-1)b} \overset{b}{\sim} \frac{n^2 p}{Lb} \text{ for large } b.$$

It remains to estimate $\mathbb{E}\left(\frac{1}{U}\right)$ in order to complete $\mathbb{E}(\tilde{\mu}^2), \mathbb{V}(\tilde{\mu})$. Using the binomial distribution $Bin(r, p)$ does not lead to a tractable expression. But, as R is large whp, we can use the Gaussian approximation for U as follows: conditioned on $R = r$, we have

$$\mathbb{E}\left(\frac{1}{U}\right) \sim \int_1^r \frac{\exp\left(-\frac{(u-rp)^2}{2rpq}\right)}{\sqrt{2\pi rpq}u} du$$

$$\sim \int_{-rp}^{rq} \frac{\exp\left(-\frac{v^2}{2rpq}\right)}{\sqrt{2\pi rpq}} \frac{1}{rp}\left(1-\frac{v}{rp}+\frac{v^2}{r^2p^2}+\dots\right)$$
$$\sim \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{v^2}{2rpq}\right)}{\sqrt{2\pi rpq}} \frac{1}{rp}\left(1-\frac{v}{rp}+\frac{v^2}{r^2p^2}+\dots\right)$$
$$\sim \frac{1}{rp}\left(1+\frac{q}{rp}\right).$$

Unconditionning, this gives

$$\mathbb{E}\left(\frac{1}{U}; R>0\right) \overset{b}{\sim} \frac{1}{Lbp}+\frac{3q}{4p^2Lb^2} \overset{b}{\sim} \frac{1}{Lbp}$$

that we insert now into Thm 7.1.

8 The Black-Green Sampling

This analysis was motivated by an oral question by P. Flajolet.

In this section, We analyze a case in some sense opposite to the one of Sec. 6: here we do *not* observe the color of each key. At first sight, all colors are black. Nevertheless, observing their multiplicities, we want to recover their colors at the end of the algorithm.

One difficulty encountered in clustering multidimensional data streams is in maintaining summaries of each cluster which are often space intensive. Methods such as *CSketch* have been developed which use a count-min sketch to store the frequencies of attribute-value combinations in each cluster, see the paper by C. Aggarwal [2] and [1]. This motivates the model we study below, where from a slight variant of AS, we recover estimates of the number of keys appearing with each frequency (hence their colors).

We have two models: in Model I: only multiplicities can be observed, in Model II AS is speeding up when we can observe *one* extra color

Model I: only multiplicities can be observed

Assume that there are n distinct keys, among which $n_i = np_i (0 < p_i < 1)$ do have color $C_i, i = 1 \dots k$. Assume also that each *distinct* color C_i key κ_i appears μ_i times in the sample, all $\mu_i's$ being *different*, but we can't observe the keys colors. So $N = \sum_1^k n_i\mu_i$. We want to estimate n, n_i, μ_i, all unknown. (Here, we consider only the mean of our estimates). We attach to each *distinct* key κ_ℓ a counter ν_ℓ giving the number of times (multiplicity) this key appears in the sample. At the end of AS, we have in the cache U_i distinct color C_i keys, $i = 1..k$, and each one will display the *same* value for ν_i, which is obviously equal to μ_i. Hence the colors are now distinguished. This leads to the unbiased estimates $\tilde{n}_i = 2^D U_i, \tilde{n} = 2^D R, \tilde{p}_i = U_i/R$. (see Sec. 6.3).

Model II: AS speeding up when we can observe *one* extra color

Now we assume that we have an extra color Green (G), with *known* multiplicity $\mu_G > \mu_i, i = 1 \dots k$. ($\mu_i$ still unknown). We can improve the speed of AS as follows: at each depth $d, d = 0..D$, each time a key obtains the value $\nu = \mu_G$, it is obviously G, and it is extracted from the cache. We have a vector counter H such that, each time a G key is extracted at depth d from the cache, $H[d]$ is increased by 1. At the end of this new AS, the final number of process executions is D^*, say, and we have the estimates $\tilde{n}_i = 2^{D^*} U_i, \mu_G \tilde{n}_G = N - \sum_1^k \mu_i \tilde{n}_i$, hence $\tilde{n}_G$, $\tilde{n} = \sum_1^k \tilde{n}_i + \tilde{n}_G, \tilde{p}_i = \tilde{n}_i/\tilde{n}$. A more precise estimate $\tilde{n}_G$ is obtained as follows: we use $\tilde{n}_G = \sum_0^{D^*} 2^d H[d]$ (see the detailed explanation below).

Intuitively, $D^* < D$. To evaluate the difference $D - D^*$, we turn to an example.

Assume that among the n distinct keys, np $(0 < p < 1)$ are Black (B), with multiplicity μ_B and $nq, q := 1-p$ are Green (G), with *known* multiplicity $\mu_G > \mu_B$. For instance, assume that each B key is unique and each G key is present in triplicate. So we have a total of $N = np + 3qn = n(3-2p)$ keys. At the end of AS, each key with $\nu = 1$ is obviously B. As all N keys are assumed to be distributed according to the uniform permutation distribution, we can consider the effect of each key on the cache

as a Markov process: with probability $\frac{p}{3-2p}$, the key is B and it is inserted, with probability $\frac{3(1-p)}{3-2p}$, the key is G and three cases can occur: assume that the observed key appears in position $v, 1 \leq v \leq N$. Set $\tau := v/N$. Then

- With probability τ^2, the key was the third one among the three G keys with the same value, so it is deleted from the cache
- With probability $2\tau(1-\tau)$, the key was the second one, and it remains in the cache
- With probability $(1-\tau)^2$, the key is the first one, and it is inserted in the cache.

This can be seen as a Random walk on the cache. So the mean effect (on the cache size) of a G key at position v is given by

$$-\tau^2 + 0 \times 2\tau(1-\tau) + 1 \times (1-\tau)^2 = 1 - 2\tau.$$

Finally, the mean effect on the cache size of a key at position v is given by

$$\pi(\tau) = \frac{p}{3-2p} + \frac{3(1-p)}{3-2p}(1-2\tau) = \frac{3-6\tau-2p+6p\tau}{3-2p}.$$

Consider now the process beginning ($d = 0$). How many keys (in the mean) must be read in order to fill up the b positions in the cache? This is given by v_0, where $b = V(0, v_0)$ and

$$V(u_1, u_2) = \int_{u_1}^{u_2} \pi(\tau)dv = N \int_{u_1/N}^{u_2/N} \pi(\tau)d\tau = \frac{3(u_1^2 - u_2^2) + 3p(u_2^2 - u_1^2) - 3N(u_1 - u_2) + 2Np(u_1 - u_2)}{N(3-2p)}.$$

This leads to

$$v_0 = \frac{-3N + 2Np + [N(-3+2p)(2Np - 12bp - 3N + 12b)]^{1/2}}{6(p-1)}.$$

An average of $b/2$ keys (starting with bit 1) are killed for the next execution $d = 1$. But the mean number of available keys still to be read is also divided by 2. So the mean number of keys necessary to fill up the $b/2$ remaining positions in the cache is given by $v_1 - v_0$, where $b/2 = \frac{1}{2}V(v_0, v_1)$. This leads to

$$v_1 = \frac{-3N + 2Np + [N(-3+2p)(2Np - 24bp - 3N + 24b)]^{1/2}}{6(p-1)}.$$

More generally, the mean number of keys necessary to fill up the $b/2$ remaining positions in the cache at depth d is given by $v_d - v_{d-1}$, where $b/2 = 2^{-d}V(v_{d-1}, v_d)$. This leads to

$$v_d = \frac{-3N + 2Np + \left[N(-3+2p)(2Np - 2^d 12bp - 3N + 2^d 12b)\right]^{1/2}}{6(p-1)},$$

and finally, the mean total number $\mathbb{E}(D)$ is given by $\mathbb{E}(D) = \lceil D^* \rceil$, where D^* is the solution of

$$N = \frac{-3N + 2Np + \left[N(-3+2p)(2Np - 2^{D^*} 12bp - 3N + 2^{D^*} 12b)\right]^{1/2}}{6(p-1)}.$$

This gives

$$D^* = \lg\left(\frac{Np}{(3-2p)b}\right) = \lg N - \lg b + \lg p - \lg((3-2p)) = \lg n - \lg b + \lg p.$$

and

$$D - D^* = -\lg p > 0.$$

This is the more positive the less p is.

Note that, at the end, the number of B keys in the sample is estimated by

$$2^{D^*} \times \text{ number of B keys in the cache,}$$

obviously only B keys (with $\nu = 1$) remain in the cache and the number of G keys in the sample is estimated by

$$3. \sum_{d=0}^{D^*} 2^d H[d].$$

Indeed, imagine that we mark a key with a $*$ as soon as it is decided to be G (because it is the third time we observe it). At depth 0, $v \in [0, v_0)]$, all marked keys are counted in $H[0]$. At depth 1, $v \in [v_0, v_1]$, all marked keys (starting with 0) are counted in $H[1]$, this corresponds in the mean, to $2H[1]$ G keys, etc. Actually, the vector counter H could be replaced by a single counter H into which, at each depth d, we add the number of extracted G keys $\times 2^d$.

9 Conclusion

Once again, the techniques using Gumbel-like distributions and Stein methodology proved to be quite efficient in the analysis of algorithms such as Adaptive Sampling.

Acknowledgements

We would like to thank J. Lumbroso with whom we had many interesting discussions. Matthew Drescher is supported by ERC Consolidator Grant 615640-ForEFront. We would also like to thank the referee for careful reading and many useful suggestions that improved the paper. We would also like to thank Ben Karsin for the helpful discussions.

References

[1] C. Aggarwal *Data Mining*, p 417 Springer-Verlag, 2015.

[2] C. Aggarwal A Framework for Clustering Massive-Domain Data Streams IEEE 25th International Conference on Data Engineering, 2009

[3] M.M. Astrahan, M. Schkolnick, and K.Y. Whang. Approximating the number of unique values of an attribute without sorting. *Information Sciences*, 12:11–15, 1987.

[4] A. D. Barbour and L. H. Y. Chen. *An introduction to Stein's method,* volume 4 of Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. Singapore University Press, 2005.

[5] A. D. Barbour and L. H. Y. Chen. *Stein's method and applications,* volume 5 of Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. Singapore University Press, 2005.

[6] J. Chae and D. Thom and H. Bosch and Y. Jang and R. Maciejewski and D. S. Ebert and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. IEEE VAST, pages 143–152, 2012

[7] P. Flajolet. On adaptive sampling. *Computing*, 34:391–400, 1990.

[8] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[9] P. B. Gibbons. Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *Proceedings of the 27th VLDB Conference ,Roma*, 2001.

[10] P. Hitczenko and G. Louchard. Distinctness of compositions of an integer: a probabilistic analysis. *Random Structures and Algorithms*, 19(3,4):407–437, 2001.

[11] P. Jacquet and W. Szpankowski. Analytic depoissonization and its applications. *Theoretical Computer Science*, 201(1-2):1–62, 1998.

[12] D. Kane and J. Nelson and D. Woodruff An Optimal Algorithm for the Distinct Elements Problem Proceedings of the 29th Annual ACM Symposium on Principles of Database Systems, 2010

[13] M. Loève. *Probability Theory, 3rd ed.* D. Van Nostrand, 1963.

[14] G. Louchard. Probabilistic analysis of adaptative sampling. *Random Structures and Algorithms*, 10:157–168, 1997.

[15] G. Louchard and H. Prodinger. Asymptotics of the moments of extreme-value related distribution functions. *Algorithmica*, 46:431–467, 2006.

[16] G. Louchard and H. Prodinger. On gaps and unoccupied urns in sequences of geometrically distributed random variables. *Discrete Mathematics*, 308,9:1538–1562, 2008. Long version: http://www.ulb.ac.be/di/mcs/louchard/gaps18.ps.

[17] G. Louchard, H. Prodinger, and M.D. Ward. The number of distinct values of some multiplicity in sequences of geometrically distributed random variables. *Discrete Mathematics and Theoretical Computer Science*, AD:231–256, 2005. 2005 International Conference on Analysis of Algorithms.

[18] G. Louchard and Y. Swan. The adaptive sampling revisited. Technical report, 2017. Long version: http://www.ulb.ac.be/di/mcs/louchard/louchard.papers/asnew5.pdf.

[19] A. Rajaraman and J.D. Ullman. *Mining of Massive Datasets.* Cambridge University Press, 2012.

[20] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

[21] Milad Shokouhi. Detecting seasonal queries by time-series analysis. Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China,July 25-29, 2011

[22] Open Source. Feature request estimating distinct. Technical report, https://wiki.postgresql.org/wiki/Estimating_Distinct, 2015.

[23] M. Wegman. Sample counting. 1984. Private communication to P. Flajolet.

[24] Jeffrey Wong, Chris Colburn, Elijah Meeks, and Shankar Vedaraman. Rad – outlier detection on big data. Technical report, https://medium.com/netflix-techblog/rad-outlier-detection-on-big-data-d6b0494371cc, 2015.

A Asymmetric Adaptive Sampling

For the sake of completeness, we analyze in this section the Asymmetric Adaptive Sampling. Assume that the hashing function gives asymmetric distributed bits. Let p denote the probability of bit 1 ($q := 1 - p$). Now, the number of keys in the cache is asymptotically Poisson with parameter nq^d and the number of keys in the twin bucket is asymptotically Poisson with parameter $npq^{d-1} = n\frac{p}{q}q^d$. So we set here

$$\begin{aligned} Q &:= 1/q, \\ Z &:= \frac{RQ^D}{n}, \\ \log &:= \log_Q, \end{aligned}$$

$$\begin{aligned}\eta &:= d - \log n,\\ L &:= \ln Q,\\ \tilde{\alpha} &:= \alpha/L,\\ \{x\} &:= \text{ fractional part of } x,\\ \chi_l &:= \frac{2l\pi\mathbf{i}}{L}.\end{aligned}$$

So the asymptotic distribution is now given by

$$p(r,d) \sim f(r,\eta) = \exp(-e^{-L\eta})\frac{e^{-Lr\eta}}{r!}\left[1 - \exp(-e^{-L\eta}p/q)\sum_{k=0}^{b-r}\frac{e^{-Lk\eta}(p/q)^k}{k!}\right], \tag{17}$$

and

$$p(.,d) := \mathbb{P}(D = d) = \sum_{r=0}^{b} p(r,d).$$

This leads to

$$\phi(r,\alpha) = \int_{-\infty}^{\infty} e^{\alpha\eta} f(r,\eta)d\eta = \frac{\Gamma(r-\tilde{\alpha})}{Lr!} - \sum_{k=0}^{b-r}\frac{\Gamma(r+k-\tilde{\alpha})q^{r+k-\tilde{\alpha}}(p/q)^k}{Lr!k!}.$$

In the sequel, we only provide the main related theorems. All detailed proofs can be found in this paper long version: [18].

A.1 Moments of $D - \log n$

We have

Theorem A.1 *The asymptotic moments of $D - \log n$ are given by*

$$\begin{aligned}
\tilde{m}_{1,k} &= -\frac{\psi(k)}{L^2 k} + \sum_{i=0}^{b-k}\frac{(\psi(k+i) - L)q^k p^i \Gamma(k+i)}{L^2\Gamma(k+1)\Gamma(i+1)}, \quad k > 0,\\
\tilde{m}_{1,0} &= \frac{1}{2} + \frac{\gamma}{L} + \sum_{i=1}^{b}\frac{(\psi(i) - L)p^i}{iL^2},\\
\tilde{m}_{2,k} &= \frac{\psi(1,k) + \psi(k)^2}{L^3 k} + \sum_{i=0}^{b-k} - \frac{(-2\psi(k+i)L + L^2 + \psi(1,k+i) + \psi(k+i)^2)q^k p^i \Gamma(k+i)}{L^3\Gamma(k+1)\Gamma(i+1)}, \quad k > 0,\\
\tilde{m}_{1,0} &= \frac{1}{3} + \frac{\gamma}{L} + \frac{\pi^2}{6L^2} + \frac{\gamma^2}{L^2} + \sum_{i=1}^{b} - \frac{(-2\psi(i)L + L^2 + \psi(1,i) + \psi(i)^2)p^i}{iL^3},\\
\tilde{w}_{1,k} &= \sum_{l\neq 0}\left[-\frac{\psi(k+\chi_l)\Gamma(k+\chi_l)}{L^2\Gamma(k+1)} + \sum_{i=0}^{b-k}\frac{(\psi(k+i+\chi_l) - L)\Gamma(k+i+\chi_l)q^{k+i}}{L^2\Gamma(k+1)\Gamma(i+1)}\right]e^{-2l\pi\mathbf{i}\log n}, \quad k > 0,\\
\tilde{w}_{1,0} &= \sum_{l\neq 0}\left[-\frac{\psi(\chi_l)\Gamma(\chi_l)}{L^2} + \sum_{i=0}^{b}\frac{(\psi(i+\chi_l) - L)\Gamma(i+\chi_l)q^i}{L^2\Gamma(i+1)}\right]e^{-2l\pi\mathbf{i}\log n}, \quad k > 0.
\end{aligned}$$

A.2 Moments of Z

Theorem A.2 *The non-periodic components of the moments of Z are given by*

$$m_{1,k} = 1 + \frac{(b-k)!}{L}\sum_{i=1}^{k-1}\left\{{k \atop i}\right\}\frac{q^{i-k} - 1}{(k-i)(b-i)!},$$

$$\mathbb{V}(Z) \sim \frac{p}{(b-1)qL}.$$

The periodic component is obtained as follows

$$w_{1,k} = \sum_{l \neq 0} \frac{1}{L} \sum_{j=1}^{k-1} \left\{ {k \atop j} \right\} \left[(1 - q^{j-k} \right] \Gamma(j - k + \chi_l) \binom{b - k + \chi_l}{b - j} e^{-2l\pi \mathbf{i} \log n}.$$

A.3 Distribution of R

Theorem A.3 *The asymptotic distribution of R is given by*

$$\mathbb{P}(R = r) = p(r, .) \sim \phi(r, 0) = \frac{1}{L} \left[\frac{1}{r} - \sum_{u=r}^{b} \frac{(u-1)!}{r!(u-r)!} p^u (q/p)^r \right], r \geq 1,$$

$$\begin{aligned} \mathbb{E}(R) &\sim \frac{1}{L} \left[b - qb \right] = \frac{pb}{L}, \\ \mathbb{E}(R^2) &\sim \frac{1}{L} \left[b(b+1)/2 - q^2 \frac{b(b-1)}{2} - qb \right]. \end{aligned}$$

A.4 Moments of $1/R, R > 0$ for large b

Theorem A.4 *The asymptotic moments of $1/R, R > 0$ for large b, with $R > 0$ are given by*

$$\mathbb{E}\left(\frac{1}{R}; R > 0 \right) \overset{b}{=} \frac{1}{Lb} \frac{p}{q}.$$