Approximating the Probability of an Itemset being Frequent

Nele Dexters

Abstract

In the literature, there exist an analytical and empirical study for the behaviour of the Apriori Algorithm, the best known frequent itemset mining algorithm [PVGG04]. For the analytical part, a very simple shopping model is used where every item has the same probability and all the items and all the transactions are independent. The notion of S_l, the probability that a certain set consisting of l elements is a frequent set, is introduced and approximated using Chernoff bounds. This technical report discusses a new, statistically inspired approximation of S_l that is easier to compute than the Chernoff result. This new approach is based on the approximation of the Binomial Distribution.

1 Introduction

The frequent itemset problem, introduced in [AIS93, AS94], is a well known and interesting basic problem at the core of many data mining problems [AIS93, AS94, Goe03, GZ03]. The problem is, given a large database of basket data, i.e. subsets of a fixed set of items $\mathcal{I}$, and a user-defined support threshold k, determine which sets of items occur together in at least k baskets. In the last two decades, several different algorithms for solving this problem were proposed [AIS93, AS94, HPY00, Zak00]. The best known algorithm is the Apriori Algorithm, introduced in [AS94].

In the literature [PVGG04], there exist an analytical and empirical study for the behaviour of Apriori. In this theoretical study, the notions of C_l, S_l and F_l were introduced to gain more insight in the average case performance of Apriori. C_l is the probability that a certain set consisting of l items is a candidate set, S_l is the probability that such a candidate set is frequent, and F_l is the probability that such a candidate set of length l is a failure, so is not frequent. These probabilities were estimated with Chernoff bounds in the case of the simple shopping model where all the items were independent and had the same probability of being chosen, p, and all the transactions were independent as well. In this technical report, we use the same simple model of shopping behaviour. Based on statistics, a new approximation for S_l is derived. This new approximation is fast and easy to compute and gives better results than the Chernoff bounds.

The study of the average case performance of Apriori is not easy. In the expressions found for the different probabilities, C_l, S_l and F_l in [PVGG04], combinatorial sums appeared that are hard to compute. Therefore, it is good to have a computable form for these probabilities. In this technical report, we focus on such a form for S_l to estimate the size of the result. This computable form is reached by straigthforward statistical approximation and produces accurate estimates for S_l when the amount of basket data in the database, b, is large.

For the database, we assume that there are m possible items that can be bought and b baskets or transactions. The user-defined support threshold for the frequent itemset mining problem is denoted by k. The model of shopping behaviour used in the rest of this technical report is the simple model based on the following three assumptions:

- each item has the same probability p
- all the m items are independent
- all the b transactions are independent

S_l is the probability ($0 \leq S_l \leq 1$) that a set consisting of l items $\{1, \ldots, l\}$ is a frequent set. In our simple shopping model, each basket is filled at random, all the items are independent and have the same probability p, so any other set consisting of l items has the same probability of success, S_l.

Analoguously as in [PVGG04], we can now define the following conditions with respect to a single basket:

- condition M_0: the basket contains l items $\{1, \ldots, l\}$
- condition M_h $(1 \leq h \leq l)$: the basket contains all items from $\{1, \ldots, l\}$, except a fixed item h, so contains $l - 1$ items

Each basket obeys at most one of these $l + 1$ disjoint conditions M_h, $0 \leq h \leq l$.

With these definitions of the different conditions and the above knowledge of the used shopping model, we can now write down some basic probabilities. The probability that a randomly filled basket obeys condition M_0 is

$$P(l) = p^l.$$

The probability that a randomly filled basket obeys condition M_h ($1 \leq h \leq l$) is

$$Q(l) = p^{l-1}(1 - p).$$

The probability that at least k baskets obey condition M_0 can be found by

$$S_l = \sum_{j \geq k} \binom{b}{j} [P(l)]^j [1 - P(l)]^{b-j}$$

and describes the probability that the set $\{1, \ldots, l\}$ is a frequent set.

Outline

The rest of this Technical Report is organized as follows. In Section 2, some notions and distributions in statistics are shortly revisisted. They are necessary to understand the new approach. This section is the statistical foundation of our approximation. In Section 3, the new results for S_l are presented. Section 4 discusses these results w.r.t. the experimental settings used in [PVGG04]. Here it is shown that the new approach yields fast and easy computations and better results. Section 5 concludes and points out future work.

2 Statistical Background Information

This section gives a general background on the statistical components used and can therefore be seen as the statistical foundation of the new approach presented in Section 3. For more information, see [DK95], [OGD80], [JK69] or any other reference book on statistics.

2.1 The Binomial Distribution

Distribution

Consider stochast $X = X_1 + \ldots + X_n$, where the stochasts X_j $(1 \leq j \leq n)$ are independent and identically distributed (i.i.d.), following a Bernoulli Distribution $B(1, p)$. The Bernoulli Distribution is the distribution that is used to describe an experiment with two possible outcomes, a "success" outcome with probability p and a "failure" outcome with probability $1 - p$. The classical example of this distribution is tossing a coin where both sides have equal probability $1/2$ to be on top. Each stochast X_j $(1 \leq j \leq n)$ equals 1 when it represents a success and 0 when it represents a failure. X, the result of the Bernoulli sum, is now defined to follow the Binomial Distribution $X \sim B(n, p)$. This is a discrete distribution where X represents the amount of successes (so the amount of 1-occurrences) in n independent Bernoulli experiments with success probability p . The Binomial Distribution is therefore the result of n independent repetitions of a random experiment with two possible outcomes, success with probability p and failure with probability $1 - p$. There are two parameters: n, the number of repetitions, and p, the probability of success in the repeated Bernoulli experiment. This p has to be the same for all the n Bernoulli trials.

The probability that there are j successes in the n successive Bernoulli $B(1, p)$ experiments is

$$P(X = j) = \left(\begin{array}{c} n \\ j \end{array} \right) p^j \ (1-p)^{n-j}.$$

The probability of having at least k ($0 \leq k \leq n$) successes in the n successive $B(1, p)$ experiments is

$$P(X \geq k) = \sum_{j \geq k} \left(\begin{array}{c} n \\ j \end{array} \right) p^j \ (1-p)^{n-j} = \sum_{j=k}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) p^j \ (1-p)^{n-j}.$$

Approximation
In the statistics literature there exist good approximations for the Binomial Distribution. An overview is given in Figure 1. It is not the purpose of this technical report to cover the proofs of these properties. They can be found in the better statistical handbooks.

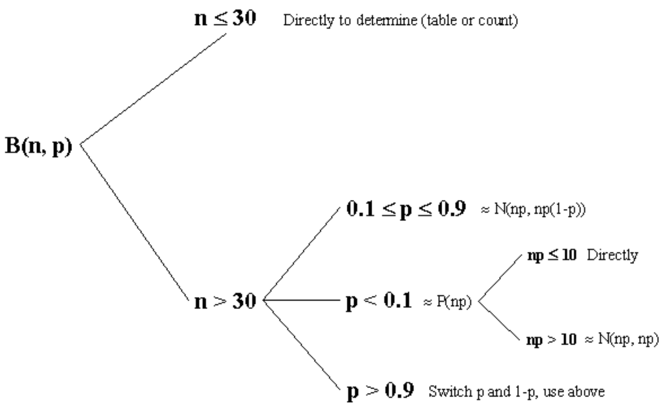

Figure 1: Overview approximations

When approximating a discrete distribution by a continous distribution, we have to take care of the continuity correction by adding or subtracting 0.5.

Remark
There exist other ways of approximating a Binomial Distribution. An overview can be found in [JK69]. For this technical report, we have chosen the most commonly used, simple and straightforward approximations.

The difference in approximating the Binomial Distribution by the Normal or the Poisson Distribution is the role p plays. In the Normal approximation, $n \to \infty$ and p is fixed. In the Poisson approximation, $n \to \infty$, $p \to 0$ but np, the Poisson parameter, stays constant.

2.2 The Poisson Distribution

Distribution
A random variable X is said to follow a discrete Poisson Distribution $P(\lambda)$ with parameter $0 < \lambda < \infty$ if

$$P[X = j] = p_j = \frac{\lambda^j}{j!} e^{-\lambda}.$$

Approximation
The Poisson Distribution is the limit of a Binomial Distribution, as the number of Bernoulli trials, n, gets large and the probability of success, p, gets small. Formally, a Poisson distribution approaches a Binomial Distribution if $n \to \infty$ and $p \to 0$ in such a way that their product remains constant, $np = \lambda$. This value is called the Poisson parameter.

2.3 The (Standard) Normal Distribution

First, the Standard Normal Distribution is considered. Then it is extended to the Normal Distribution. In contrast with the previously considered discrete distributions, these distributions are continuous.

Standard Normal Distribution
The Standard Normal Distribution $N(0, 1)$ is defined by the probability density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

The distribution function (or cumulative probability distribution function) of $N(0, 1)$ is defined by

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\mathrm{d}t, \qquad x \in \mathbb{R}$$

and describes the surface under the graph of ϕ from $-\infty$ to the point x (Figure 2).

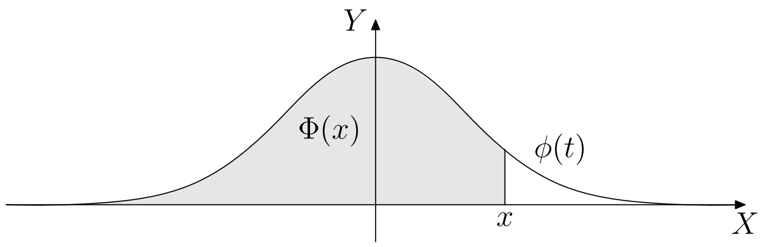

Figure 2: Grafical illustration of Φ.

It is easy to see that $\Phi(-x) = 1 - \Phi(x)$ because of the symmetry around the Y-axis and the fact that ϕ is a density function, so the total surface under the graph is 1. We can use this property when we need to find Φ in large values of x. When just computing $\Phi(x)$ when x is large, it is possible that this results in 1, while it is known that the result is close to 1 but not equal to it. A better way to compute a more accurate value for $\Phi(x)$ is to compute $1 - \Phi(-x)$. In this expression, $\Phi(-x)$ is a very small value close to 0 but not equal to it, so $1 - \Phi(-x)$ results in an expression close to 1 but not equal to 1.

Normal Distribution
The general Normal Distribution $N(\mu, \sigma^2)$ is the distribution of $X = \sigma Z + \mu$ where $Z \sim N(0,1)$. The Standard Normal Distribution is in fact a special case of the Normal Distribution $N(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$, where $\mu = 0$ and $\sigma = 1$.

The density function is

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad x, \mu \in \mathbb{R} \text{ and } \sigma > 0.$$

The distribution function

$$F_X(x) = P[X \leq x] = F_Z\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

can be computed by using the knowledge of the Standard Normal Distribution:

$$\frac{X - E[X]}{\sqrt{Var[X]}} = \frac{X-\mu}{\sigma} = Z \sim N(0,1).$$

3 Efficient computation of S_l

Based on the theory of Section 2.1,

$$S_l = \sum_{j=k}^{b} \binom{b}{j} [P(l)]^j [1 - P(l)]^{b-j}$$

can be seen as $P(X \geq k) = 1 - P(X < k)$ with $X \sim B(b, P(l))$. We now use the appropriate approximation for the Binomial Distribution and investigate the three different situations that can appear. The case $b \leq 30$ is not considered because b is the amount of tuples in the database and this is supposed to be larger than 30.

If the above formula for S_l is compared with the new results found by the approximation, it is clear that the new approach yields fast and easy computations. The binomial sums do not have to be computed but are approximated by simple formulas using Φ and Poisson.

3.1 $b > 30$ and $0.1 \leq P(l) \leq 0.9$

We approximate the Binomial distributed $X \sim B(b, P(l))$ by the Normal distributed $Y \sim N(bP(l), bP(l)(1 - P(l)))$, so

$$Z = \frac{Y - bP(l)}{\sqrt{bP(l)(1 - P(l))}} \sim N(0,1).$$

Because we approximate a discrete distribution by a continuous distribution, we have to take care of the continuity correction. In details:

$$\begin{aligned}
P(X < k) &\approx P(Y \le k - 0.5) \\
&= P\left(Z \le \frac{(k-0.5) - bP(l)}{\sqrt{bP(l)(1-P(l))}}\right) \\
&= \Phi\left(\frac{(k-0.5) - bP(l)}{\sqrt{bP(l)(1-P(l))}}\right)
\end{aligned}$$

so

$$\begin{aligned}
S_l = P(X \ge k) &= 1 - P(X < k) \\
&= 1 - \Phi\left(\frac{(k-0.5) - bP(l)}{\sqrt{bP(l)(1-P(l))}}\right).
\end{aligned}$$

3.2 $b > 30$ and $P(l) < 0.1$

In this case, the Binomial distributed $X \sim B(b, P(l))$ will be approximated by the Poisson distributed $Y \sim P(bP(l))$. Dependent of the value of $bP(l)$ we can distinguish two different cases.

3.2.1 $bP(l) \le 10$

In this case, the approximation of the discrete Binomial Distribution by the discrete Poisson Distribution is used. A continuity correction is not necessary.

$$\begin{aligned}
P(X < k) &= P(Y < k) \\
&= P(Y \le k - 1) \\
&= F(k-1) \\
&= \sum_{j=0}^{k-1} \frac{(bP(l))^j e^{-bP(l)}}{j!}
\end{aligned}$$

so

$$\begin{aligned}
S_l = P(X \ge k) &= 1 - P(X < k) \\
&= 1 - F(k-1) \\
&= 1 - \sum_{j=0}^{k-1} \frac{(bP(l))^j e^{-bP(l)}}{j!}
\end{aligned}$$

3.2.2 $bP(l) > 10$

In this case, the discrete Poisson Distribution is approximated by the continuous Normal Distribution and we have to take care of the continuity correction.

$$\begin{aligned} X \sim B(b, P(l)) &\approx Y \sim P(bP(l)) \\ &\approx T \sim N(bP(l), bP(l)) \end{aligned}$$

so

$$Z = \frac{T - bP(l)}{\sqrt{bP(l)}} \sim N(0, 1).$$

Therefore

$$\begin{aligned} P(X < k) &\approx P(T \leq k - 0.5) \\ &= P\left(Z \leq \frac{(k - 0.5) - bP(l)}{\sqrt{bP(l)}}\right) \\ &= \Phi\left(\frac{(k - 0.5) - bP(l)}{\sqrt{bP(l)}}\right) \end{aligned}$$

so

$$\begin{aligned} S_l = P(X \geq k) &= 1 - P(X < k) \\ &= 1 - \Phi\left(\frac{(k - 0.5) - bP(l)}{\sqrt{bP(l)}}\right). \end{aligned}$$

3.3 $b > 30$ and $P(l) > 0.9$

In this case, $X \sim B\left(b, P(l)\right)$ with $P(l) > 0.9$. $X' = b - X \sim B(b, 1 - P(l))$ with $1 - P(l) < 0.1$ is constructed. We are now in the previous case (see Section 3.2) with X' instead of X. Therefore

$$\begin{aligned} P(X < k) &= P(b - X > b - k) \\ &= P(X' > b - k) \\ &= 1 - P(X' \leq b - k) \end{aligned}$$

and

$$\begin{aligned} S_l = P(X \geq k) &= 1 - P(X < k) \\ &= 1 - \big(1 - P(X' \leq b - k)\big) \\ &= P(X' \leq b - k). \end{aligned}$$

We know that $X' \sim B\left(b, (1 - P(l))\right)$ with $1 - P(l) < 0.1$ so

$$X' \approx Y \sim P\left(b(1 - P(l))\right)$$

as seen in Section 3.2. Again, there can occur two situations that have to be considered.

3.3.1 $b\,(1-P(l)) \leq 10$

$$\begin{aligned} P(X' \leq b-k) &\approx P(Y \leq b-k) \\ &= F(b-k) \\ &= \sum_{j=0}^{b-k} \frac{(b(1-P(l)))^j e^{-b(1-P(l))}}{j!} \end{aligned}$$

For S_l this gives:

$$S_l = F(b-k) = \sum_{j=0}^{b-k} \frac{(b(1-P(l)))^j e^{-b(1-P(l))}}{j!}$$

3.3.2 $b\,(1-P(l)) > 10$

In this case, $Y \sim P\,(b(1-P(l)))$ will be approximated by $T \sim N\,(b(1-P(l))\,,$ $b\,(1-P(l)))$, so

$$Z = \frac{T - b(1-P(l))}{\sqrt{b(1-P(l))}} \sim N(0,1).$$

Therefore

$$\begin{aligned} P(X' \leq b-k) &\approx P(T \leq b-k+0.5) \\ &= P\left(Z \leq \frac{(b-k+0.5) - b(1-P(l))}{\sqrt{b(1-P(l))}}\right) \\ &= \Phi\left(\frac{(b-k+0.5) - b(1-P(l))}{\sqrt{b(1-P(l))}}\right). \end{aligned}$$

For S_l this gives:

$$S_l = \Phi\left(\frac{(b-k+0.5) - b(1-P(l))}{\sqrt{b(1-P(l))}}\right).$$

4 Experimental Results For S_l

4.1 Quality of the New Approach

In this section, the results of the new approach are compared to the exact values. We show that our results are quite good and that they can be computed very fast and easy.

4.1.1 $b = 1024$ and $p = 1/2$

$b > 30$
$P(l) = p^l$, so for every value of l ($l = 1, 2, 3, 4, 5$), $P(l)$ has a different value.

- $\boldsymbol{l = 1} \rightarrow 0.1 \leq P(l) = 1/2 \leq 0.9 \Rightarrow$ Section 3.1
- $\boldsymbol{l = 2} \rightarrow 0.1 \leq P(l) = 1/4 \leq 0.9 \Rightarrow$ Section 3.1
- $\boldsymbol{l = 3} \rightarrow 0.1 \leq P(l) = 1/8 \leq 0.9 \Rightarrow$ Section 3.1
- $\boldsymbol{l = 4} \rightarrow P(l) = 1/16 < 0.1$ and $bP(l) = 64 > 10 \Rightarrow$ Section 3.2.2
- $\boldsymbol{l = 5} \rightarrow P(l) = 1/32 < 0.1$ and $bP(l) = 32 > 10 \Rightarrow$ Section 3.2.2

The results can be found in Table 2. The computation of this table is based on the following example. The calculations are performed in Maple. The original values for S_l can be found in Table 1.

Example
Let us consider $l = 1$ and $k = 1$. In this case $P(1) = 1/2$ and the approach of Section 3.1 has to be followed. $bP(1) = 512$ and $1 - P(1) = 1/2$, so $bP(1)(1 - P(1)) = 256$.

$$\begin{aligned} S_1 &= 1 - \Phi\left(\frac{(k-0.5) - bP(1)}{\sqrt{bP(1)(1-P(1))}}\right) \\ &= 1 - \Phi\left(\frac{(1-0.5) - 512}{\sqrt{256)}}\right) \\ &= 1 - \Phi(-31.97) \\ &= 1 - 1.48\ 10^{-224} \end{aligned}$$

When $l = 5$ and $k = 533$, $P(5) = 1/32$ and the approach of Section 3.2.2 has to be followed. $bP(5) = 32$.

$$\begin{aligned} S_5 &= 1 - \Phi\left(\frac{(k-0.5) - bP(5)}{\sqrt{bP(5)}}\right) \\ &= 1 - \Phi\left(\frac{(533-0.5) - 32}{\sqrt{32}}\right) \\ &= 1 - \Phi(88.48) \\ &= \Phi(-88.48) \\ &= 6.26\ 10^{-1703} \end{aligned}$$

□

k	S_1	S_2	S_3	S_4	S_5
1	$1.0 - 5.6\ 10^{-309}$	$1.0 - 1.2\ 10^{-128}$	$1.0 - 4.1\ 10^{-60}$	$1.0 - 2.0\ 10^{-29}$	$1.0 - 7.6\ 10^{-15}$
2	$1.0 - 5.7\ 10^{-306}$	$1.0 - 4.0\ 10^{-126}$	$1.0 - 6.1\ 10^{-58}$	$1.0 - 1.4\ 10^{-27}$	$1.0 - 2.6\ 10^{-13}$
3	$1.0 - 2.9\ 10^{-303}$	$1.0 - 6.8\ 10^{-124}$	$1.0 - 4.5\ 10^{-56}$	$1.0 - 4.8\ 10^{-26}$	$1.0 - 4.4\ 10^{-12}$
4	$1.0 - 1.0\ 10^{-300}$	$1.0 - 7.7\ 10^{-122}$	$1.0 - 2.2\ 10^{-54}$	$1.0 - 1.1\ 10^{-24}$	$1.0 - 5.0\ 10^{-11}$
5	$1.0 - 2.5\ 10^{-298}$	$1.0 - 6.6\ 10^{-120}$	$1.0 - 8.1\ 10^{-53}$	$1.0 - 1.9\ 10^{-23}$	$1.0 - 4.3\ 10^{-10}$
22	$1.0 - 1.5\ 10^{-265}$	$1.0 - 3.1\ 10^{-95}$	$1.0 - 2.3\ 10^{-34}$	$1.0 - 1.5\ 10^{-10}$	$1.0 - 2.4\ 10^{-2}$
32	$1.0 - 9.2\ 10^{-250}$	$1.0 - 3.3\ 10^{-84}$	$1.0 - 5.4\ 10^{-27}$	$1.0 - 2.0\ 10^{-6}$	$5.2\ 10^{-1}$
33	$1.0 - 2.9\ 10^{-248}$	$1.0 - 3.4\ 10^{-83}$	$1.0 - 2.4\ 10^{-25}$	$1.0 - 4.3\ 10^{-6}$	$4.5\ 10^{-1}$
45	$1.0 - 2.4\ 10^{-231}$	$1.0 - 5.7\ 10^{-72}$	$1.0 - 1.6\ 10^{-19}$	$1.0 - 4.2\ 10^{-3}$	$1.6\ 10^{-2}$
57	$1.0 - 6.8\ 10^{-216}$	$1.0 - 3.1\ 10^{-62}$	$1.0 - 3.7\ 10^{-14}$	$8.3\ 10^{-1}$	$3.1\ 10^{-5}$
58	$1.0 - 1.2\ 10^{-214}$	$1.0 - 1.7\ 10^{-61}$	$1.0 - 9.2\ 10^{-14}$	$8.0\ 10^{-1}$	$1.6\ 10^{-5}$
64	$1.0 - 1.9\ 10^{-207}$	$1.0 - 4.0\ 10^{-57}$	$1.0 - 1.4\ 10^{-11}$	$5.2\ 10^{-1}$	$2.5\ 10^{-7}$
65	$1.0 - 2.9\ 10^{-206}$	$1.0 - 2.0\ 10^{-56}$	$1.0 - 3.0\ 10^{-11}$	$4.7\ 10^{-1}$	$1.2\ 10^{-7}$
91	$1.0 - 6.2\ 10^{-178}$	$1.0 - 1.9\ 10^{-40}$	$1.0 - 1.1\ 10^{-4}$	$5.8\ 10^{-4}$	$2.3\ 10^{-18}$
120	$1.0 - 1.5\ 10^{-150}$	$1.0 - 7.6\ 10^{-27}$	$7.9\ 10^{-1}$	$5.3\ 10^{-11}$	$2.0\ 10^{-34}$
121	$1.0 - 1.2\ 10^{-149}$	$1.0 - 1.9\ 10^{-26}$	$7.6\ 10^{-1}$	$2.6\ 10^{-11}$	$4.7\ 10^{-35}$
128	$1.0 - 1.3\ 10^{-143}$	$1.0 - 9.8\ 10^{-24}$	$5.1\ 10^{-1}$	$1.4\ 10^{-13}$	$1.7\ 10^{-39}$
129	$1.0 - 8.8\ 10^{-143}$	$1.0 - 2.3\ 10^{-23}$	$4.8\ 10^{-1}$	$6.6\ 10^{-14}$	$3.8\ 10^{-40}$
186	$1.0 - 2.8\ 10^{-100}$	$1.0 - 7.1\ 10^{-8}$	$1.3\ 10^{-7}$	$8.9\ 10^{-39}$	$6.4\ 10^{-83}$
247	$1.0 - 3.7\ 10^{-65}$	$7.5\ 10^{-1}$	$2.0\ 10^{-24}$	$1.2\ 10^{-75}$	$5.2\ 10^{-139}$
248	$1.0 - 1.2\ 10^{-64}$	$7.3\ 10^{-1}$	$8.7\ 10^{-25}$	$2.4\ 10^{-76}$	$5.3\ 10^{-140}$
256	$1.0 - 9.6\ 10^{-61}$	$5.1\ 10^{-1}$	$1.1\ 10^{-27}$	$7.2\ 10^{-82}$	$4.7\ 10^{-148}$
257	$1.0 - 2.9\ 10^{-60}$	$4.8\ 10^{-1}$	$4.8\ 10^{-28}$	$1.4\ 10^{-82}$	$4.6\ 10^{-149}$
377	$1.0 - 8.1\ 10^{-18}$	$3.8\ 10^{-17}$	$1.4\ 10^{-87}$	$9.8\ 10^{-182}$	$4.9\ 10^{-286}$
503	$7.2\ 10^{-1}$	$6.9\ 10^{-62}$	$1.5\ 10^{-178}$	$2.2\ 10^{-314}$	$2.1\ 10^{-458}$
504	$7.0\ 10^{-1}$	$2.4\ 10^{-62}$	$2.3\ 10^{-179}$	$1.5\ 10^{-315}$	$7.0\ 10^{-460}$
512	$5.1\ 10^{-1}$	$4.0\ 10^{-66}$	$4.4\ 10^{-186}$	$6.6\ 10^{-325}$	$9.3\ 10^{-472}$
513	$4.9\ 10^{-1}$	$1.3\ 10^{-66}$	$6.3\ 10^{-187}$	$4.4\ 10^{-326}$	$3.0\ 10^{-473}$
533	$1.0\ 10^{-1}$	$1.6\ 10^{-76}$	$3.3\ 10^{-204}$	$5.6\ 10^{-350}$	$1.9\ 10^{-503}$

Table 1: Exact values for S_l for $b = 1024$, $p = 1/2$ and selected values of k.

k	S_1	S_2	S_3	S_4	S_5
1	$1.0 - 1.48\ 10^{-224}$	$1.0 - 3.19\ 10^{-76}$	$1.0 - 9.98\ 10^{-34}$	$1.0 - 1.03\ 10^{-15}$	$1.0 - 1.28\ 10^{-8}$
2	$1.0 - 1.09\ 10^{-223}$	$1.0 - 1.21\ 10^{-75}$	$1.0 - 3.13\ 10^{-33}$	$1.0 - 2.80\ 10^{-15}$	$1.0 - 3.49\ 10^{-8}$
3	$1.0 - 8.03\ 10^{-223}$	$1.0 - 4.55\ 10^{-75}$	$1.0 - 9.71\ 10^{-33}$	$1.0 - 7.50\ 10^{-15}$	$1.0 - 9.20\ 10^{-8}$
4	$1.0 - 5.88\ 10^{-222}$	$1.0 - 1.71\ 10^{-74}$	$1.0 - 2.99\ 10^{-32}$	$1.0 - 1.98\ 10^{-14}$	$1.0 - 2.35\ 10^{-7}$
5	$1.0 - 4.28\ 10^{-221}$	$1.0 - 6.37\ 10^{-74}$	$1.0 - 9.11\ 10^{-32}$	$1.0 - 5.13\ 10^{-14}$	$1.0 - 5.83\ 10^{-7}$
22	$1.0 - 1.09\ 10^{-206}$	$1.0 - 1.51\ 10^{-64}$	$1.0 - 4.01\ 10^{-24}$	$1.0 - 5.41\ 10^{-8}$	$1.0 - 3.17\ 10^{-2}$
32	$1.0 - 1.92\ 10^{-198}$	$1.0 - 2.44\ 10^{-59}$	$1.0 - 3.81\ 10^{-20}$	$1.0 - 2.43\ 10^{-5}$	$1.0 - 4.65\ 10^{-1}$
33	$1.0 - 1.25\ 10^{-197}$	$1.0 - 7.89\ 10^{-59}$	$1.0 - 9.08\ 10^{-20}$	$1.0 - 4.12\ 10^{-5}$	$4.65\ 10^{-1}$
45	$1.0 - 5.60\ 10^{-188}$	$1.0 - 6.67\ 10^{-53}$	$1.0 - 1.51\ 10^{-15}$	$1.0 - 7.39\ 10^{-3}$	$1.36\ 10^{-2}$
57	$1.0 - 1.43\ 10^{-178}$	$1.0 - 2.68\ 10^{-47}$	$1.0 - 7.09\ 10^{-12}$	$1.0 - 1.74\ 10^{-1}$	$7.42\ 10^{-6}$
58	$1.0 - 8.46\ 10^{-178}$	$1.0 - 7.58\ 10^{-47}$	$1.0 - 1.35\ 10^{-11}$	$1.0 - 2.08\ 10^{-1}$	$3.28\ 10^{-6}$
64	$1.0 - 3.38\ 10^{-173}$	$1.0 - 3.52\ 10^{-44}$	$1.0 - 5.48\ 10^{-10}$	$1.0 - 4.75\ 10^{-1}$	$1.28\ 10^{-8}$
65	$1.0 - 1.95\ 10^{-172}$	$1.0 - 9.61\ 10^{-44}$	$1.0 - 9.85\ 10^{-10}$	$4.75\ 10^{-1}$	$4.59\ 10^{-9}$
91	$1.0 - 3.03\ 10^{-153}$	$1.0 - 3.49\ 10^{-33}$	$1.0 - 1.97\ 10^{-4}$	$4.62\ 10^{-4}$	$2.29\ 10^{-25}$
120	$1.0 - 3.43\ 10^{-133}$	$1.0 - 3.39\ 10^{-23}$	$1.0 - 2.11\ 10^{-1}$	$2.00\ 10^{-12}$	$2.85\ 10^{-54}$
121	$1.0 - 1.59\ 10^{-132}$	$1.0 - 6.94\ 10^{-23}$	$1.0 - 2.39\ 10^{-1}$	$8.18\ 10^{-13}$	$1.80\ 10^{-55}$
128	$1.0 - 6.56\ 10^{-128}$	$1.0 - 8.99\ 10^{-21}$	$1.0 - 4.81\ 10^{-1}$	$1.03\ 10^{-15}$	$3.04\ 10^{-64}$
129	$1.0 - 2.95\ 10^{-127}$	$1.0 - 1.76\ 10^{-20}$	$4.81\ 10^{-1}$	$3.74\ 10^{-16}$	$1.50\ 10^{-65}$
186	$1.0 - 7.36\ 10^{-93}$	$1.0 - 1.81\ 10^{-7}$	$2.77\ 10^{-8}$	$2.14\ 10^{-52}$	$1.89\ 10^{-162}$
247	$1.0 - 3.87\ 10^{-62}$	$1.0 - 2.46\ 10^{-1}$	$2.10\ 10^{-29}$	$1.72\ 10^{-115}$	$6.35\ 10^{-315}$
248	$1.0 - 1.09\ 10^{-61}$	$1.0 - 2.70\ 10^{-1}$	$7.21\ 10^{-30}$	$9.82\ 10^{-117}$	$7.64\ 10^{-318}$
256	$1.0 - 3.87\ 10^{-58}$	$1.0 - 4.86\ 10^{-1}$	$9.98\ 10^{-34}$	$6.24\ 10^{-127}$	$1.08\ 10^{-341}$
257	$1.0 - 1.05\ 10^{-57}$	$4.86\ 10^{-1}$	$3.16\ 10^{-34}$	$3.09\ 10^{-128}$	$9.85\ 10^{-345}$
377	$1.0 - 1.24\ 10^{-17}$	$1.71\ 10^{-18}$	$3.19\ 10^{-122}$	$4.66\ 10^{-334}$	$2.95\ 10^{-808}$
503	$1.0 - 2.76\ 10^{-1}$	$4.25\ 10^{-71}$	$1.36\ 10^{-274}$	$2.90\ 10^{-655}$	$3.13\ 10^{-1505}$
504	$1.0 - 2.98\ 10^{-1}$	$1.17\ 10^{-71}$	$4.76\ 10^{-276}$	$3.04\ 10^{-658}$	$1.26\ 10^{-1511}$
512	$1.0 - 4.88\ 10^{-1}$	$3.19\ 10^{-76}$	$7.86\ 10^{-288}$	$2.50\ 10^{-682}$	$2.94\ 10^{-1563}$
513	$4.88\ 10^{-1}$	$8.37\ 10^{-77}$	$2.54\ 10^{-289}$	$2.28\ 10^{-685}$	$8.97\ 10^{-1570}$
533	$1.00\ 10^{-1}$	$6.83\ 10^{-89}$	$6.15\ 10^{-320}$	$1.30\ 10^{-747}$	$6.26\ 10^{-1703}$

Table 2: Approximations for S_l ($l = 1, 2, 3, 4, 5$) for $b = 1024$, $p = 1/2$ and selected values for k.

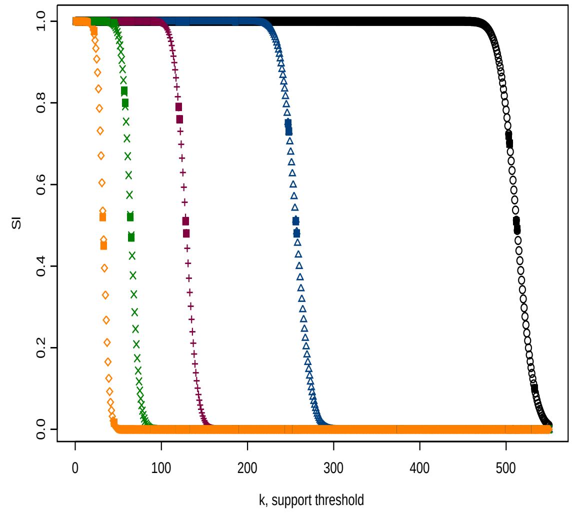

Figure 3: Exact values of S_l and approximations for $p = 1/2$ for all values of k.

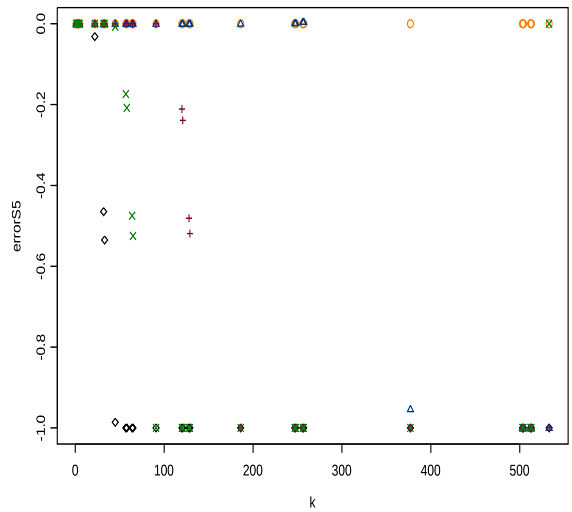

Figure 4: Relative error for the approximations for $p = 1/2$ for all values of k.

In Figure 3, the original values for S_l for selected values for k are plotted in massive cubes, together with their approximations. These selected values for k are the same as the values used in [PVGG04]. The rightmost curve (○) is for S_1, △ for

S_2, $+$ for S_3, $\times$ for S_4 and the leftmost curve ($\diamond$) is for S_5. The figure shows that the approximations are very close to the exact values.

To express the accuracy, some error computation is done. In Figure 4, the relative error is plotted. The relative error is computed as the absolute error (approximation minus exact value) divided by the exact value. The results for S_1 are again plotted as $\circ$, S_2 as $\triangle$, S_3 as $+$, S_4 as $\times$ and S_5 as $\diamond$.

If one l is fixed ($l = 1, 2, 3, 4, 5$), the approximation is more inaccurate when k increases. For increasing l, the approximations are getting worse for smaller and smaller values of k. Especially for $l = 4$ and $l = 5$ and big values of k the approximations are underestimating the real values. This happens because Φ in big negative values is close to zero.

4.1.2 $b = 1024$ and $p = 1/16$

$b > 30$
$P(l) = p^l$, so for every value of l ($l = 1, 2, 3, 4, 5$), $P(l)$ has a different value.

- $\boldsymbol{l = 1} \rightarrow P(l) = 1/16 < 0.1$ and $bP(l) = 64 \Rightarrow$ Section 3.2.2
- $\boldsymbol{l = 2} \rightarrow P(l) = 1/256 < 0.1$ and $bP(l) = 4 \Rightarrow$ Section 3.2.1
- $\boldsymbol{l = 3} \rightarrow P(l) = 1/4096 < 0.1$ and $bP(l) = 0.25 \Rightarrow$ Section 3.2.1
- $\boldsymbol{l = 4} \rightarrow P(l) = 1/65536 < 0.1$ and $bP(l) = 0.016 \Rightarrow$ Section 3.2.1
- $\boldsymbol{l = 5} \rightarrow P(l) = 1/1048576 < 0.1$ and $bP(l) = 9.5\ \ 10^{-7} \Rightarrow$ Section 3.2.1

The results can be found in Table 4. The exact values can be found in Table 3.

Example
Let us consider $l = 1$ and $k = 1$. In this case $P(1) = 1/16$ and the approach of Section 3.2.2 has to be followed. $bP(1) = 64$.

$$\begin{aligned} S_1 &= 1 - \Phi\left(\frac{(k-0.5) - bP(1)}{\sqrt{bP(1)}}\right) \\ &= 1 - \Phi\left(\frac{(1-0.5) - 64}{\sqrt{64}}\right) \\ &= 1 - \Phi(-7.94) \\ &= 1 - 1.03\ \ 10^{-15} \end{aligned}$$

When $l = 5$ and $k = 257$, $P(5) = 1/1048576$, the approach of Section 3.2.1 has to be followed. $bP(5) = 0.0009765625$.

$$\begin{aligned} S_5 &= 1 - F(k-1) \\ &= 1 - P_{0.0009765625}(256) \\ &= 1 - 2.69\ \ 10^{-1278} \end{aligned}$$

$\square$

k	S_1	S_2	S_3	S_4	S_5
1	$1.0 - 2.0\ 10^{-29}$	$1.0 - 1.8\ 10^{-2}$	$2.2\ 10^{-1}$	$1.6\ 10^{-2}$	$9.8\ 10^{-4}$
2	$1.0 - 1.4\ 10^{-27}$	$1.0 - 9.1\ 10^{-2}$	$2.6\ 10^{-2}$	$1.2\ 10^{-4}$	$4.8\ 10^{-7}$
3	$1.0 - 4.8\ 10^{-26}$	$7.6\ 10^{-1}$	$2.2\ 10^{-3}$	$6.3\ 10^{-7}$	$1.5\ 10^{-10}$
4	$1.0 - 1.1\ 10^{-24}$	$5.7\ 10^{-1}$	$1.3\ 10^{-4}$	$2.4\ 10^{-9}$	$3.8\ 10^{-14}$
5	$1.0 - 1.9\ 10^{-23}$	$3.7\ 10^{-1}$	$6.6\ 10^{-6}$	$7.6\ 10^{-12}$	$7.3\ 10^{-18}$
22	$1.0 - 1.5\ 10^{-10}$	$3.0\ 10^{-10}$	$3.2\ 10^{-35}$	$1.3\ 10^{-61}$	$4.2\ 10^{-88}$
32	$1.0 - 2.0\ 10^{-6}$	$1.0\ 10^{-18}$	$1.0\ 10^{-55}$	$3.7\ 10^{-94}$	$1.1\ 10^{-132}$
33	$1.0 - 4.3\ 10^{-6}$	$1.2\ 10^{-19}$	$7.3\ 10^{-58}$	$1.7\ 10^{-97}$	$3.1\ 10^{-137}$
45	$1.0 - 4.2\ 10^{-3}$	$9.2\ 10^{-32}$	$2.0\ 10^{-84}$	$1.6\ 10^{-142}$	$1.1\ 10^{-192}$
57	$8.3\ 10^{-1}$	$2.5\ 10^{-45}$	$1.9\ 10^{-112}$	$5.5\ 10^{-181}$	$1.3\ 10^{-249}$
58	$8.0\ 10^{-1}$	$1.7\ 10^{-46}$	$7.8\ 10^{-115}$	$1.4\ 10^{-184}$	$2.1\ 10^{-254}$
64	$5.2\ 10^{-1}$	$8.9\ 10^{-54}$	$2.5\ 10^{-129}$	$2.6\ 10^{-206}$	$2.3\ 10^{-283}$
65	$4.7\ 10^{-1}$	$5.1\ 10^{-55}$	$8.9\ 10^{-132}$	$5.9\ 10^{-210}$	$3.3\ 10^{-288}$
91	$5.8\ 10^{-4}$	$2.0\ 10^{-89}$	$1.6\ 10^{-197}$	$5.1\ 10^{-307}$	$1.4\ 10^{-416}$
120	$5.3\ 10^{-11}$	$5.6\ 10^{-132}$	$4.8\ 10^{-275}$	$1.9\ 10^{-419}$	$6.1\ 10^{-564}$
121	$2.6\ 10^{-11}$	$1.6\ 10^{-133}$	$8.7\ 10^{-278}$	$2.1\ 10^{-423}$	$4.3\ 10^{-569}$
128	$1.4\ 10^{-13}$	$2.3\ 10^{-144}$	$4.5\ 10^{-297}$	$4.1\ 10^{-451}$	$3.1\ 10^{-605}$
129	$6.6\ 10^{-14}$	$6.3\ 10^{-146}$	$7.7\ 10^{-300}$	$4.4\ 10^{-455}$	$2.1\ 10^{-610}$
186	$8.9\ 10^{-39}$	$7.9\ 10^{-241}$	$1.8\ 10^{-463}$	$2.4\ 10^{-687}$	$2.6\ 10^{-911}$
247	$1.2\ 10^{-75}$	$1.0\ 10^{-352}$	$6.7\ 10^{-649}$	$3.0\ 10^{-945}$	$1.2\ 10^{-1243}$
248	$2.4\ 10^{-76}$	$1.3\ 10^{-354}$	$5.1\ 10^{-652}$	$1.5\ 10^{-950}$	$3.5\ 10^{-1249}$
256	$7.1\ 10^{-82}$	$5.4\ 10^{-370}$	$4.9\ 10^{-677}$	$3.3\ 10^{-985}$	$1.8\ 10^{-1293}$
257	$1.4\ 10^{-82}$	$6.3\ 10^{-372}$	$3.6\ 10^{-680}$	$1.5\ 10^{-989}$	$5.2\ 10^{-1299}$

Table 3: Exact values for S_l for $b = 1024$, $p = 1/16$ and selected values of k.

k	S_1	S_2	S_3	S_4	S_5
1	$1.0 - 1.03\ 10^{-15}$	$1.0 - 1.83\ 10^{-2}$	$1.0 - 7.79\ 10^{-1}$	$1.0 - 9.84\ 10^{-1}$	$1.0 - 9.99\ 10^{-1}$
2	$1.0 - 2.80\ 10^{-15}$	$1.0 - 7.33\ 10^{-2}$	$1.0 - 1.95\ 10^{-1}$	$1.0 - 1.54\ 10^{-2}$	$1.0 - 9.76\ 10^{-4}$
3	$1.0 - 7.50\ 10^{-15}$	$1.0 - 1.47\ 10^{-1}$	$1.0 - 2.43\ 10^{-2}$	$1.0 - 1.20\ 10^{-4}$	$1.0 - 4.76\ 10^{-7}$
4	$1.0 - 1.98\ 10^{-14}$	$1.0 - 1.95\ 10^{-1}$	$1.0 - 2.03\ 10^{-3}$	$1.0 - 6.26\ 10^{-7}$	$1.0 - 1.55\ 10^{-10}$
5	$1.0 - 5.13\ 10^{-14}$	$1.0 - 1.95\ 10^{-1}$	$1.0 - 1.27\ 10^{-4}$	$1.0 - 2.45\ 10^{-9}$	$1.0 - 3.79\ 10^{-14}$
22	$1.0 - 5.41\ 10^{-8}$	$1.0 - 1.58\ 10^{-9}$	$1.0 - 3.47\ 10^{-33}$	$1.0 - 2.27\ 10^{-58}$	$1.0 - 1.19\ 10^{-83}$
32	$1.0 - 2.43\ 10^{-5}$	$1.0 - 1.03\ 10^{-17}$	$1.0 - 2.05\ 10^{-53}$	$1.0 - 1.22\ 10^{-90}$	$1.0 - 5.82\ 10^{-128}$
33	$1.0 - 4.12\ 10^{-5}$	$1.0 - 1.28\ 10^{-18}$	$1.0 - 1.60\ 10^{-55}$	$1.0 - 5.96\ 10^{-94}$	$1.0 - 1.78\ 10^{-132}$
45	$1.0 - 7.39\ 10^{-3}$	$1.0 - 2.13\ 10^{-30}$	$1.0 - 9.47\ 10^{-82}$	$1.0 - 1.25\ 10^{-134}$	$1.0 - 1.32\ 10^{-187}$
57	$1.0 - 1.74\ 10^{-1}$	$1.0 - 1.34\ 10^{-43}$	$1.0 - 2.11\ 10^{-109}$	$1.0 - 9.89\ 10^{-177}$	$1.0 - 3.72\ 10^{-244}$
58	$1.0 - 2.08\ 10^{-1}$	$1.0 - 9.39\ 10^{-45}$	$1.0 - 9.25\ 10^{-112}$	$1.0 - 2.71\ 10^{-180}$	$1.0 - 6.38\ 10^{-249}$
64	$1.0 - 4.75\ 10^{-1}$	$1.0 - 7.86\ 10^{-52}$	$1.0 - 4.62\ 10^{-126}$	$1.0 - 8.07\ 10^{-202}$	$1.0 - 1.13\ 10^{-277}$
65	$4.75\ 10^{-1}$	$1.0 - 4.91\ 10^{-53}$	$1.0 - 1.80\ 10^{-128}$	$1.0 - 1.79\ 10^{-205}$	$1.0 - 1.73\ 10^{-282}$
91	$4.62\ 10^{-3}$	$1.0 - 1.89\ 10^{-86}$	$1.0 - 3.42\ 10^{-193}$	$1.0 - 1.84\ 10^{-301}$	$1.0 - 7.96\ 10^{-410}$
120	$2.00\ 10^{-12}$	$1.0 - 1.45\ 10^{-127}$	$1.0 - 3.16\ 10^{-269}$	$1.0 - 2.05\ 10^{-412}$	$1.0 - 1.07\ 10^{-555}$
121	$8.18\ 10^{-13}$	$1.0 - 4.48\ 10^{-129}$	$1.0 - 6.59\ 10^{-272}$	$1.0 - 2.67\ 10^{-416}$	$1.0 - 8.67\ 10^{-561}$
128	$1.03\ 10^{-15}$	$1.0 - 1.76\ 10^{-139}$	$1.0 - 8.93\ 10^{-291}$	$1.0 - 1.35\ 10^{-443}$	$1.0 - 1.63\ 10^{-596}$
129	$3.74\ 10^{-16}$	$1.0 - 5.50\ 10^{-141}$	$1.0 - 1.74\ 10^{-293}$	$1.0 - 1.64\ 10^{-447}$	$1.0 - 1.24\ 10^{-601}$
186	$2.14\ 10^{-52}$	$1.0 - 1.07\ 10^{-231}$	$1.0 - 7.86\ 10^{-453}$	$1.0 - 1.72\ 10^{-675}$	$1.0 - 3.01\ 10^{-898}$
247	$1.72\ 10^{-115}$	$1.0 - 2.76\ 10^{-337}$	$1.0 - 7.18\ 10^{-632}$	$1.0 - 5.55\ 10^{-928}$	$1.0 - 3.45\ 10^{-1224}$
248	$9.82\ 10^{-117}$	$1.0 - 4.47\ 10^{-339}$	$1.0 - 7.27\ 10^{-635}$	$1.0 - 3.51\ 10^{-932}$	$1.0 - 1.36\ 10^{-1229}$
256	$6.24\ 10^{-127}$	$1.0 - 1.83\ 10^{-353}$	$1.0 - 6.93\ 10^{-659}$	$1.0 - 7.80\ 10^{-966}$	$1.0 - 7.05\ 10^{-1273}$
257	$3.09\ 10^{-128}$	$1.0 - 2.86\ 10^{-355}$	$1.0 - 6.77\ 10^{-662}$	$1.0 - 4.76\ 10^{-970}$	$1.0 - 2.69\ 10^{-1278}$

Table 4: Approximations for S_l ($l = 1, 2, 3, 4, 5$) for $b = 1024$, $p = 1/16$ and selected values of k.

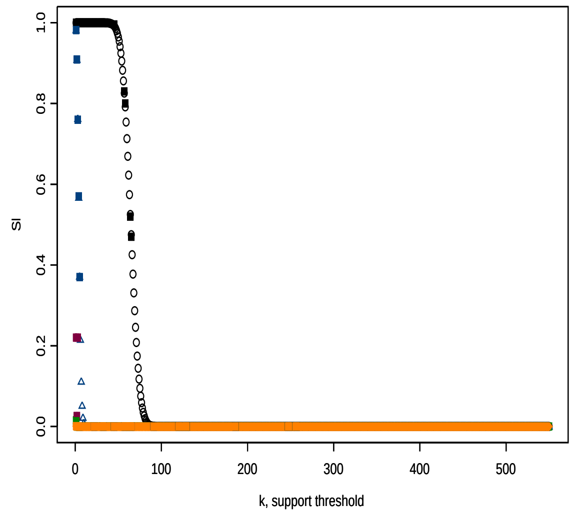

Figure 5: Exact values of S_l and approximations for $p = 1/16$.

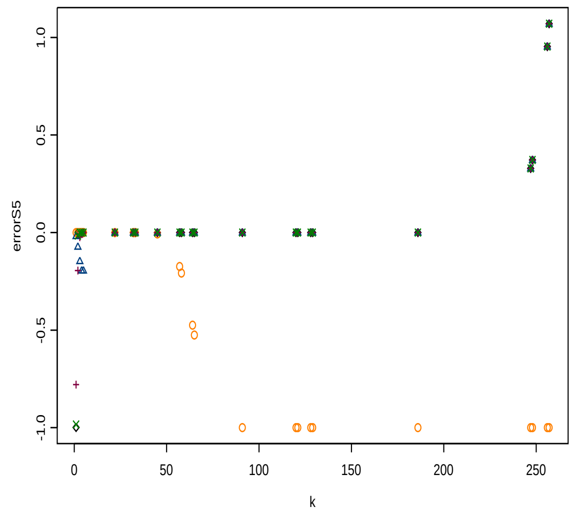

Figure 6: Relative error for approximations for $p = 1/16$.

In Figure 5, the original values for S_l for selected values for k, the values also used in [PVGG04], are plotted in massive cubes, together with their approximations. The curve $\circ$ is for S_1, $\triangle$ for S_2, $+$ for S_3, $\times$ for S_4 and $\diamond$ for S_5. As the figures show, the approximations are not that good.

The error computation in Figure 6 shows that the approximations are not so accurate. In the case that $l = 1$ the approximation is of the same quality as the aprroximations in the previous section. For $l = 2, 3, 4$ and 5, the Poisson approximation is used. For small values of k, this approximation is accurate, but for larger values of k, it overestimates the real value of S_l. It stays close to 1 while the real values decrease to zero.

Remark

In [PVGG04], they found that the values for S with $p = 1/2$ and $l = 4$ are approximately the same as S for $p = 1/16$ and $l = 1$, particularly when k is small. In our approximation we can see that these values are approximated by the same formula

$$1 - \Phi\left(\frac{(k - 0.5) - bP(l)}{\sqrt{bP(l)}}\right)$$

what in the two cases will lead to the same results.

4.2 New Approach versus Old Approach

In [PVGG04], there are no actual approximations computed for S_l but an upper and a lower bound for S_l is derived and these bounds are plotted together with the correct values for S_l. Their figures show that there is a gap between the exact values and each of the two bounds. This new approach does not suffer from this gap; it is more accurate than the introduced upper and lower bound.

Remark

When we look at the exact values for S_l in Tables 1 and 3 we can notice the following interesting things. When we fix a certain, moderate-sized value for k, S_l is close to 1 for small values of l and it is close to zero for large values of l. The transition from near 1 to near 0 is quite sharp with increasing l. The transition value of l increases when k decreases. For large k, even S_1 is near zero. For small values of k, l has to be large before S_l approaches zero.

5 Conclusion and Future Work

The first step in future work is trying to repeat the basic ideas of the analysis from this technical report for the other two important probabilities, C_l and F_l. These two probabilities are closely related, so if it is known how to treat C_l, the same can be done for F_l. By following the same kind of reasoning as with S_l, the most easy part in the approach of C_l is to find the distribution that C_l describes. This is a Multinomial Distribution, the direct multivariate generalization of the univariate Binomial Distribution. By extending this analoguous way of thinking, the challange is to find the approximating Multivariate Normal and Multivariate Poisson

Distributions. These distributions are the theoretical approximations, but are not so tractable in practice for computations. The multivariate analogon of Φ is a multivariate integral and is too hard to solve. The Multivariate Poisson is even more strange, starting from the definition by loosing one degree of freedom. The computation of the found formula is also very hard, because of the need of all the different partitions of the original transaction database obeying the conditions expressed in the formula for C_l. An important step in future work is considering in detail these two cases and trying to find a practical computation method.

The next possible topic is applying a new, more realistic and more complex model of shopping behaviour. The model used now to describe the shopping behaviour is very simple. It assumes that all the items are independent and have the same probability and that all the transactions are independent. In reality, none of these three conditions is satisfied, so our model is a strong simplification of the real world. First of all, it is not true that all the items have the same probability of being chosen. Some items are needed or wanted more than others, will therefore have bigger probabilities and will be bought more often. This brings us automatically to the next condition, independence of items. The assumption of independence of items is not true in the real world. One of the basic powers in shopping behaviour is that buying one item is influenced by buying or not buying another item(s), so the products are clearly positively or negatively correlated. A third aspect that has to be considered, is the fact that the behaviour of person i at time t is influenced by experiences that person had in the past. Loyalty to certain branches or satisfaction of a product play an important role in shopping behaviour. It is also possible that the buying pattern of person i is influenced by items that person saw in the basket of some other shopper j, shopping at the same time and place. These cases lead to transactions not being independent any more. In the future, there have to be taken care of these more realistic situations, keeping in mind that they won't simplify the analysis at all. Even the easiest step in the generalisation, jumping from the same probability for each item to a different probability for each item makes the formulas very complex.

References

[AIS93] R. Agrawal, T. Imilienski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD Int. Conf. Management of Data, pages 207-216, Washington D.C.*, 1993.

[AS94] R. Agrawal and R Srikant. Fast algorithms for mining association rules. *Proc. of the 1994 Very Large Data Bases Conference, pages 487-499*, 1994.

[DK95] H.G. Dehling and J.N. Kalma. *Kansrekening, het zekere van het onzekere*. Epsilon Uitgaven, Utrecht, 1995.

[Goe03] B. Goethals. Survey of frequent pattern mining. *http://www.adrem.ua.ac.be/ goethals/publications.html*, 2003.

[GZ03] B. Goethals and M. J. Zaki. Proc. of the workshop on frequent itemset mining implementations. *Melbourne, Florida*, 2003.

[HPY00] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *Proc. of the 2000 ACM SIGMOD Int. Conf. Management of Data, Dallas, TX. pages 1-12*, 2000.

[JK69] N.L. Johnson and S. Kotz. *Distributions in Statistics: Discrete Distributions*. Houghton Mifflin Company, Boston, 1969.

[OGD80] I. Olkin, L.J. Glese, and C. Derman. *Probability Models and Applications*. Macmillan Publishing Co., New York, Collier Macmillan Publishers, London, 1980.

[PVGG04] Paul W. Purdom, Dirk Van Gucht, and Dennis P. Groth. Average-case performance of the apriori algorithm. *SIAM J. Computing, vol. 33, No.5, pp. 1223-1260*, 2004.

[Zak00] M. J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering, 12(3), pages 372-390*, 2000.