A Appendix

A.1 Approximation Methods Section A.1.1 outlined how the probability for a certain support can be calculated exactly. This procedure can be infeasible for large datasets. Therefor we describe two possible approximation methods. We are given the Poisson binomial distribution as defined in the problem statement $X_{sup}(p) = \sum_i^n X_i(p)$ with each $X_i(p)$ describing an independent Bernoulli trial with probability $\mathbf{P}_i(p)$. The goal is to find $P[X_{sup}(p) \geq \text{SUP}(p)]$. Note that $\text{SUP}(p) > 0$ because we assume only patterns that occur are analyzed.

Before describing each method we will outline the advantages and disadvantages of each method and give an intuition as to which method is a good choice for different requirements and datasets. The three methods are the following:

1. The exact calculation of $P[X_{sup}(p) \geq \text{SUP}(p)]$ (described in section A.1.1). Aside from any rounding errors due to floating-point calculations, the resulting probabilities are exact. This computation can be done in $O(|\mathbf{P}(p)| \times \text{SUP}(p))$ time and, when calculated in a streaming fashion over the sequences, $O(\text{SUP}(p))$ space for a single pattern. Here, $|\mathbf{P}(p)|$ denotes the number of sequences in which all symbols of the pattern occur.

2. We also describe an approximation using a normal distribution $X_{sup}^{norm}(p) \overset{d}{\approx} X_{sup}(p)$ (described in section A.1.2). This method is most precise when probabilities are not very small and patterns occur often, as would be the case with shorter patterns which repeat often. This approximation has an $O(|\mathbf{P}(p)|)$ time complexity and an $O(1)$ space complexity when performed in a streaming fashion over the sequences.

3. Lastly, we describe an approximation using a Poisson distribution $X_{sup}^{Pois}(p) \overset{d}{\approx} X_{sup}(p)$ (described in section A.1.3). This method is most precise when both probability values and support values are very small, as would be the case with long patterns that rarely occur. To illustrate the lower probabilities involved with longer patterns, a pattern of length 10 where each item occurs once in a sequence has a probability of $\frac{1}{10!} < 10^{-6}$ to occur when the sequence is randomly permuted. This approximation method has an $O(|\mathbf{P}(p)|+\text{SUP}(p))$ time complexity and an $O(1)$ space complexity.

Figure 2 illustrates the difference in precision between the two approximation methods. Data is generated randomly where each item of the pattern occurs randomly in each sequence between 1 and 5 times. This leads to no probabilities of zero in any of the sequences. On this data, the probability values are then calculated using the exact method and each approximation method. The difference between the two is the absolute error which is shown in the figures. Figure 2a was created by running both approximation methods on short patterns ($2 \leq |p| \leq 4$). The error of the normal approximation is consistently lower. Figure 2b shows results for longer patterns ($7 \leq |p| \leq 10$) and a larger number of sequences ($|\mathbf{P}(p)| = 1000$). Here we see the Poisson approximation outperforming the normal approximation. These figures show the general guideline of which approximation is better in which case.

A.1.1 Exact Calculation The example in section 3 described briefly how we can calculate the desired probability $P[X_{sup}(p) \geq \text{SUP}(p)]$. Because $X_{sup}(p)$ is a Poisson binomial distribution it consists of the Bernoulli trials $X_i(p)$. We are interested in the probability of at least $\text{SUP}(p)$ of these Bernoulli trials having a successful outcome. The cases where this is true are those situations in which the pattern occurs in at least $\text{SUP}(p)$ sequences or has a support-value greater than or equal to $\text{SUP}(p)$. The well-read reader might see the similarity between our case and calculating the probabilistic support in uncertain transaction databases as introduced by Chui et al. [3]. Bernecker et al. [1] describe how to compute the exact expected support of an itemset in this case. We can use the same method based on dynamic programming to compute the expected support for our sequential patterns. For details we refer to [1].

A.1.2 Normal Distribution Approximation Calculating the probability $P[X_{sup}(p) \geq \text{SUP}(p)]$ as shown in section A.1.1 can be infeasible for very large datasets. We use Lyapunov's version of the central limit theorem [2, p. 359] to prove we can approximate our probability using the normal distribution. The algorithm can be trivially derived from the formulas.

Lyapunov's condition states: Let $X_1, X_2, ..., X_n$ be independent random variables, each with expected value μ_i and variance σ_i^2. If for some $\delta > 0$ the following holds

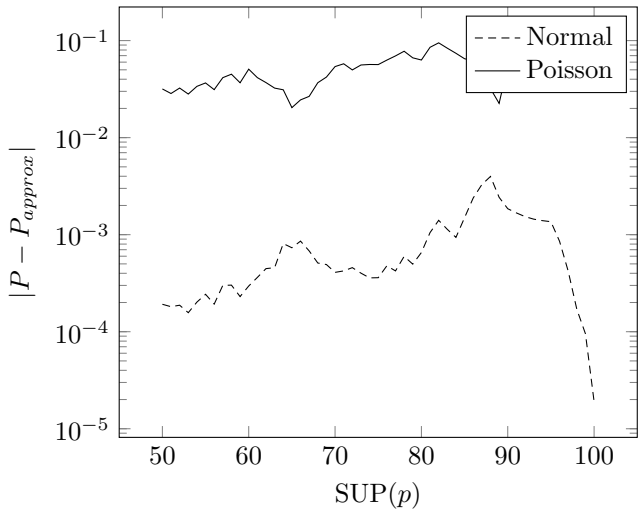
$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E\left[|X_i - \mu_i|^{2+\delta}\right]}{(\sum_{i=1}^n \sigma_i^2)^\delta} = 0$$

then it holds that

$$\sum_i X_i \overset{d}{\approx} N\left(\sum_i \mu_i, \sum_i \sigma_i\right) \Rightarrow \frac{\sum_i X_i - \mu_i}{\sqrt{\sum_i \sigma_i^2}} \sim N(1,0)$$

To make this condition valid for our case, we need to add the assumption that $\exists \epsilon > 0 : \forall n : \exists n' > n : \sigma_{n'} > \epsilon$,

(a) $2 \leq |p| \leq 4$, $50 \leq \text{SUP}(p) \leq 100$ and $|\mathbf{S}| = |\mathbf{P}| = 100$.



(b) $7 \leq |p| \leq 10$, $1 \leq \text{SUP}(p) \leq 10$ and $|\mathbf{S}| = |\mathbf{P}| = 1000$.

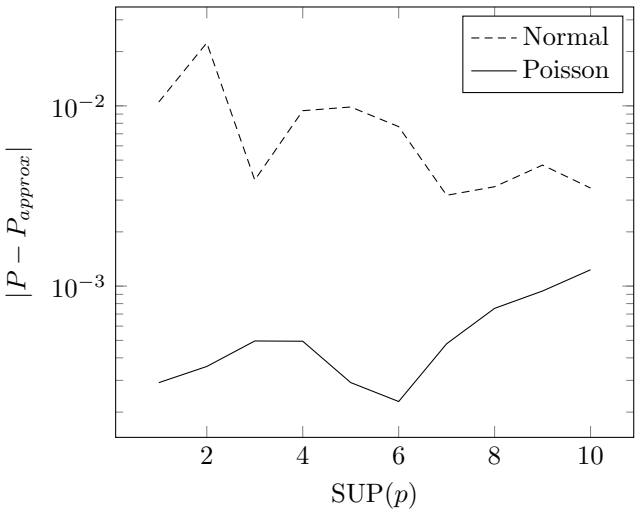


Figure 2: The precision of both approximation methods compared to the exact p-values. Each data point is an average of 100 samples. Preferable datasets were created for each method by changing the pattern length ($|p|$), support ($\text{SUP}(p)$) and dataset size ($|\mathbf{S}|$).

which means items in a pattern have to keep occurring, i.e. there is no decrease in the occurrence the items of a pattern over time. More specifically, it can't decrease so much that the items stop occurring altogether. This requirement can be combined with the requirement that enough data needs to be available. When there is, we can approximate the probability of a pattern occurring a certain number of times.

THEOREM A.1. *Given the problem setting in 3, the criteria for Lyapunov's CLT hold and thus an approximation using the normal distribution is possible.*

Proof.

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E\left[|X_i - \mu_i|^{2+\delta}\right]}{(\sum_{i=1}^{n} \sigma_i^2)^{\delta}} \\
&\leq \lim_{n\to\infty} \frac{\sum_{i=1}^{n} E\left[|X_i - \mu_i|^{2}\right]}{(\sum_{i=1}^{n} \sigma_i^2)^{\delta}} \\
&= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \sigma_i^2}{(\sum_{i=1}^{n} \sigma_i^2)^{\delta}} \qquad\qquad \sigma^2 = E\left[|X_i - \mu_i|^2\right] \\
&= \lim_{n\to\infty} (\sum_{i=1}^{n} \sigma_i^2)^{-\delta} \;=\; 0
\end{aligned}$$

□

This means we can approximate our Poisson binomial distribution $X_{sup}(p)$ using the normal distribution as such:

$$\begin{aligned}
X_{sup}(p) &\overset{d}{\approx} X_{sup}^{norm}(p) \\
&\overset{d}{=} \text{N}\left(\sum_i \mathbf{P}_i(p), \sum_i \mathbf{P}_i(p)(1 - \mathbf{P}_i(p))\right)
\end{aligned}$$

This lets us approximate the p-value as such:

$$\begin{aligned}
P[X_{sup}(p) \geq \text{SUP}(p)] &\approx P[X_{sup}^{norm}(p) \geq \text{SUP}(p)] \\
&= 1 - \Phi\left(\frac{\text{SUP}(p) - \frac{1}{2} - \sum_i \mathbf{P}_i(p)}{\sqrt{\sum_i \mathbf{P}_i(p)(1 - \mathbf{P}_i(p))}}\right)
\end{aligned}$$

Note the $-\frac{1}{2}$ term, we use this because we are approximating a discrete random variable with a continuous one.

To illustrate this method we will apply it to the example in section 3. We first calculate the expected value and variance for the support of both patterns. These can then be used to calculate the probabilities as such:

$$\begin{aligned}
P[X_{sup}(abc) \geq \text{SUP}(abc)] &\approx 1 - \Phi\left(\frac{5 - 0.5 - 4.841}{\sqrt{2.438}}\right) \\
&= 0.586 \\
P[X_{sup}(ade) \geq \text{SUP}(ade)] &\approx 1 - \Phi\left(\frac{5 - 0.5 - 2.250}{\sqrt{1.729}}\right) \\
&= 0.044
\end{aligned}$$

Note the error in this example:

$$\left|P[X_{sup}(abc) \geq \text{SUP}(abc)] - P[X_{sup}^{norm}(abc) \geq \text{SUP}(abc)]\right| = 0.002$$

$$\left|P[X_{sup}(ade) \geq \text{SUP}(ade)] - P[X_{sup}^{norm}(ade) \geq \text{SUP}(ade)]\right| = 0.008$$

A.1.3 Poisson Distribution Approximation Another method of approximating the support probability is using the Poisson distribution. The algorithm for this can be trivially derived from the formulas. We can prove that this approximation is possible using Le Cam's inequality [10] which states:

$$\sum_{i=0}^{\infty} \left| P[X = \sup(p)] - \frac{e^{-\lambda}\lambda^k}{k!} \right| < 2\sum_{i=1}^{n} \mathbf{P}_i^2$$

with $\lambda = \|\mathbf{P}\|_1$. Since the distribution $X_{sup}(p) = \sum_i X_i(p)$ is a sum of random variables $X_i(p)$ which have a Bernoulli distribution, this inequality holds for our case. This means we can approximate our Poisson binomial distribution using the Poisson distribution as such:

$$X_{sup}(p) \overset{d}{\approx} X_{sup}^{Pois}(p) \overset{d}{=} \text{Pois}\left(\|\mathbf{P}(p)\|_1\right)$$

This lets us approximate the p-value as such:

$$\begin{aligned} P[X_{sup}(p) \geq \text{SUP}(p)] &\approx P[X_{sup}^{Pois}(p) \geq \text{SUP}(p)] \\ &= 1 - e^{-\lambda_p} \sum_{i=0}^{\text{SUP}(p)-1} \frac{\lambda_p^i}{i!} \end{aligned}$$

with $\lambda_p = \|\mathbf{P}(p)\|_1$, i.e. the sum of the probabilities for pattern p for all sequences.

To illustrate this method we will again apply it to the earlier example.

$$\begin{aligned} P[X_{sup}(abc) \geq \text{SUP}(abc)] &\approx 1 - e^{-\lambda_{abc}} \sum_{i=0}^{\text{SUP}(p)-1} \frac{\lambda_{abc}^i}{i!} \\ &= 0.531 \end{aligned}$$

$$\begin{aligned} P[X_{sup}(ade) \geq \text{SUP}(ade)] &\approx 1 - e^{-\lambda_{ade}} \sum_{i=0}^{\text{SUP}(p)-1} \frac{\lambda_{ade}^i}{i!} \\ &= 0.078 \end{aligned}$$

Note the error in this example:

$$\left|P[X_{sup}(abc) \geq \text{SUP}(abc)] - P[X_{sup}^{Pois}(abc) \geq \text{SUP}(abc)]\right| = 0.053$$

$$\left|P[X_{sup}(ade) \geq \text{SUP}(ade)] - P[X_{sup}^{Pois}(ade) \geq \text{SUP}(ade)]\right| = 0.027$$

A.2 Approximation Analysis In this section, we investigate the impact of using PS^2 with the approximation methods described earlier. There are two important factors to investigate. First, we will investigate the precision of the approximations by comparing the probabilities assigned by the approximation methods and comparing them to the true probabilities. Then, we will look at how well the approximation methods maintain the order. In many cases, exact probabilities aren't the most important result of PS^2. What is more important is the ranking of patterns, when given a large set of patterns, knowing which ones are most significant. If the exact probabilities are needed for those few patterns deemed significant, a second run of the algorithms calculating exact probabilities could be done.

Table 5 shows the top 10 patterns using each method. It can be seen that the probabilities assigned to each pattern vary significantly. This is partly because these datasets are relatively small. As the derivations of the approximation methods show, theoretically, the results will improve when more data becomes available. This is further illustrated in figure 3. The true probabilities are plotted against the approximated ones. Note this is a log-log plot to enlarge the difference of small probabilities. A perfect result would mean all approximations fall on the shown line, where each approximation is exactly the true probability.

Next, we look at how well each of the approximation methods maintains the ordering of all sequential patterns. We do this by calculating Spearman's rank correlation coefficient on the list of patterns as given by the exact calculation and each approximation method. This value will tend to one as the rankings of the patterns become more similar. Table 6 shows these results. It can be seen that the normal approximation consistently performs better than the Poisson approximation. This is partly because the patterns found are fairly short, meaning the probabilities are generally fairly high, which is where the normal approximation works better than the Poisson approximation.

Exact (12s)		Normal Approximation (12s)		Poisson Approximation (11s)	
van der waal	22.8	van der waal	31.2	point view	16.5
point view	22.7	professor van hoff	28.7	cathod rai	14.3
professor van hoff	21.0	professor van hoff solut	26.9	relat between	13.3
long time	17.7	principl conserv energi	20.8	long time	12.8
principl conserv energi	17.0	point view	17.8	magnet field	12.5
cathod rai	15.9	last few year	17.2	van der waal	10.7
magnet field	14.8	principl conserv energi mechan	16.8	professor van hoff	9.91
relat between	14.8	histori wireless telegraphi	15.0	between two	9.85
between two	12.5	between two point	14.3	principl conserv energi	9.80
last few year	12.5	long time	14.2	much more	9.00

Table 5: The New Physics and Its Evolution. Numbers behind each pattern signify $-log(P[X_{sup}(p) \geq \text{SUP}(p)])$. Run-times are reported between brackets in seconds.

Dataset	Exact	Normal Approximation	Poisson Approximation
The New Physics and Its Evolution	1 (12s)	0.99956 (12s)	0.98317 (11s)
A Book About Lawyers	1 (21s)	0.99937 (19s)	0.98840 (20s)
Adventures of Huckleberry Finn	1 (5s)	0.99990 (5s)	0.99578 (6s)
A Tale of Two Cities	1 (1s)	0.99961 (1s)	0.99318 (1s)

Table 6: The Spearman's rank correlation coefficient of the patterns in several datasets as ranked by the exact method and both approximation methods.

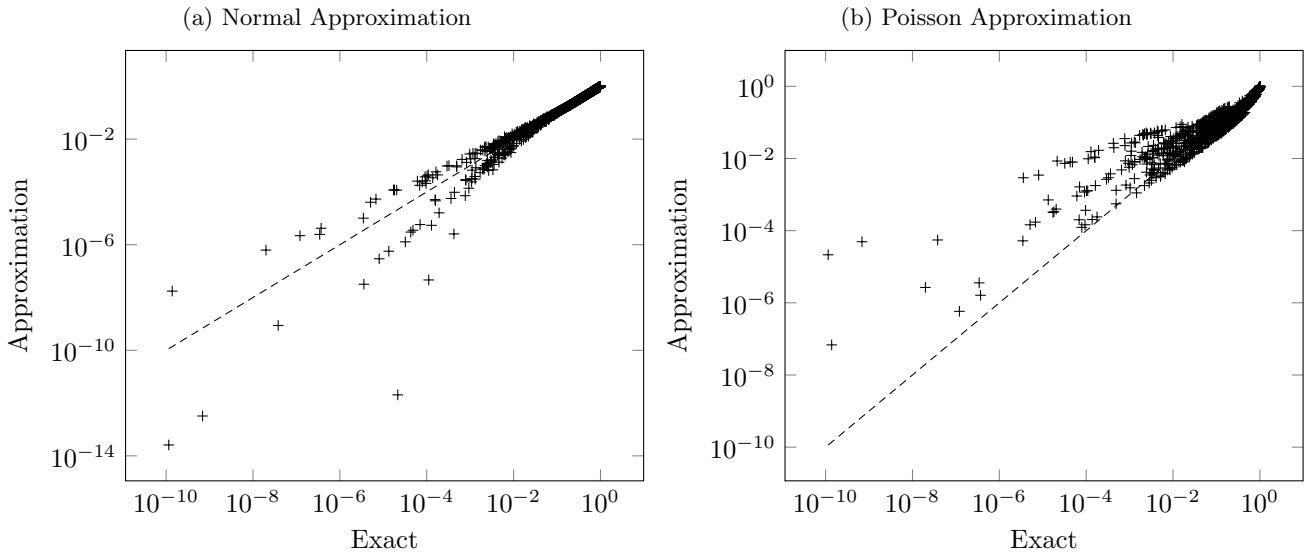


Figure 3: Exact $P[X_{sup}(p) \geq \text{SUP}(p)]$ versus approximations

A.3 Proof of Lemma 4.1

Proof. We first prove that removing symbols from s that do not occur in p does not change $\mathbf{P}_s(p)$: let a be an item that appears k times in s but that does not appear in p. Let s' be the sequence obtained by removing all occurrences of a from s. Then $|\pi(s)| = \binom{|s'|+k-1}{k}|\pi(s')|$, as there are exactly $\binom{|s'|+k-1}{k}$ ways to insert k occurrences of item a in sequence s'. Similarly, $|\{r \in \pi(s)|p \preceq r\}| = \binom{|s'|+k-1}{k}|\{r \in \pi(s')|p \preceq r\}|$ and hence $\mathbf{P}_s(p) = \frac{\binom{|s'|+k-1}{k}}{\binom{|s'|+k-1}{k}}\mathbf{P}_s(p) = \mathbf{P}_{s'}(p)$.

Secondly, it is clear that if we reorder the items in s without changing the number of occurrences of the items, $\mathbf{X}(s,p)$ does not change as we consider all permutations of s which is invariant under permutation.

Let now s^1, s^2 be two sequences and p^1, p^2 two patterns such that $\mathbf{X}(s^1,p^1) = \mathbf{X}(s^2,p^2)$. Because the count vectors have equal length (they are even exactly equal), and neither p^1, nor p^2 contain repetitions of items, we know $|p^1| = |p^2| = m$. Let $s^{i\prime}$ be s^i with all symbols not in p^i removed, and the items reordered in the order of p^i, $i = 1, 2$. Then we have as well $\mathbf{X}(s^{1\prime},p^1) = \mathbf{X}(s^{2\prime},p^2)$. Due to the reordering and the removal of items not in p^i, $s^{i\prime}$ will be exactly $\mathbf{X}(s^1,p^1)_1$ copies of item p^1_1 followed by $\mathbf{X}(s^1,p^1)_2$ copies of item p^1_2, and so on. Because $\mathbf{X}(s^1,p^1) = \mathbf{X}(s^2,p^2)$ this implies that the replacements p^2_i/p^1_i, $i = 1 \dots m$ turns $s^{1\prime}$ into $s^{2\prime}$ and p_1 into p_2. This implies that there is a 1-1 correspondence between on the one hand $\pi(s^{1\prime})$ and $\pi(s^{2\prime})$, and on the other hand between $\frac{|\{r\in\pi(s^{1\prime})|p_1\preceq r\}|}{|\pi(s^{1\prime})|}$ and $\frac{|\{r\in\pi(s^{2\prime})|p_2\preceq r\}|}{|\pi(s^{2\prime})|}$. Therefore $\mathbf{P}_{s^1}(p^1) = \mathbf{P}_{s^{1\prime}}(p^1) = \mathbf{P}_{s^{2\prime}}(p^2) = \mathbf{P}_{s^2}(p^2)$. ∎