Formal Verification of Tail Distribution Bounds in the HOL Theorem Prover

Osman Hasan and Sofiène Tahar

Department of Electrical and Computer Engineering,

Concordia University, Montreal, Canada

Email: {o_hasan, tahar}@ece.concordia.ca

Abstract

Tail distribution bounds play a major role in the estimation of failure probabilities in performance and reliability analysis of systems. They are usually estimated using the Markov and Chebyshev's inequalities, which represent tail distribution bounds for a random variable in terms of its mean or variance. This paper presents the formal verification of Markov's and Chebyshev's inequalities for discrete random variables using a higher-order-logic theorem prover (HOL). The paper also provides the formal verification of mean and variance relations for some of the widely used discrete random variables, such as Uniform(m), Bernoulli(p), Geometric(p) and Binomial(m, p) random variables. This infrastructure allows us to precisely reason about the tail distribution properties and thus turns out to be quite useful for the analysis of systems used in safety-critical domains, such as space, medicine or transportation. For illustration purposes, we present the performance analysis of the Coupon Collector's problem, a well known commercially used algorithm.

Keywords: Higher-Order-Logic, Mechanization of Proofs, Probabilistic Analysis of Algorithms, Probability Theory, Theorem Proving.
Mathematics Subject Classification: 03B15, 03B35, 60A05, 68T15.

1 Introduction

Probability theory is a tool of fundamental importance in the areas of performance and reliability analysis. The random and unpredictable elements, found in a system that needs to be analyzed, are mathematically modeled by appropriate random variables and performance

and reliability issues are then judged based on the corresponding probabilistic properties. Statistical characteristics, such as mean and variance, are the major decision making factors as they tend to summarize the distribution functions of random variables as single numbers that can be compared easily. During performance and reliability analysis while looking at the failure rates of a system, it is often the case that we are interested in the probability that a random variable assumes values that are far from its expectation or mean value. Instead of characterizing this probability by a distribution function, it is a common practice to rely upon bounds on this distribution, termed as tail distribution bounds, which are usually calculated using the Markov's or the Chebyshev's inequalities [1].

The Markov's inequality gives an upper bound for the probability that a non-negative random variable X is greater than or equal to some positive constant

$$Pr(X \geq a) \leq \frac{Ex[X]}{a} \tag{1}$$

where Pr and Ex denote the probability and expectation functions, respectively. Markov's inequality gives the best tail bound possible, for a nonnegative random variable, using the expectation for the random variable only [2]. This bound can be improved upon if more information about the distribution of random variable is taken into account. Chebyshev's inequality is based on this principle and it presents a significantly stronger tail bound in terms of variance of the random variable

$$Pr(|X - Ex[X]| \geq a) \leq \frac{Var[X]}{a^2} \tag{2}$$

where Var denotes the variance function. The Chebyshev's inequality allows us to bound the deviation of the random variable from its expectation and it can be calculated using the random variable's mean and variance only. Due to the widespread interest in failure probabilities and the ease of calculation of tail distribution bounds using Equations 1 and 2, Markov and Chebyshev's inequalities have now become one of the core techniques in modern probabilistic analysis.

Today, simulation is the most commonly used computer based probabilistic analysis technique. Most simulation softwares provide a programming environment for defining functions that approximate random variables for probability distributions. The random or unpredictable elements in a given system are modeled by these functions and the system is analyzed using computer simulation techniques, such as the Monte Carlo method [3], where the main idea is to approximately answer a query on a probability distribution by analyzing a large number of samples. Statistical quantities, such as mean and variance, and tail distribution bounds may then be calculated, based on the data collected during the sampling

process, using their mathematical relations in a computer. Due to the inaccuracies introduced by computer arithmetic operations and the inherent nature of simulation techniques, the simulation based probabilistic analysis results can never be termed as 100% accurate. McCullough [4, 5] proposed a collection of intermediate-level tests for assessing the numerical reliability of simulation based probabilistic analysis tools and uncovered flaws in some of the mainstream statistical packages. This inaccuracy poses a serious problem in highly sensitive and safety critical applications, such as space travel, medicine or transportation, where a mismatch between the predicted and the actual system performance may result in either inefficient usage of the available resources or paying higher costs to meet some performance or reliability criteria unnecessarily. Besides the inaccuracy of the results, another major limitation of simulation based probabilistic analysis is the enormous amount of CPU time requirement for attaining meaningful estimates. This approach generally requires hundreds of thousands of simulations to calculate the probabilistic quantities and becomes impractical when each simulation step involves extensive computations.

In order to overcome the limitations of the simulation based approaches, it has been proposed in [6] to conduct probabilistic analysis in a higher-order logic interactive theorem prover HOL [7]. Higher-order logic is a system of deduction with a precise semantics and can be used for the development of almost all classical mathematics theories. Interactive theorem proving is the field of computer science and mathematical logic concerned with computer based formal proof tools that require some sort of human assistance. Both discrete [8] and continuous [9] random variables can be formalized in higher-order-logic and their probabilistic and statistical characteristics, such as mean and variance, can be verified using an interactive theorem prover [6, 10]. Due to the inherent soundness of this approach, the probabilistic analysis carried out in this way is capable of providing exact answers. In order to be able to formally reason about tail distribution properties, we outlined an approach in [11] that allows us to formalize and verify the Markov's and Chebyshev' inequalities for discrete random variables in HOL. In the current paper, we mainly extend upon this approach and present the HOL proof steps in detail for the verification of Markov's and Chebyshev' inequalities. We also verify the mean and variance relations for the widely used discrete random variables: Uniform(m), Bernoulli(p), Geometric(p) and Binomial(m, p), in HOL. Thus, the main contribution of this paper is to extend the HOL libraries for probabilistic analysis with the ability to precisely reason about tail distribution bounds and thus enhance the capabilities of HOL as a successful probabilistic analysis framework.

In order to illustrate the practical effectiveness of the formalization presented in this paper, we utilize the above results to conduct the performance analysis of the Coupon Collector's problem [2], which is a well known commercially used algorithm in computer science, in HOL. Coupon Collector's problem is motivated by "*collect all n coupons and*

win" contests. The problem is to find the number of trials that we need to find all the n coupons, assuming that a coupon is drawn independently and uniformly at random from n possibilities. We first present a formalization of the Coupon Collector's problem using the Geometric random variable. Using this model, we illustrate the process of formally reasoning about the tail distribution properties of the Coupon Collector's problem using the formally verified mean and variance relations along with the Markov's and Chebyshev's inequalities, in HOL.

The rest of the paper is organized as follows. Section 2 gives a review of the related work. In Section 3, we provide some preliminaries including a brief introduction to the HOL theorem prover and an overview of modeling random variables and verifying their probabilistic and statistical properties in HOL. Next, we present the HOL formalization and verification of the Markov's and the Chebyshev's inequalities for discrete random variables in Section 4. The results are found to be in good agreement with existing theoretical paper-and-pencil counterparts. Then, we present the verification of mean and variance relations for some commonly used discrete random variables in Section 5. The analysis of the Coupon Collector's problem is presented in Section 6. Finally, Section 7 concludes the paper.

2 Related Work

Nędzusiak [12] and Bialas [13] were among the first ones to formalize some probability theory in higher-order-logic. Hurd [8] extended their work and developed a framework for the verification of probabilistic algorithms in the HOL theorem prover. He demonstrated the practical effectiveness of his formal framework by successfully verifying the sampling algorithms for four discrete probability distributions, some optimal procedures for generating dice rolls from coin flips, the symmetric simple random walk and the Miller-Rabin primality test based on the corresponding probability distribution properties. Hurd *et. al* [14] also formalized the *probabilistic guarded-command language* ($pGCL$) in HOL. The $pGCL$ contains both demonic and probabilistic nondeterminism and thus makes it suitable for reasoning about distributed random algorithms. Celiku [15] built upon the formalization of the $pGCL$ to mechanize the quantitative Temporal Logic (qtl) and demonstrated the ability to verify temporal properties of probabilistic systems in HOL. An alternative method for probabilistic verification in higher-order logic has been presented by Audebaud *et. al* [16]. Instead of using the measure theoretic concepts of probability space, as is the case in Hurd's approach, Audebaud *et. al* based their methodology on the monadic interpretation of randomized programs as probabilistic distribution. This approach only uses functional and algebraic properties of the unit interval and has been successfully used to verify a sampling algorithm

of the Bernoulli distribution and the termination of various probabilistic programs in the Coq theorem prover.

Building upon Hurd's formalization framework [8], we have been able to successfully verify the sampling algorithms of a few continuous random variables [9] and the classical *Cumulative Distribution Function* (CDF) properties [17], which play a vital role in verifying arbitrary probabilistic properties of both discrete and continuous random variables. The sampling algorithms for discrete random variables are either guaranteed to terminate or they satisfy probabilistic termination, meaning that the probability that the algorithm terminates is 1. Thus, they can be expressed in HOL by either well formed recursive functions or the *probabilistic while loop* [8]. On the other hand, the implementation of continuous random variables requires non-terminating programs and hence calls for a different approach. In [9], we presented a methodology that can be used to formalize any continuous random variable for which the inverse of the CDF can be expressed in a closed mathematical form. The core components of our methodology are the Standard Uniform random variable and the Inverse Transform method [18], which is a well known nonuniform random generation technique for generating nonuniform random variates for continuous probability distributions for which the inverse of the CDF can be represented in a closed mathematical form. Using the formalized Standard Uniform random variable and the Inverse Transform method, we were able to formalize continuous random variables, such as Exponential, Rayleigh, etc. and verify their correctness by proving the corresponding CDF properties in HOL.

The formalization, mentioned so far, allows us to express random behaviors as random variables in a higher-order-logic theorem prover and verify the corresponding quantitative probability distribution properties, which is a significant aspect of a probabilistic analysis framework. With the probability distribution properties of a random variable, such as the *Probability Mass Function* (PMF) and the CDF, we are able to completely characterize the behavior of their respective random variables. Though for comparison purposes, it is frequently desirable to summarize the characteristic of the distribution of a random variable by a single number, such as its expectation or variance, rather than an entire function. For example, it is more interesting to find out the expected value of the runtime of an algorithm for an NP-hard problem, rather than the probability of the event that the algorithm succeeds within a certain number of steps. In [6, 10], we tackled the verification of mean and variance in HOL for the first time. We extended Hurd's formalization framework with a formal definition of expectation, which can be utilized to formalize and verify the mean and variance characteristics associated with discrete random variables that attain values in positive integers only. In the current paper, we take the HOL probabilistic analysis framework further ahead by presenting the verification of Markov and Chebyshev's inequalities, which allows us to verify tail distribution bounds in HOL and is thus a novelty that has not been

available so far.

Besides theorem proving, another formal approach that is capable of providing exact solutions to probabilistic properties is probabilistic model checking [19, 20]. The most promising feature of probabilistic model checking is the ability to perform the analysis automatically. On the other hand, it is limited to systems that can only be expressed as a probabilistic finite state machine. In contrast, the theorem proving based probabilistic verification is an interactive approach but is capable of handling all kinds of probabilistic systems including the *unbounded* ones. Similarly, to the best of our knowledge, it is not possible to precisely evaluate statistical quantities, such as mean or variance, and tail distribution bounds, using probabilistic model checking so far. The most that has been reported in this domain is the approximate evaluation of mean values. Some probabilistic model checkers, such as PRISM [21] and VESTA [22], offer the capability of verifying expected values in a semi-formal manner. For example, in the PRISM model checker, the basic idea is to augment probabilistic models with cost or rewards: real values associated with certain states or transitions of the model. This way, the expected value properties, related to these rewards, can be analyzed by PRISM. The expectation values computed are expressed in a computer based notation, such as fixed or floating point numbers, which introduces some degree of approximation in the results. Similarly, the meaning ascribed to expected properties is, of course, dependent on the definitions of the rewards themselves and thus there is always some risk of verifying false properties. On the other hand, the proposed theorem proving based approach allows us to formally verify the statistical quantities, such as mean or variance, or tail distribution bounds related to the random variables without suffering from the above mentioned issues. Another major limitation of the probabilistic model checking approach is the state space explosion [23], which is not an issue with the proposed theorem proving based probabilistic analysis approach.

3 Preliminaries

In this section, we provide an overview of the HOL theorem prover and of modeling random variables and verifying their probabilistic and statistical properties in HOL. The intent is to provide a brief introduction to these topics along with some notation that is going to be used in the next sections.

3.1 HOL Theorem Prover

The HOL theorem prover, developed at the University of Cambridge, UK, is an interactive theorem prover which is capable of conducting proofs in higher-order logic. It utilizes the simple type theory of Church [24] along with Hindley-Milner polymorphism [25] to implement higher-order logic. HOL has been successfully used as a verification framework for both software and hardware as well as a platform for the formalization of pure mathematics. It supports the formalization of various mathematical theories including sets, natural numbers, *real* numbers, measure and probability. The HOL theorem prover includes many proof assistants and automatic proof procedures. The user interacts with a proof editor and provides it with the necessary tactics to prove goals while some of the proof steps are solved automatically by the automatic proof procedures.

In order to ensure secure theorem proving, the logic in the HOL system is represented in the strongly-typed functional programming language ML [26]. The ML abstract data types are then used to represent higher-order-logic theorems and the only way to interact with the theorem prover is by executing ML procedures that operate on values of these data types. Users can prove theorems using a natural deduction style by applying inference rules to axioms or previously generated theorems. The HOL core consists of only 5 basic axioms and 8 primitive inference rules, which are implemented as ML functions. Soundness is assured as every new theorem must be created from these basic axioms and primitive inference rules or any other pre-existing theorems/inference rules.

We selected the HOL theorem prover for the proposed formalization mainly because of its inherent soundness and ability to handle higher-order logic and in order to benefit from the built-in mathematical theories for conducting probabilistic analysis. Table 1 summarizes some of the HOL symbols used in this paper and their corresponding mathematical interpretation [27].

3.2 Probabilistic Analysis in HOL

Random variables are the core component of conducting probabilistic performance analysis. They can be formalized in higher-order logic as deterministic functions with access to an infinite Boolean sequence $\mathbb{B}^\infty$; a source of infinite random bits [8]. These deterministic functions make random choices based on the result of popping the top most bit in the infinite Boolean sequence and may pop as many random bits as they need for their computation. When the functions terminate, they return the result along with the remaining portion of the infinite Boolean sequence to be used by other programs. Thus, a random variable which

takes a parameter of type α and ranges over values of type β can be represented in HOL by the function.

$$\mathcal{F} : \alpha \to B^\infty \to \beta \times B^\infty$$

As an example, consider the $Bernoulli(\frac{1}{2})$ random variable that returns 1 or 0 with equal probability $\frac{1}{2}$. It can be formalized in HOL as follows

```
⊢ bit = λs.  (if shd s then 1 else 0, stl s)
```

where `s` is the infinite Boolean sequence and `shd` and `stl` are the sequence equivalents of the list operation *'head'* and *'tail'*. The probabilistic programs can also be expressed in the more general state-transforming monad where states are infinite Boolean sequences.

```
⊢ ∀ a s.  unit a s = (a,s)
⊢ ∀ f g s.  bind f g s = let (x,s')← f(s) ∈ g x s'
```

The `unit` operator is used to lift values to the monad, and the `bind` is the monadic analogue of function application. All monad laws hold for this definition, and the notation allows us to write functions without explicitly mentioning the sequence that is passed around, e.g., function `bit` can be defined as

```
⊢ bit_monad = bind sdest (λb.  if b then unit 1 else unit 0)
```

where `sdest` gives the head and tail of a sequence as a pair (*shd* `s`,*stl* `s`). [8] also presents some formalization of the mathematical measure theory in HOL, which can be used to define a probability function $\mathbb{P}$ from sets of infinite Boolean sequences to *real* numbers between 0 and 1. The domain of $\mathbb{P}$ is the set $\mathcal{E}$ of events of the probability. Both $\mathbb{P}$ and $\mathcal{E}$ are defined using the Carathéodory's Extension theorem, which ensures that $\mathcal{E}$ is a σ-algebra: closed under complements and countable unions. The formalized $\mathbb{P}$ and $\mathcal{E}$ can be used to prove probabilistic properties for random variables such as

```
⊢ ℙ {s | fst (bit s) = 1} = 1/2
```

where the function `fst` selects the first component of a pair and $\{x|C(x)\}$ represents a set of all x that satisfy the condition C in HOL.

The measurability and independence of a probabilistic function are important concepts in probability theory. A property `indep`, called *strong function independence*, is introduced

in [8] such that if $f \in$ `indep`, then f will be both measurable and independent. It has been shown in [8] that a function is guaranteed to preserve *strong function independence*, if it accesses the infinite Boolean sequence using only the `unit`, `bind` and `sdest` primitives. All reasonable probabilistic programs preserve *strong function independence*, and these extra properties are a great aid to verification.

The above mentioned approach has been successfully used to formalize both discrete [8, 6] and continuous random variables [9] and verify their correctness in terms of their probability distribution properties, such as PMF or CDF relations. It is often the case that we are more interested in verifying statistical quantities, such as mean or variance, rather than the distribution function of a random variable. For this purpose, [10] presents a higher-order-logic formalization of the following definition of expectation for a function of a random variable

$$Ex[f(R)] = \sum_{n=0}^{\infty} f(n) Pr(R = n) \tag{3}$$

where Ex denotes the expectation function, R is the random variable and f represents a function of the random variable R. Equation 3 has been formalized, for a discrete random variable that attains values in positive integers only and a function that maps this random variable to a *real* value, in [10] as follows

`Definition 1.` *Expectation of Function of a Discrete Random Variable*

`expec_fn:` $(num \rightarrow real) \rightarrow ((num \rightarrow bool) \rightarrow num \times (num \rightarrow bool)) \rightarrow real$

$\vdash \forall$ `f R. expec_fn f R = suminf (`λ`n.  (f n)` $\mathbb{P}$`{s | fst(R s) = n})`

where the mathematical notions of the probability function $\mathbb{P}$ and random variable `R` have been inherited from [8], as presented above, and `suminf` represents the HOL formalization of the infinite summation of a *real* sequence [28]. The function `expec_fn` accepts two parameters, the function f and the *positive integer* valued random variable R and returns a *real* number. The expected value of a discrete random variable that attains values in positive integers can now be defined as a special case of the above definition

`Definition 2.` *Expectation of Discrete Random Variable*

`expec:` $((num \rightarrow bool) \rightarrow num \times (num \rightarrow bool)) \rightarrow real$

$\vdash \forall$ `R. expec R = expec_fn (`λ` n.  n) R`

The function, `expec`, accepts a *positive integer* valued random variable R and returns a *real* number. Using the above two definitions, [10] also presents a formal definition of variance in

HOL for the case of discrete random variables that can attain values in the positive integers only.

Definition 3. *Variance of a Discrete Random Variable*

`variance:` $((num \rightarrow bool) \rightarrow num \times (num \rightarrow bool)) \rightarrow real$

$\vdash \forall$ `R. variance R = expec_fn (`λ`n. (n - expec R)`^{2}`) R`

The function `variance` accepts a discrete random variable R that attains values in the positive integers only and returns a *real* number.

The verification of some useful properties related to expectation and variance of discrete random variables is also presented in [10]. One such property (Equation 4) gives an alternate relationship for variance that is quite useful for the verification of variance properties for discrete random variables in HOL, as will be seen in Section 5 of this paper

$$\forall\ R.\ Var[R] = E[R^2] - (E[R])^2 \tag{4}$$

where Var denotes variance and R is a discrete random variable that can attain values in the positive integers only. This property can be stated in HOL using the formal definitions of variance and expectation as follows.

Theorem 1. *Variance in Terms of Moments*

$\vdash \forall$ `R. (R` $\in$ `indep_fn)` $\wedge$ `(summable(`λ`n. n` $\mathbb{P}${`s | fst (R s) = n`}`))` $\wedge$
`(summable(`λ`n. n`^{2} $\mathbb{P}${`s | fst (R s) = n`}`)))` $\Rightarrow$
`(variance R = expec_fn (`λ`n. n`^{2}`) R - (expec R)`^{2}`)`

The assumptions in Theorem 1 ensure that the random variable R is measurable and its expectation and second moment are well defined, i.e., the summations corresponding to the expectation and second moment of variable R are convergent.

The other two properties that are verified in [10], which will be used in this paper, are linearity of expectation and variance properties [29]. By these properties, the expectation or variance of a sum of independent random variables equals the sum of their individual expectations or variances, respectively.

$$Ex[\sum_{i=1}^{n} R_i] = \sum_{i=1}^{n} Ex[R_i] \tag{5}$$

$$Var[\sum_{i=1}^{n} R_i] = \sum_{i=1}^{n} Var[R_i] \tag{6}$$

The HOL versions of these properties are as follows

Theorem 2. *Linearity of Expectation Property*

$$\vdash \forall \texttt{L}.\ (\forall \texttt{R}.\ (\texttt{mem R L}) \Rightarrow ((\texttt{R} \in \texttt{indep\_fn}) \wedge (\texttt{summable}\ (\lambda \texttt{n}.\ \ \texttt{n}\ \mathbb{P}\{\texttt{s} \mid \texttt{fst(R s)} = \texttt{n}\})))) \Rightarrow (\texttt{expec (sum\_rv\_lst L)} = \sum_{\texttt{n}=0}^{\texttt{length L}} \texttt{(expec (el (length L - (n+1)) L))}$$

Theorem 3. *Linearity of Variance Property*

$$\vdash \forall \texttt{L}.\ (\forall \texttt{R}.\ (\texttt{mem R L}) \Rightarrow ((\texttt{R} \in \texttt{indep\_fn}) \wedge (\texttt{summable}\ (\lambda \texttt{n}.\ \ \texttt{n}\ \mathbb{P}\{\texttt{s} \mid \texttt{fst(R s)} = \texttt{n}\}))) \wedge (\texttt{summable}\ (\lambda \texttt{n}.\ \ \texttt{n}^2\ \mathbb{P}\{\texttt{s} \mid \texttt{fst(R s)} = \texttt{n}\}))) \Rightarrow (\texttt{variance (sum\_rv\_lst L)} = \sum_{\texttt{n}=0}^{\texttt{length L}} \texttt{(variance (el (length L - (n+1)) L))}$$

where the function `length`, defined in the HOL *list* theory, returns the length of its list argument. The function `el`, defined in the *list* theory, accepts a *positive integer* number, say n, and a list and returns the n^{th} element of the given list. The function `mem`, also defined in the *list* theory, accepts a list and an element and returns $True$ if the element is a member of the given list. The function `sum_rv_lst`, given in [10], accepts a list of discrete random variables and returns their sum such that the outcome of each random variable is independent of all the others and is defined as follows

Definition 4. *Summation of n Random Variables*

$$\texttt{sum\_rv\_lst}: \ ((num \rightarrow bool) \rightarrow num \times (num \rightarrow bool))\ list \rightarrow ((num \rightarrow bool) \rightarrow num \times (num \rightarrow bool))$$

$$\vdash \texttt{(sum\_rv\_lst [] = unit 0)} \wedge \forall \texttt{h t.}\ \ \texttt{(sum\_rv\_lst (h::t) =}\ \texttt{bind h (}\lambda\texttt{a.}\ \ \texttt{bind (sum\_rv\_lst t) (}\lambda\texttt{b.}\ \ \texttt{unit (a + b)))}$$

where `::` is the list *cons* operator in HOL that allows us to add a new element to a list. The assumptions in Theorems 2 and 3 ensure that all random variables in the list of random variables, L, are measurable and their expectation is well-defined, in the case of Theorem 2, and their expectation and the second moment is well-defined in the case of Theorem 3.

4 Verification of Markov and Chebyshev's Inequalities

In this section, we present the verification of Markov and Chebyshev's inequalities in HOL using the probabilistic analysis framework, outlined in the previous section.

4.1 Verification of Markov's Inequality in HOL

Markov's inequality, given in Equation 1, utilizes the definition of expectation to obtain a weak tail bound and can be expressed in HOL for a measurable discrete random variable, which attains values in positive integers only, with a well-defined expectation as follows.

Theorem 4. *Markov's Inequality*

$$\vdash \forall\ \texttt{R a.}\ \ (\texttt{0} < \texttt{a}) \wedge (\texttt{R} \in \texttt{indep\_fn}) \wedge (\texttt{summable}(\lambda \texttt{n.}\ \ \texttt{n}\ \mathbb{P}\{\texttt{s}\mid \texttt{fst (R s) = n}\})) \Rightarrow \mathbb{P}\ \{\texttt{s}\mid \texttt{fst (R s)} \geq \texttt{a}\} \leq \frac{\texttt{(expec R)}}{\texttt{a}}$$

where a represents a *real* number.

We proceed with the proof of Theorem 4 in HOL by rewriting its proof goal with the definition of expectation, given in Definition 2,

$$\mathbb{P}\{\texttt{s}|\texttt{fst}(\texttt{R s}) \geq \texttt{a}\} \leq \frac{\lim\limits_{\texttt{k}\to\infty}(\sum_{\texttt{n}=0}^{\texttt{k}}(\texttt{n}\ \mathbb{P}\{\texttt{s}|\texttt{fst}(\texttt{R s}) = \texttt{n}\}))}{\texttt{a}} \tag{7}$$

Now, the set on the left hand side (LHS) of the above inequality can be expressed as follows

$$\{\texttt{s}|\texttt{fst}(\texttt{R s}) \geq \texttt{a}\} = \{\texttt{s}|\texttt{fst}(\texttt{R s}) \geq \lceil \texttt{a} \rceil\} \tag{8}$$

where $\lceil \texttt{x} \rceil$ denotes the ceiling of x, which represents the closest integer for a *real* number x that is greater than or equal to x. The above equation is $True$ because the random variable R acquires values in positive integers only. Thus, all possible values of the random variable R that are greater than a are also greater than or equal to $\lceil a \rceil$ and vice versa. Equation 8 can now be used, along with some arithmetic reasoning in HOL, to rewrite our proof goal (Equation 7) as follows

$$\mathbb{P}\{\texttt{s}|\texttt{fst}(\texttt{R s}) \geq \lceil \texttt{a} \rceil\} \leq \lim_{\texttt{k}\to\infty}(\sum_{\texttt{n}=0}^{\texttt{k}}(\frac{\texttt{n}}{\lceil \texttt{a} \rceil}\ \mathbb{P}\{\texttt{s}|\texttt{fst}(\texttt{R s}) = \texttt{n}\})) \tag{9}$$

Next, we use the complement law of the probability function $P(\overline{A}) = 1 - P(A)$, which is formally verified in [8], to rewrite the LHS of the above inequality as $\texttt{1} - \mathbb{P}\{\texttt{s}|\texttt{fst}(\texttt{R s}) < \lceil \texttt{a} \rceil\}$. The expression $\mathbb{P}\{\texttt{s}|\texttt{fst}(\texttt{R s}) < \lceil \texttt{a} \rceil\}$ can be further simplified using the additive law of probability $P(A \cup B) = P(A) + P(B)$, also verified in [8], as $\sum_{\texttt{n}=0}^{\lceil \texttt{a} \rceil} \mathbb{P}\{\texttt{s}|\texttt{fst}(\texttt{R s}) = \texttt{n}\}$. This simplification allows us to rewrite the subgoal, given in Equation 9, as follows

$$1 - \sum_{\mathtt{n}=0}^{\lceil \mathtt{a} \rceil} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\} \leq \lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=0}^{\mathtt{k}} (\frac{\mathtt{n}}{\lceil \mathtt{a} \rceil}\ \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\})) \tag{10}$$

It can be proved in HOL that $\lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=0}^{\mathtt{k}} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}) = 1$, which allows us to rewrite the LHS of the above inequality as the limit value of the *real* sequence $\sum_{\mathtt{n}=\lceil \mathtt{a} \rceil}^{\mathtt{k}} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}$ as k approaches infinity. Similarly, the expression $\lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=\lceil \mathtt{a} \rceil}^{\mathtt{k}} (\frac{\mathtt{n}}{\lceil \mathtt{a} \rceil}\ \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}))$ can be proved to be less than or equal to the right hand side (RHS) of the above inequality, which allows us to rewrite the subgoal, given in Equation 10, as follows

$$\lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=\lceil \mathtt{a} \rceil}^{\mathtt{k}} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}) \leq \lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=\lceil \mathtt{a} \rceil}^{\mathtt{k}} (\frac{\mathtt{n}}{\lceil \mathtt{a} \rceil}\ \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\})) \tag{11}$$

Now, we verified in HOL that for all values of k, the expression $(\sum_{\mathtt{n}=\lceil \mathtt{a} \rceil}^{\mathtt{k}} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\})$, found on the LHS of the above inequality, is less than or equal to the expression $(\sum_{\mathtt{n}=\lceil \mathtt{a} \rceil}^{\mathtt{k}} (\frac{\mathtt{n}}{\lceil \mathtt{a} \rceil}\ \mathbb{P}\ \{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}))$, found on its RHS. This reasoning allows us to prove the limit relationship, given in Equation 11, between these expressions using the properties of limit of a *real* sequence, formalized in [28], and thus concludes the proof of Markov's inequality, given in Theorem 4.

4.2 Verification of Chebyshev's Inequality in HOL

Chebyshev's inequality (Equation 2) utilizes the variance and the mean characteristics to derive a significantly stronger tail bound than the one obtained by Markov's inequality. We verified the Chebyshev's inequality in HOL by first verifying one of its variants [1]

$$Pr(|X - Ex[X]| \geq a.\sigma[X]) \leq \frac{1}{a^2} \tag{12}$$

where σ denotes the standard deviation function, which returns the square root of variance for the given random variable. This property can be expressed in HOL for a measurable discrete random variable, which attains values in positive integers only, with well-defined first and second moments as follows

Theorem 5. *Chebyshev's Inequality in terms of Standard Deviation*

$$\begin{aligned} \vdash \forall\ \mathtt{R}\ \mathtt{a}.\ \ & (\mathtt{0} < \mathtt{a}) \wedge (\mathtt{0} < \mathtt{variance\ R}) \wedge (\mathtt{R} \in \mathtt{indep\_fn}) \wedge \\ & (\mathtt{summable}(\lambda \mathtt{n}.\ \ \mathtt{n}\ \mathbb{P}\{\mathtt{s} \mid \mathtt{fst}\ (\mathtt{R}\ \mathtt{s}) = \mathtt{n}\})) \wedge \\ & (\mathtt{summable}(\lambda \mathtt{n}.\ \ \mathtt{n}^2\ \mathbb{P}\{\mathtt{s} \mid \mathtt{fst}\ (\mathtt{R}\ \mathtt{s}) = \mathtt{n}\})) \Rightarrow \\ & \quad \mathbb{P}\ \{\mathtt{s} \mid \mathtt{abs}\ (\mathtt{fst}\ (\mathtt{R}\ \mathtt{s}) - \mathtt{expec\ R}) \geq \mathtt{a\ std\_dev\ R}\} \leq \frac{1}{\mathtt{a}^2} \end{aligned}$$

where the HOL function `abs`, defined in [28], returns the absolute value of a *real* number. The HOL function `std_dev`, defined as follows, returns the square root of the variance for a discrete random variable, which attains values in positive integers only

Definition 5. *Standard Deviation of a Discrete Random Variable*

```
std_dev:  ((num → bool) → num × (num → bool)) → real
⊢ ∀ R. std_dev R = sqrt (variance R)
```

where the HOL function `sqrt`, defined in [28], returns the square root of a *real* number. It is important to note that we have used the assumption `0 < variance R` in Theorem 5 because variance is a positive quantity and there is no point in calculating the tail distribution bound for random variables with variance equal to 0.

We proceed with the proof of Theorem 5 in HOL by splitting its proof goal, using the transitivity property of $\leq$, i.e., $(a \leq b \wedge b \leq c \Rightarrow a \leq c)$, into two subgoals as follows

$$\mathbb{P}\{\mathtt{s}|\mathtt{abs}(\mathtt{fst}(\mathtt{R\ s}) - \mu_{\mathtt{R}}) \geq \mathtt{a}\sigma_{\mathtt{R}}\} \leq \mathbb{P}\{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \geq \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}}\} + \mathbb{P}\{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \leq \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\} \tag{13}$$

$$\mathbb{P}\{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \geq \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}}\} + \mathbb{P}\{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \leq \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\} \leq \frac{\mathtt{1}}{\mathtt{a}^{\mathtt{2}}} \tag{14}$$

where the symbols μ_R and σ_R denote the HOL functions for expectation and standard deviation for a random variable R.

The sets $\{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \geq \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}}\}$ and $\{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \leq \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\}$, found on the RHS of Equation 13, can be proved to be disjoint because the term $a\sigma_R$ is greater than 0. This fact along with the additive law of probability $P(A \cup B) = P(A) + P(B)$ allows us to rewrite Equation 13 as follows

$$\mathbb{P}\{\mathtt{s}|\mathtt{abs}(\mathtt{fst}(\mathtt{R\ s}) - \mu_{\mathtt{R}}) \geq \mathtt{a}\sigma_{\mathtt{R}}\} \leq \mathbb{P}\{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \geq \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}}\} \cup \{\mathtt{s}|(\mathtt{fst}(\mathtt{R\ s})) \leq \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\} \tag{15}$$

Now, using arithmetic reasoning, it can be proved in HOL that the set on the LHS of the inequality in Equation 15 is a subset of the set that appears on the RHS. We used this fact along with the increasing probability law $P(A \subseteq B) \Rightarrow P(A) \leq P(B)$ to verify Equation 15 and this concludes the proof of Equation 13.

The next step in the verification of Theorem 5 is to prove the inequality given in Equation 14. We proceed in this direction by replacing the terms on the LHS of the inequality

in Equation 14 as follows

$$\mathbb{P}\{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) \geq \lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil\} + \mathbb{P}\{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) < \lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil\} \cup \{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\} \leq \frac{1}{\mathtt{a}^2} \tag{16}$$

The above step is valid due to the transitivity property of $\leq$, as the sum of the terms on the LHS of the inequality in Equation 16 is greater than the sum of the terms on the LHS of the inequality in Equation 14. This is the case because of the increasing probability law and the fact that the set $\{\mathtt{s}|(\mathtt{fst}(\mathtt{R}\ \mathtt{s})) \geq \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}}\}$ is a subset of the set $\{\mathtt{s}|(\mathtt{fst}(\mathtt{R}\ \mathtt{s})) \geq \lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil\}$ and the set $\{\mathtt{s}|(\mathtt{fst}(\mathtt{R}\ \mathtt{s})) \leq \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\}$ is a subset of the set $\{\mathtt{s}|(\mathtt{fst}(\mathtt{R}\ \mathtt{s})) < \lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil\} \cup \{\mathtt{s}|(\mathtt{fst}(\mathtt{R}\ \mathtt{s})) = \mu_{\mathtt{R}} - \ \mathtt{a}\sigma_{\mathtt{R}}\}$. Next, we can rewrite Equation 16, using arithmetic reasoning, as follows

$$\begin{aligned} \sigma^2_{\mathtt{R}}\mathtt{a}^2(\mathbb{P}\{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) \geq \lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil\} + \\ \mathbb{P}\{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) < \lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil\} \cup \{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\}) \leq \sigma^2_{\mathtt{R}} \end{aligned} \tag{17}$$

where the symbol $\sigma^2_{\mathtt{R}}$ denotes the variance of random variable R. In order to prove the above inequality we try to verify the following relationship regarding its second term on the LHS.

$$\begin{aligned} \sigma^2_{\mathtt{R}}\mathtt{a}^2(\mathbb{P}\{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) < \lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil\} \cup \{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\}) \leq \\ \sum_{\mathtt{n}=0}^{\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil} (\mathtt{n} - \mu_{\mathtt{R}})^2 \mathbb{P}\{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\} \end{aligned} \tag{18}$$

The two sets, in the union, on the LHS of the above inequality are disjoint, which allows us to rewrite the expression on the LHS as a sum of two probabilities, using the additive law of probability. The first probability term, out of these two terms, can then be expressed as a sum $\sum_{\mathtt{n}=0}^{\lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil} \sigma^2_{\mathtt{R}}\mathtt{a}^2\mathbb{P}\{\mathtt{s}|\mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}$ using the additive law of probability. Whereas, the expression on the RHS of the above inequality can be split into the sum of two terms, using the definition of the summation function in HOL, as follows

$$\sum_{\mathtt{n}=0}^{\lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil} \sigma_{\mathtt{R}}^2 \mathtt{a}^2 \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\} + \sigma_{\mathtt{R}}^2 \mathtt{a}^2 \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}}\}) \leq$$
$$\sum_{\mathtt{n}=0}^{\lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil} (\mathtt{n} - \mu_{\mathtt{R}})^2 \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\} + \sum_{\mathtt{n}=\lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil}^{\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil - \lceil \mu_{\mathtt{R}} - \mathtt{a}\sigma_{\mathtt{R}} \rceil} (\mathtt{n} - \mu_{\mathtt{R}})^2 \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\} \tag{19}$$

Now the above inequality can be proved in HOL, as both the terms on the LHS of the above equation are less than or equal to the corresponding two terms on the RHS. This result allows us to rewrite the inequality, given in Equation 17, as follows

$$\sigma_{\mathtt{R}}^2 \mathtt{a}^2 \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) \geq \lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil\} + \sum_{\mathtt{n}=0}^{\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil} (\mathtt{n} - \mu_{\mathtt{R}})^2 \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\} \leq \sigma_{\mathtt{R}}^2 \tag{20}$$

using the transitivity property of $\leq$. Now, using the definition of variance and rearranging the terms, based on arithmetic reasoning, the above equation can be rewritten as follows

$$\mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) \geq \lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil\} \leq$$
$$\lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=0}^{\mathtt{k}} \frac{(\mathtt{n} - \mu_{\mathtt{R}})^2}{\sigma_{\mathtt{R}}^2 \mathtt{a}^2}\ \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}) - \sum_{\mathtt{n}=0}^{\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil} \frac{(\mathtt{n} - \mu_{\mathtt{R}})^2}{\sigma_{\mathtt{R}}^2 \mathtt{a}^2} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\} \tag{21}$$

The probability term on the LHS of the above inequality can be expressed in terms of the limit of the *real* sequence $\sum_{\mathtt{n}=\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil}^{\mathtt{k}} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}$ as k approaches infinity, using the same reasoning as was used for the case of the proof of Markov's inequality in Equations 9 to 11. Similarly, the expression on the RHS of the above inequality can also be expressed in terms of a limit of a *real* sequence, which allows us to rewrite Equation 21 as follows

$$\lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil}^{\mathtt{k}} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}) \leq \lim_{\mathtt{k} \to \infty} (\sum_{\mathtt{n}=\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil}^{\mathtt{k}} \frac{(\mathtt{n} - \mu_{\mathtt{R}})^2}{\sigma_{\mathtt{R}}^2 \mathtt{a}^2}\ \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}) \tag{22}$$

It can be verified in HOL that for all values of k, the expression $\sum_{\mathtt{n}=\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil}^{\mathtt{k}} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}$ found on the LHS of the above inequality, is less than or equal to the expression $\sum_{\mathtt{n}=\lceil \mu_{\mathtt{R}} + \mathtt{a}\sigma_{\mathtt{R}} \rceil}^{\mathtt{k}} \frac{(\mathtt{n} - \mu_{\mathtt{R}})^2}{\sigma_{\mathtt{R}}^2 \mathtt{a}^2} \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{R}\ \mathtt{s}) = \mathtt{n}\}$, found on its RHS. This reasoning allows us to prove the limit relationship, given in Equation 22, between these expressions using the properties of limit

of a *real* sequence, formalized in [28], which completes the proof of the inequality given in Equation 14 and thus concludes the proof of Theorem 5 as well.

Theorem 5 can now be used to verify the Chebyshev's inequality, given in Equation 2, in HOL as a special case when the constant a is assigned the value $\frac{\texttt{a}}{\texttt{std\_dev R}}$. The corresponding HOL theorem can be expressed for a measurable discrete random variable, which attains values in positive integers only, with well-defined first and second moments as follows

Theorem 6. *Chebyshev's Inequality*

$$\vdash \forall\ \texttt{R a.}\ \ \texttt{(0 < a)} \wedge \texttt{(0 < variance R)} \wedge \texttt{(R} \in \texttt{indep\_fn)} \wedge \texttt{(summable(}\lambda\texttt{n.}\ \ \texttt{n}\ \mathbb{P}\{\texttt{s | fst (R s) = n}\})) \wedge \texttt{(summable(}\lambda\texttt{n.}\ \ \texttt{n}^2\ \mathbb{P}\{\texttt{s | fst (R s) = n}\})) \Rightarrow \mathbb{P}\ \{\texttt{s | abs (fst (R s) - expec R)} \geq \texttt{a}\} \leq \frac{\texttt{variance R}}{\texttt{a}^2}$$

Theorems 5 and 6 represent the HOL theorems corresponding to Markov's and Chebyshev's inequalities and the results are found to be in good agreement with the existing theoretical paper-and-pencil counterparts given in Equations 1 and 2, respectively. These formally verified theorems allow us to reason about tail distribution bounds within the HOL theorem prover as will be demonstrated in Section 6 of this paper.

5 Verification of Mean and Variance for Discrete Distributions

In this section, we utilize the formal definitions of expectation and variance, given in Definitions 2 and 3, respectively, to verify the mean and variance properties of Uniform(m), Bernoulli(p), Geometric(p) and Binomial(m, p) random variables in HOL. The formally verified mean and variance relations of these discrete random variables can in turn be used, along with the formally verified Markov and Chebyshev's inequalities presented in the last section, to formally reason about the tail distribution properties of their respective random variables.

5.1 Uniform(m) Random Variable

The Uniform(m) random variable assigns equal probability to each element in the set $\{0, 1, \cdots, (m-1)\}$ and thus ranges over a finite number of positive integers. A sampling algorithm for the Uniform(m) can be found in [8], which has been proven correct by verifying the corresponding PMF property in HOL

$$\vdash \forall\ \texttt{m x.}\ \ \texttt{x} < \texttt{m} \Rightarrow \mathbb{P}\ \{\texttt{s}\ |\ \texttt{fst (prob\_unif m s) = x}\} = \frac{1}{\texttt{m}}$$

where `prob_unif` represents the higher-order-logic function for the Uniform(m) random variable.

Now, we want to formally verify the mean characteristic for the Uniform(m), which can be expressed in HOL as follows.

Theorem 7. *Expectation of Uniform(m) Random Variable*

$$\vdash \forall\ \texttt{m.}\ \ \texttt{expec}\ (\lambda\texttt{s.}\ \ \texttt{prob\_unif (m+1) s}) = \frac{\texttt{m}}{2}$$

We proceed with the proof of this theorem in HOL by rewriting it with the definition of expectation

$$\lim_{\texttt{k}\to\infty}(\sum_{\texttt{n=0}}^{\texttt{k}} \texttt{n}\ \mathbb{P}\{\texttt{s}\mid \texttt{fst}(\texttt{prob\_unif}\ (\texttt{m}+1)\ \texttt{s}) = \texttt{n}\}) = \frac{\texttt{m}}{2} \tag{23}$$

Next, we verified in HOL that the Uniform(m) random variable can never acquire a value greater than or equal to m using its PMF property.

$$\vdash \forall\ \texttt{m x.}\ \ \texttt{(m + 1)} \leq \texttt{x} \Rightarrow \mathbb{P}\{\texttt{s}\ |\ \texttt{fst(prob\_unif (m + 1) s) = x}\} = 0$$

This property allows us to rewrite the infinite summation of Equation 23 in terms of a finite summation over $(m+1)$ values using the properties verified in the HOL theory of *limit of a real sequence*.

$$\sum_{\texttt{n=0}}^{\texttt{m+1}} \texttt{n}\ \mathbb{P}\{\texttt{s}\mid \texttt{fst}(\texttt{prob\_unif}\ (\texttt{m}+1)\ \texttt{s}) = \texttt{n}\} = \frac{\texttt{m}}{2} \tag{24}$$

The above equation can be verified using the PMF of the Uniform(m) random variable along with some basic properties of the summation function in HOL.

Next, we formally verify the variance characteristic for the Uniform(m) random variable, which can be expressed in HOL as follows.

Theorem 8. *Variance of Uniform(m) Random Variable*

$$\vdash \forall\ \texttt{m.}\ \ \texttt{variance}\ (\lambda\texttt{s.}\ \ \texttt{prob\_unif (m+1) s}) = \frac{(\texttt{m}+1)^2-1}{12}$$

The proof goal of Theorem 8 can be simplified using the variance relation given in Theorem 1, and the definition of expectation of a function of a random variable (Definition 1) as follows

$$\sum_{\mathtt{n}=0}^{\infty} \mathtt{n}^2 \mathbb{P}\{\mathtt{s} | \mathtt{fst}(\mathtt{prob\_unif}\ (\mathtt{m}+1)\ \mathtt{s}) = \mathtt{n}\} - (\mathtt{expec}(\lambda \mathtt{s}.\ \mathtt{prob\_unif}\ (\mathtt{m}+1)\ \mathtt{s}))^2 = \frac{(\mathtt{m}+1)^2 - 1}{12} \tag{25}$$

Now, the second moment of the Uniform(m) random variable, i.e., the first term on the LHS of the above equation, can be verified in HOL to be equal to $\frac{\mathtt{m}(2\mathtt{m}+1)}{2}$, using the same approach as was used for the verification of its expectation relation in Theorem 7. This result and some arithmetic reasoning, allows us to verify Equation 25 and thus Theorem 8 in HOL.

5.2 Bernoulli(p) Random Variable

The Bernoulli(p) random variable models an experiment with two outcomes; success and failure, whereas the parameter p represents the probability of success. A sampling algorithm of the Bernoulli(p) random variable has been formalized in [8] as the function `prob_bern` such that it returns $True$ with probability p and $False$ otherwise. It has also been verified to be correct by proving the corresponding PMF property in HOL.

$$\vdash\ \forall\ \mathtt{p}.\ \ 0 \leq \mathtt{p} \wedge \mathtt{p} \leq 1 \Rightarrow \mathbb{P}\ \{\mathtt{s}\ |\ \mathtt{fst}\ (\mathtt{prob\_bern}\ \mathtt{p}\ \mathtt{s})\}\ =\ \mathtt{p}$$

The Bernoulli(p) random variable ranges over 2 values of *Boolean* data type. The expectation property of these kind of discrete random variables, which range over a finite number of values of a different data type than positive integers, can be verified in HOL by mapping all their values to distinct positive integers. In the case of Bernoulli(p) random variable, we redefined the function `prob_bern` such that it returns positive integers 1 and 0 instead of the Boolean quantities $True$ and $False$, respectively, i.e., the range of the random variable was changed from *Boolean* data type to positive integers. It is important to note that this redefinition does not change the distribution properties of the given random variable. The expectation property for this alternate definition of Bernoulli(p) random variable, `prob_bernN`, can be expressed in HOL as follows

`Theorem 9.` *Expectation of Bernoulli(p) Random Variable*

$$\vdash\ \forall\ \mathtt{p}.\ \ 0 \leq \mathtt{p} \wedge \mathtt{p} \leq 1 \Rightarrow \mathtt{expec}\ (\lambda \mathtt{s}.\ \ \mathtt{prob\_bernN}\ \mathtt{p}\ \mathtt{s})\ =\ \mathtt{p}$$

Theorem 9 can now be verified using the same procedure used for the case of random variables that range over a finite number of positive integers, such as the Unform(m) random variable. In the case of Bernoulli(p) random variable, we were able replace the infinite

summation in the definition of expectation with the summation of the first two values of the corresponding *real* sequence using the HOL theory of *limit of a real sequence*. This substitution along with the PMF property of the Bernoulli(p) random variable and some arithmetic reasoning allowed us to verify Theorem 9 in HOL.

We also verified the variance of the Bernoulli(p) random variable in HOL, using a similar approach that we used for the verification of the variance relation for the Unform(m) random variable and the HOL theorem is given below

`Theorem 10.` *Variance of Bernoulli(p) Random Variable*

$$\vdash \forall \texttt{p.}\ \ 0 \leq \texttt{p} \wedge \texttt{p} \leq 1 \Rightarrow \texttt{variance}\ (\lambda \texttt{s.}\ \ \texttt{prob\_bernN p s}) = \texttt{p}\ (1\texttt{-p})$$

5.3 Geometric(p) Random Variable

The Geometric(p) random variable can be defined as the index of the first success in an infinite sequence of Bernoulli(p) trials [30]. Therefore, the Geometric(p) distribution may be sampled by extracting random bits from the function `prob_bern`, explained in the previous section, and stopping as soon as the first $False$ is encountered and returning the number of trials performed till this point. Thus, the Geometric(p) random variable ranges over a countably infinite number of positive integers numbers. This fact makes it different from other random variables that we have considered so far. Based on the above sampling algorithm, the Geometric(p) random variable has been formalized in [6] as the function `prob_geom`, which has also been verified to be correct by proving the corresponding PMF property in HOL.

$$\vdash \forall \texttt{n p.}\ \ 0 < \texttt{p} \wedge \texttt{p} \leq 1 \Rightarrow \mathbb{P}\ \{\texttt{s} \mid \texttt{fst (prob\_geom p s) = (n + 1)}\} = \texttt{p}\ (1 - \texttt{p})^{\texttt{n}}$$

It is important to note that p, which represents the probability of success for the Geometric(p) or the probability of obtaining $False$ from the Bernoulli(p) random variable, cannot be assigned a value equal to 0 as this will lead to a non-satisfying success condition for the Geometric random variable.

The expectation theorem for the Geometric(p) random variable can now be expressed in HOL as follows

`Theorem 11.` *Expectation of Geometric(p) Random Variable*

$$\vdash \forall \texttt{p.}\ \ 0 < \texttt{p} \wedge \texttt{p} \leq 1 \Rightarrow \texttt{expec}\ (\lambda \texttt{s.}\ \ \texttt{prob\_geom p s}) = \frac{1}{\texttt{p}}$$

Rewriting the above proof goal with the definition of expectation and simplifying using the PMF relation for the Geometric(p) random variable along with some arithmetic reasoning, we reach the following subgoal.

$$\lim_{k\to\infty}(\sum_{n=0}^{k}((n+1)p(1-p)^n)) = \frac{1}{p} \tag{26}$$

Substituting $1-q$ for p and after some rearrangement of the terms, based on arithmetic reasoning, the above subgoal can be rewritten as follows.

$$\lim_{k\to\infty}(\sum_{n=0}^{k}((n+1)q^n)) = \frac{1}{(1-q)^2} \tag{27}$$

Now, using the properties of summation of a *real* sequence in HOL, we proved the following relationship

$$\forall q\ k.\sum_{n=0}^{k}((n+1)q^n) = \sum_{n=0}^{k}(\sum_{i=0}^{k}q^i - \sum_{i=0}^{n}q^i) \tag{28}$$

which allows us to rewrite the subgoal under consideration, given in Equation 27 as follows.

$$\lim_{k\to\infty}(\sum_{n=0}^{k}(\sum_{i=0}^{k}q^i - \sum_{i=0}^{n}q^i)) = \frac{1}{(1-q)^2} \tag{29}$$

The above subgoal can now be proved using the summation of a finite geometric series along with some properties of summation and limit of *real* sequences available in the *real* number theories in HOL. This also concludes the proof of Theorem 11 in HOL.

The variance property of Geometric(p) random variable can be stated in HOL as follows.

Theorem 12. *Variance of Geometric(p) Random Variable*

$\vdash \forall$ p. 0 < p $\wedge$ p $\leq$ 1 $\Rightarrow$ (variance (λs. prob_geom p s) = $\frac{1-p}{p^2}$)

We utilize the variance property, proved in Theorem 1, to verify Theorem 12. The foremost step in this regard is to verify the second moment relationship for the Geometric(p) random variable.

$\vdash \forall$ p. 0 < p $\wedge$ p $\leq$ 1 $\Rightarrow$
(expec_fn (λn. (n^2) (λs. prob_geom p s)) = $\frac{2}{p^2} - \frac{1}{p}$)

Rewriting the above proof goal with the definition of function `expec_fn` and simplifying using the PMF relation of the Geometric random variable along with some properties from HOL *real* number theories, we reach the following subgoal.

$$\lim_{k\to\infty}(\sum_{n=0}^{k}((n+1)^2 p(1-p)^n)) = \frac{2}{p^2} - \frac{1}{p} \tag{30}$$

Now, substituting $1-q$ for p and after some rearrangement of the terms, based on arithmetic reasoning, the above subgoal can be rewritten as follows.

$$\lim_{k\to\infty}(\sum_{n=0}^{k}((n+1)^2 q^n)) = \frac{2}{(1-q)^3} - \frac{1}{(1-q)^2} \tag{31}$$

Using the properties of summation of a *real* sequence in HOL, we proved the following

$$\forall q\ k.\ \sum_{n=0}^{k}((n+1)^2 q^n) = \sum_{n=0}^{k}((2n+1)(\sum_{i=0}^{k} q^i - \sum_{i=0}^{n} q^i)) \tag{32}$$

which allows us to rewrite the subgoal under consideration, given in Equation 31 as follows.

$$\lim_{k\to\infty}(\sum_{n=0}^{k}((2n+1)(\sum_{i=0}^{k} q^i - \sum_{i=0}^{n} q^i))) = \frac{2}{(1-q)^3} - \frac{1}{(1-q)^2} \tag{33}$$

The above subgoal can now be proved using the summation of a finite geometric series along with some properties of summation and limit of *real* sequences available in the *real* number theories in HOL. This concludes the proof of the second moment relation for the Geometric(p) random variable, which can now be used along with Theorems 1 and 11 and some arithmetic reasoning to prove Theorem 12 in HOL.

5.4 Binomial(m, p) Random Variable

The Binomial(m, p) random variable models an experiment which counts the number of successes in a finite number, m, of independent Bernoulli trials, with a success probability equal to p [30]. Therefore, the Binomial(m, p) distribution may be sampled by an algorithm in HOL that sums m independent outcomes of the `prob_bernN` random variable, which models the Bernoulli(p) random variable with outcomes 0 and 1, as described in Section 5.2. We formalized it in HOL by first defining a function that recursively returns a list of m Bernoulli(p) random variables.

`Definition 6.` *List of m Bernoulli(p) Random Variables*

```
bern_lst:  num → real → ((num → bool) → num × (num → bool)) list
⊢ ∀ p.  bern_lst 0 p = [] ∧
 (⊢ ∀ n p.  bern_lst (n + 1) p = prob_bernN p ::  (bern_lst n p))
```

Now, the Binomial(m, p) random variable can be modeled as the sum of all elements in the list modeled by the HOL function `bern_lst`, such that the result of each one of these random variables is independent of one another. This can be done using the function `sum_rv_lst`, given in Definition 4, as follows

Definition 7. *Binomial(m,p) Random Variable*

```
prob_bino:  num → real → ((num → bool) → num × (num → bool))
 ⊢ ∀ m p.  prob_bino m p = sum_rv_lst (bern_lst m p)
```

We verified the correctness of the above definition by verifying its PMF characteristic in HOL using the properties verified in the HOL libraries corresponding to the *probability* and *set* theories.

Theorem 13. *PMF of Binomial(m,p) Random Variable*

```
⊢ ∀ m p n.  0 ≤ p ∧ p ≤ 1 ⇒ ℙ {s | fst (prob_bino m p s) = n} =
    (binomial m n) (p^n) ((1 - p)^(m−n))
```

where the HOL function (`binomial m n`) represents the term $\frac{m!}{n!(m-n)!}$.

The expectation theorem for the Binomial(m, p) random variable can now be expressed in HOL

Theorem 14. *Expectation of Binomial(m,p) Random Variable*

```
⊢ ∀ m p.  0 ≤ p ∧ p ≤ 1 ⇒ expec (λs.  prob_bino m p s) = m p
```

Instead of using the definition of expectation directly, we use the linearity of expectation property, given in Theorem 2, to prove the above theorem. This way, we do not need to deal with the summation involving the `binomial` function in HOL, which saves a considerable amount of proof effort. Since, the Binomial(m, p) random variable represents the sum of m Bernoulli(p) random variables, the linearity of expectation property allows us to rewrite the LHS of the proof goal in Theorem 14 as the sum of m expectation values of the Bernoulli(p) random variable. Now, using the fact that the expectation of the Bernoulli(p) random variable is equal to p, as given in Theorem 9, Theorem 14 can be verified in HOL.

In a similar way, we can also verify the variance relation for the Binomial(m, p) in HOL using the linearity of variance property, given in Theorem 3.

Theorem 15. *Variance of Binomial(m,p) Random Variable*

$\vdash \forall$ m p. 0 $\leq$ p $\wedge$ p $\leq$ 1 $\Rightarrow$
variance (λs. prob_bino m p s) = m p (1 - p)

The formalization and verification of the Binomial(m, p), presented in this section, illustrates one of the main strengths of mechanical theorem proving, i.e., the reusability of existing definitions and theorems to develop and prove new and more complex definitions and theorems. This approach greatly speeds up the formal verification process and allows us to take the work further than would have been possible starting from scratch, without compromising on the soundness of the results.

6 Performance Analysis of Coupon Collector's Problem in HOL

In this section, we utilize the HOL formalization presented so far to formally analyze the tail distribution properties of the Coupon Collector's problem [2]. Firstly, we present a brief overview of the algorithm and present its formalization in HOL.

The Coupon Collector's problem refers to the problem of probabilistically evaluating the number of trials required to acquire all unique, say n, coupons from a collection of multiple copies of these coupons that are independently and uniformly distributed. The problem is similar to the example when each box of cereal contains one of n different coupons and once you obtain one of every type of coupon, you win a prize. This simple problem arises in many different scenarios. For example, suppose that packets are sent in a stream from source to destination host along a fixed path of routers. It is often the case that the destination host would like to know all routers that the stream of data has passed through. This may be done by appending the identification of each router to the packet header but this is not a practical solution as usually we do not have this much room available. An alternate way of meeting this requirement is to store the identification of only one router, uniformly selected at random between all routers on the path, in each packet header. Then, from the point of view of the destination host, determining all routers on the path is like a Coupon Collector's problem. An approach for the formalization of the Coupon Collector's problem as a probabilistic algorithm in higher-order-logic and the verification of its expectation relationship has been presented in [6]. We mainly build upon this model to verify its variance and tail distribution bounds in this Section.

The Coupon Collector's problem can be formalized by modeling the total number of trials required to obtain all n unique coupons, say X, as a sum of the number of trials required

to obtain each distinct coupon, i.e., $X = \sum_{i=1}^{n} X_i$, where X_i represents the number of trials to obtain the i^{th} coupon, while $i-1$ distinct coupons have already been acquired. The advantage of breaking the random variable X into the sum of n random variables $X_1, X_2 \cdots, X_n$ is that each X_i can be modeled as a Geometric(p) random variable. Based on the above model, the expectation relation for the Coupon Collector's problem can be verified using the linearity of expectation property, given in Theorems 2, and the expectation of the Geometric(p) random variable, given in Theorem 3.

The Coupon Collector's problem is modeled in HOL by identifying the coupons with unique positive integers, such that the first coupon acquired by the coupon collector is identified as number 0 and after that each different kind of a coupon acquired with subsequent numbers in numerological order. The coupon collector saves these coupons in a list of positive integers. The following function accepts the number of distinct coupons acquired by the coupon collector and recursively generates the corresponding coupon collector's list.

Definition 8. *Coupon Collector's List*

```
coupon_lst:  (num → num list)
   ⊢ (coupon_lst 0 = []) ∧
   ∀ n.  (coupon_lst (n + 1) = n ::  (coupon_lst n))
```

The next step is to define a list of Geometric random variables, such that each one of its elements represents an X_i, mentioned above. It is important to note that the probability of success for each one of these Geometric random variables is different from one another and depends on the number of different coupons acquired so far. Since, every coupon is drawn independently and uniformly at random from the n possibilities and the coupons are identified with positive integers, we can use the Uniform(n) random variable to model each trial of acquiring a coupon. Now we can define the probability of success for a particular Geometric random variable as the probability of the event when the Uniform(n) random variable generates a new value, i.e., a value that is not already present in the coupon collector's list. Using this probability of success, the following function generates the required list of Geometric random variables

Definition 9. *Geometric Variable List for Coupon Collector's Problem*

```
geom_rv_lst:  (num list   → num → ((num → bool) → num × (num → bool)) list)
  ⊢ ∀ n.  (geom_rv_lst [] n = [prob_geom 1]) ∧
    ∀ h t n.  (geom_rv_lst (h::t) n =
      (prob_geom ℙ{s | ∼(mem (fst(prob_unif n s)) (h::t))}) ::
                    (geom_rv_lst t n))
```

The `geom_rv_lst`, accepts two arguments; a list of positive integers that represents the coupon collector's list and a *positive integer* number that represents the total number of coupons in the Coupon Collector's problem. It returns, a list of Geometric random variables, whose sum would model the total number of trials required to acquire all coupons. The base case in the above recursive definition corresponds to the condition when the coupon collector does not have any coupon and thus the probability of success, i.e., the probability of acquiring a new coupon is 1.

Using the above definitions along with the function `sum_rv_lst`, given in Definition 4, the Coupon Collector's problem has been formally represented in HOL as follows.

`Definition 10.` *Probabilistic Algorithm for Coupon Collector's Problem*

`coupon_collector:` $(num \rightarrow ((num \rightarrow bool) \rightarrow num \times (num \rightarrow bool)))$

```
⊢ ∀ n.  (coupon_collector (n + 1) =
          (sum_rv_lst (geo_rv_lst (coupon_lst n) (n + 1)))
```

The function, `coupon_collector`, accepts a positive integer greater than 0, i.e., $n+1$, which represents the total number of different coupons that are required to be collected. It returns the number of trials for acquiring these $n+1$ distinct coupons.

The first step towards the verification of statistical properties for the above algorithm of the Coupon Collector's problem is to verify the relation for the probability of acquiring a new coupon.

`Theorem 16.` *Probability of Acquiring a New Coupon*

$$\vdash \forall\ \texttt{L n.}\ \ (\texttt{dist\_lst L}) \wedge (\forall \texttt{a.}\ \ \texttt{mem a L} \Rightarrow (\texttt{a < (n + 1)})) \Rightarrow (\mathbb{P}\ \{\texttt{s} \mid \sim(\texttt{mem (fst(prob\_unif (n + 1) s)) L})\} = 1 - \frac{(\texttt{length L})}{(\texttt{n+1})})$$

where the predicate `dist_lst` returns $True$ if all elements in its argument list are distinct. Thus, the assumption in the above theorem ensures that all elements in the given list of positive integers are distinct and are less than $(n+1)$. The coupon collector's list, modeled by the function `coupon_lst`, satisfies both assumptions in Theorem 16 for any given argument. Therefore, the probability of succuss for the Geometric random variable, which models the acquiring process of a new coupon when the coupon collectors list is exactly equal to L, is $1-\frac{length\ L}{(n+1)}$. The expectation of such a Geometric random variable can be easily verified to be equal to $\frac{n+1}{(n+1)-(length\ L)}$, by Theorem 11. This result along with the linearity of expectation property, given in Theorem 2, has been used in [6] to verify the expectation or mean of the number of trials to collect all distinct coupons.

Theorem 17. *Expectation of Coupon Collector's Problem*

$\vdash \forall$ n. expec (coupon_collector (n + 1)) = (n + 1) $(\sum_{i=0}^{n+1} \frac{1}{i+1})$

In this paper, we build upon the above infrastructure to formally reason about the tail distribution properties of the number of trials required to acquire all coupons in HOL. For this purpose, we utilize the formally verified Markov's and Chebyshev's inequalities, which have been verified in Theorems 4 and 5, respectively. The first step in this regard is to have access to formal proofs for the mean and variance relations for the events of interest. The mean has already been verified, given in Theorem 17, and thus we proceed by verifying a relationship for the variance first.

Instead of verifying the exact value of the variance for the number of trials required to acquire all coupons, we verify an upper bound for this variance

Theorem 18. *Variance Upper Bound of Coupon Collector's Problem*

$\vdash \forall$ n. variance (coupon_collector (n + 1))
$\leq$ ((n + 1)^2) $(\sum_{i=0}^{n+1} (\frac{1}{(i+1)^2}))$

The formal proof for the above theorem is based on the definition of the function coupon_collector, the linearity of variance property, given in Theorem 3, the result of Theorem 16, and the variance of Geometric random variable, verified in Theorem 12, along with some arithmetic reasoning.

Now, using the above mentioned results, we can formally verify the following two tail distribution bounds for the Coupon Collector's problem based on the formally verified Markov's and Chebyshev's inequalities, respectively.

Theorem 19. *Weak Tail Distribution Bound for the Coupon Collector's Problem*

$\vdash \forall$ n a. 0 < a $\Rightarrow \mathbb{P}$ {s | (fst (coupon_collector (n + 1) s)) $\geq$ a}
$\leq$ $(\frac{(n+1)}{a} (\sum_{i=0}^{n+1} \frac{1}{(i+1)}))$

Theorem 20. *Stronger Tail Distribution Bound for the Coupon Collector's Problem*

$\vdash \forall$ n a. 0 < a $\Rightarrow \mathbb{P}$ {s | abs ((fst (coupon_collector (n + 1) s)) -
expec (coupon_collector (n + 1))) $\geq$ a}
$\leq$ $(\frac{(n+1)^2}{a^2} (\sum_{i=0}^{n+1} \frac{1}{(i+1)^2}))$

With these results, we have been able to formally verify the tail distribution bounds for the number of trials required to acquire all distinct coupons in the Coupon Collector's problem. These bounds reveal the tail distribution characteristics for the Coupon Collector's problem, which is something that cannot be inferred by just the mean or variance quantities.

We were able to obtain this information using of the formally verified Markov or Chebyshev's inequalities, which is the main contribution of this paper. It is also important to note here that our results exactly match the results of the analysis based on paper-and-pencil proof techniques [2] and are thus 100 % precise, which is a novelty that cannot be achieved, to the best of our knowledge, by any existing computer based probabilistic analysis tool.

7 Conclusions

In this paper, we presented an approach that allows us to precisely reason about the tail distribution properties of random systems within the higher-order-logic theorem prover HOL. Bounding the tail distribution plays a vital role in determining the failure probabilities in the domain of probabilistic analysis, e.g., [2] utilizes the tail distribution bounds, estimated using the Chebyshev's inequality, to find the probability of failure for a randomized algorithm for computing the median of a given set of numbers. The formalization presented in this paper allows us to handle such problems in HOL as has been shown for the case of the Coupon Collector's problem. Due to the inherent soundness of the theorem-proving based analysis, our approach ensures accurate and precise results and thus can prove to be quite useful for the performance and reliability optimization of safety critical and highly sensitive application domains, such as medicine, military or transportation.

The main contributions of this paper are the verification of Markov's and Chebyshev's inequalities and the mean and variance relations for some commonly used discrete random variables. These formally verified results can be reused for the verification of tail distribution properties in a number of different probabilistic analysis domains. In order to illustrate the practical effectiveness of our work, we presented the analysis for the Coupon Collector's problem in this paper. To the best of our knowledge, this is the first time that it has been possible to reason about the tail distribution properties in a mechanized formal methods environment.

The infrastructure presented in this paper can be extended further by verifying the expectation and variance properties of a number of other random variables, which attain values in positive integers only, e.g., Binomial, Logarithmic and Poisson [29]. The verification of Chernoff bounds [2], which are extremely powerful and give exponentially decreasing bounds on the tail distribution, would also be of great benefit. The formally verified Markov's inequality and the formal definition of expectation of a function of a random variable, presented in this paper, can be utilized for this purpose. Another very promising future direction could be to link the formal definition of expectation, presented in this paper, with the higher-order-logic formalization of Lebesgue integration theory [31], which would further

strengthen the soundness of the definitions presented in this paper. This would also pave the way for the verification of statistical quantities, such as mean and variance, for continuous random variables, such as Normal, Exponential, etc.

Finally, it is important to note that higher-order-logic theorem proving cannot be regarded as the golden solution in performing probabilistic analysis because of its own limitations. Even though theorem provers have been successfully used for a variety of tasks, including some that have eluded human mathematicians for a long time, but these successes are sporadic, and work on hard problems usually requires a proficient user and a lot of formalization. On the other hand, simulation based techniques are at least capable of offering approximate solutions to these problems. Therefore, we consider simulation and higher-order-logic theorem proving as complementary techniques, i.e., the methods have to play together for a successful probabilistic analysis framework. For example, theorem proving can be used for the safety critical parts of the design, which can be expressed in closed mathematical forms, and simulation based approaches can handle the rest.

References

[1] P. Billingsley. *Probability and Measure*. John Wiley, 1995.

[2] M. Mitzenmacher and E. Upfal. *Probability and Computing*. Cambridge University Press, 2005.

[3] D.J.C. MacKay. Introduction to Monte Carlo Methods. In *Learning in Graphical Models, NATO Science Series*, pages 175–204. Kluwer Academic Press, 1998.

[4] B.D. McCullough. Assessing the Reliability of Statistical Software: Part I. *The American Statistician*, 52(4):358–366, 1998.

[5] B.D. McCullough. Assessing the Reliability of Statistical Software: Part II. *The American Statistician*, 53(2):149–159, 1999.

[6] O. Hasan and S. Tahar. Verification of Expectation Properties for Discrete Random Variables in HOL. In *Theorem Proving in Higher-Order Logics*, volume 4732 of *LNCS*, pages 119–134. Springer, 2007.

[7] M.J.C. Gordon. Mechanizing Programming Logics in Higher-0rder Logic. In *Current Trends in Hardware Verification and Automated Theorem Proving*, pages 387–439. Springer, 1989.

[8] J. Hurd. *Formal Verification of Probabilistic Algorithms.* PhD Thesis, University of Cambridge, Cambridge, UK, 2002.

[9] O. Hasan and S. Tahar. Formalization of the Continuous Probability Distributions. In *Automated Deduction*, volume 4603 of *LNAI*, pages 3–18. Springer, 2007.

[10] O. Hasan and S. Tahar. Formal Verification of Expectation and Variance for Discrete Random Variables. Technical Report, Concordia University, Montreal, Canada, June 2007; http://hvg.ece.concordia.ca/Publications/TECH_REP/FVEVDR_TR07.

[11] O. Hasan and S. Tahar. Verification of Tail Distribution Bounds in a Theorem Prover. In *Numerical Analysis and Applied Mathematics*, volume 936, pages 259–262. American Institute of Physics, 2007.

[12] A. Nedzusiak. σ-fields and Probability. *Journal of Formalized Mathematics*, 1, 1989.

[13] J. Bialas. The σ-Additive Measure Theory. *Journal of Formalized Mathematics*, 2, 1990.

[14] J. Hurd, A. McIver, and C. Morgan. Probabilistic Guarded Commands Mechanized in HOL. *Theoretical Computer Science*, 346:96–112, 2005.

[15] O. Celiku. Quantitative Temporal Logic Mechanized in HOL. In *International Colloquium Theoretical Aspects of Computing*, volume 3722 of *LNCS*, pages 439–453. Springer, 2005.

[16] P. Audebaud and C. Paulin-Mohring. Proofs of Randomized Algorithms in Coq. In *Mathematics of Program Construction*, volume 4014 of *LNCS*, pages 49–68. Springer, 2006.

[17] O. Hasan and S. Tahar. Verification of Probabilistic Properties in HOL using the Cumulative Distribution Function. In *Integrated Formal Methods*, volume 4591 of *LNCS*, pages 333–352. Springer, 2007.

[18] L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, 1986.

[19] C. Baier, B. Haverkort, H. Hermanns, and J.P. Katoen. Model Checking Algorithms for Continuous time Markov Chains. *IEEE Transactions on Software Engineering*, 29(4):524–541, 2003.

[20] J. Rutten, M. Kwaiatkowska, G. Normal, and D. Parker. *Mathematical Techniques for Analyzing Concurrent and Probabilisitc Systems*, volume 23 of *CRM Monograph Series.* American Mathematical Society, 2004.

[21] M. Kwiatkowska, G. Norman, and D. Parker. Quantitative Analysis with the Probabilistic Model Checker PRISM. *Electronic Notes in Theoretical Computer Science*, 153(2):5–31, 2005. Elsevier.

[22] K. Sen, M. Viswanathan, and G. Agha. VESTA: A Statistical Model-Checker and Analyzer for Probabilistic Systems. In *Proc. IEEE International Conference on the Quantitative Evaluation of Systems*, pages 251–252, 2005.

[23] E.M. Clarke, O. Grumberg, and D.A. Peled. *Model Checking*. The MIT Press, 2000.

[24] A. Church. A Formulation of the Simple Theory of Types. *Journal of Symbolic Logic*, 5:56–68, 1940.

[25] R. Milner. A Theory of Type Polymorphism in Programming. *Journal of Computer and System Sciences*, 17:348–375, 1977.

[26] L.C. Paulson. *ML for the Working Programmer*. Cambridge University Press, 1996.

[27] M.J.C. Gordon and T.F. Melham. *Introduction to HOL: A Theorem Proving Environment for Higher-Order Logic*. Cambridge University Press, 1993.

[28] J. Harrison. *Theorem Proving with the Real Numbers*. Springer, 1998.

[29] R. Khazanie. *Basic Probability Theory and Applications*. Goodyear, 1976.

[30] M. DeGroot. *Probability and Statistics*. Addison-Wesley, 1989.

[31] S. Richter. *Formalizing Integration Theory, with an Application to Probabilistic Algorithms*. Diploma Thesis, Technische Universitat Munchen, Department of Informatics, Germany, 2003.

Table 1: HOL Symbols

HOL Symbol	**Standard Symbol**	**Meaning**
$\wedge$	*and*	Logical *and*
$\vee$	*or*	Logical *or*
$\sim$ t	$\neg t$	Not t
::	*cons*	Adds a new element to a list
num	$\{0, 1, 2, \ldots\}$	Positive Integers data type
real	All Real numbers	Real data type
λx.t	$\lambda x.t$	Function that maps x to $t(x)$
{x|P(x)}	$\{\lambda x.P(x)\}$	Set of all x that satisfy the property P
(a, b)	a x b	A pair of two elements
fst	fst (a, b) = a	First component of a pair
snd	snd (a, b) = b	Second component of a pair
suminf f(n)	$\lim_{k\to\infty}(\sum_{n=0}^{k} f(n))$	Infinite summation of a *real* sequence f
summable f(n)	$\exists x.\ \lim_{k\to\infty}(\sum_{n=0}^{k} f(n)) = x$	Infinite summation of f exists