Accepted for publication in **Software Tools for Technology Transfer**
(version 2014-09-11)

On Hypothesis Testing for Statistical Model Checking

Daniël Reijsbergen[1] **Pieter-Tjerk de Boer**[2] **Werner Scheinhardt**[2] **Boudewijn Haverkort**[2]

[1]University of Edinburgh, [2]University of Twente

(2014-09-11)

Abstract Hypothesis testing is an important part of Statistical Model Checking (SMC). It is typically used to verify statements of the form $p > p_0$ or $p < p_0$, where p is an unknown probability intrinsic to the system model and p_0 is a given threshold value. Many techniques for this have been introduced in the SMC literature. We give a comprehensive overview and comparison of these techniques, starting by introducing a framework in which they can all be described. We distinguish between three classes of techniques, differing in what type of output correctness guarantees they give when the true p is very close to the threshold p_0. For each technique, we show how to parametrise it in terms of quantities that are meaningful to the user. Having parametrised them consistently, we graphically compare the boundaries of their decision thresholds, and numerically compare the correctness, power and efficiency of the tests. A companion website allows users to get more insight in the properties of the tests by interactively manipulating the parameters.

1 Introduction

Statistical model checking (SMC) [27] is increasingly seen as a powerful alternative to numerical model checking. This is witnessed by two main developments. The first is the implementation of SMC techniques in classical model checking tools such as UPPAAL [8], PRISM [25] and MRMC [24]. The second is that several libraries explicitly dedicated to SMC techniques have recently been developed, e.g., COSMOS [7] and PLASMA [22]. The main reason behind this increase in popularity is the fact that SMC in many cases can avoid problems that have long plagued numerical model checking. These include the state space explosion problem (the memory requirements of SMC only depend on the high level description of the model) and the fact that numerical techniques that deal with more complicated models — e.g., Markov reward models or probabilistic timed automata with uniformly distributed transition times — quickly become computationally, i.e., numerically infeasible.

The core idea underlying statistical model checking is to use a computer program to repeatedly simulate the behaviour of the system model in order to say something about the system's performance in terms of a given performance measure. Throughout this paper this will be some probability of interest p.[1] The exact way in which these simulation runs are then interpreted depends on the interests of the investigator. First of all, she could be interested in a *quantitative* statement, consisting of an estimate of the performance measure with a corresponding *confidence interval* (e.g., with 95% confidence, the probability of deadlock before termination is 10% with a 2% margin of error). Secondly, she could be interested in a *qualitative* statement about a performance property, specified as a *hypothesis* that asserts that the true probability p is larger (or smaller) than some boundary value p_0 (e.g., with 95% confidence, the probability of deadlock before termination is greater than 5%).

The two approaches are closely related. Given a procedure to construct confidence intervals, one obtains a hypothesis test in the following way: construct the confidence interval, then check whether the boundary value p_0 is inside the interval. If not, accept or reject the assertion $p > p_0$ depending on whether p_0 is to the 'left' or 'right' of the interval. Despite this relationship, procedures for constructing confidence intervals are sometimes implemented completely in parallel to procedures exclusively focused on hypothesis testing.[2]

[1] A typical choice for p is the probability that some formula in a temporal logic such as PCTL [17] or CSL [3,5] is satisfied.

[2] An example of this is UPPAAL v4.1.18, in which the Sequential Probability Ratio Test (SPRT) of Section 3.4 is used for

In this paper we present a general framework for hypothesis testing that allows for a clear and intuitive comparison of both '*pure*' hypothesis tests and tests based on confidence intervals. We use the framework to describe, classify and parametrise the tests so far implemented in model checking tools, and introduce two tests that are new to the field of SMC, whose output guarantees are fundamentally different from those of the other tests. We also compare the tests empirically in a comprehensive case study. To help the reader get a feeling for the diffence between tests and the influence of parameters, we have built a companion website to this paper, where investigators can interactively modify parameters; see [1].

The structure of this paper is as follows. We present a single framework that allows comparison of the hypothesis tests discussed in this paper in Section 2, and discuss the main criteria by which to judge a test. In Section 3, we present an overview of these tests using the framework of Section 2. In Section 4, we discuss how these tests must be parametrised to ensure that the output guarantees are satisfied. We compare the performance of all these tests empirically in Section 5. Section 6 concludes the paper.

2 General Framework

In this section we discuss the framework that we use to compare the tests described in Section 3. We start in Section 2.1 with a discussion of the model setting; we focus particularly on the generality of the framework. Section 2.2 begins with a summary of elementary statistical methodology in order to fix terminology and notation; we then move on to discussing the features specific to this paper. Having discussed the framework for hypothesis tests, we focus on criteria for comparing hypothesis tests in Section 2.3, and introduce a classification of tests.

2.1 Model Setting

As mentioned in the introduction, we are interested in comparing the probability p to a given boundary value p_0. Typically, p denotes the probability that a formula expressed in a temporal logic is satisfied. With ϕ denoting the temporal logic formula, the performance property that we seek to evaluate is then often[3] expressed as $\mathcal{P}_{>p_0}(\phi)$.[4] Formally, this performance property holds in a state if, from that state, the probability that an *execution path* (generated randomly using the system specification) satisfies the property ϕ is greater than p_0. The only requirement on the system model that we want to apply to our tests is that we can randomly generate execution paths in order to obtain information about whether or not $\mathcal{P}_{>p_0}(\phi)$ is satisfied. This can be rewritten into the following requirements:

1. we can generate execution paths from the model, according to a well-defined probability measure on the execution paths;
2. with probability 1, these paths are generated in a finite amount of time and we can test, also in a finite amount of time, whether the property ϕ holds on a path; and
3. we either do not encounter nondeterminism [6], or we have a well-defined policy or scheduler to resolve it.

In principle, we do not need *any* additional information about the system model as long as we can obtain execution paths that satisfy these three requirements. A system about which additional information is not available is commonly called a *black-box* system [34].

In practice, a system model is often available that allows us to write a computer program that can generate execution paths, which means that the system is not completely black-box. Popular modelling formalisms include Generalized Semi-Markov Processes (GSMPs, [14, 28]) and stochastic (possibly *non-Markovian*) Petri nets [16]. Requirements 2) and 3) will not be satisfied in all GSMPs or stochastic Petri nets. For example, if the property ϕ does not involve a time bound then 2) may be violated, e.g., when the system reaches a bottom strongly connected component that does not contain termination states. Also, 3) may be violated in a GSMP when two transitions are scheduled to occur at the same time, e.g., when some of the transitions have deterministic delays. However, even in such cases it might still be possible to apply a refined form of statistical model checking (see, for example, [12], [35] or [40], in which 2) is not satisfied). Judging whether requirements 2) and 3) are satisfied given a system model and performance property is a field of study in itself. As this paper is not about generating sample paths but about the interpretation of the results, we refer the interested reader to the vast literature on stochastic simulation [13,31], and from now on assume that we draw samples from a black-box system in order to say something about $\mathcal{P}_{>p_0}(\phi)$.

In this paper, we will not consider nested probabilistic operators. To read about how nested operators are treated in other settings, see, e.g., Section 3.2 of [35] or [41], in which a combined numerical/statistical procedure is used.

qualitative statements, while the Chow-Robbins procedure of Section 3.1 is used only for quantitative statements.

[3] For example, in the logics pCTL [17], CSL [2], UTSL [43] and CSRL [4].

[4] By treating $\mathcal{P}_{>p_0}(\phi)$, we treat, without loss of generality, all possible variations of the probabilistic path operator $\mathcal{P}$, because, using statistical model checking, we cannot differentiate between $\mathcal{P}_{>p_0}(\phi)$ and $\mathcal{P}_{\geq p_0}(\phi)$ (more on that later in Section 2.2), and because $\mathcal{P}_{<p_0}(\phi)$ and $\mathcal{P}_{\geq 1-p_0}(\neg\phi)$ are equivalent.

2.2 Statistical Framework

With $i = 1, \ldots, N$, let ω_i be the execution path in the ith sample, and define

$$X_i \triangleq \mathbf{1}_\phi(\omega_i) = \begin{cases} 1 & \text{if } \phi \text{ holds on } \omega_i, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then X_i has a Bernoulli distribution with parameter p, where p denotes the true probability that ϕ is satisfied. This means that

$$\mathbb{P}(X_i = x) = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = 0. \end{cases}$$

The total sample $\mathbf{X} \triangleq (X_i)_{i=1,\ldots,N}$ will be used to perform a statistical *test*. To do this, we combine all relevant information from the individual sample paths into a function that maps $\{0,1\}^N$ onto $\mathbb{R}$, called the *test statistic*. We use the test statistic to falsify claims about p, called *hypotheses*. If we can show that, under the condition that some hypothesis H is true, the probability that the observed outcome of the test statistic occurs is smaller than some given $\alpha \in (0, \frac{1}{2})$, then we *reject* H. The parameter α is called the *significance parameter*, and $1-\alpha$ is called the *confidence* of the test. A hypothesis that can be rejected this way is called a *null hypothesis*, while a hypothesis that can be accepted through the rejection of a null hypothesis is called an *alternative hypothesis*. Rejecting a valid null hypothesis is called an *error of the first kind* (or a *false positive*). Not accepting a valid alternative hypothesis is called an *error of the second kind* (or a *false negative*).

Since we are interested in checking whether $\mathcal{P}_{>p_0}(\phi)$ holds, there are two relevant claims: $p > p_0$ and $p \leq p_0$. There is no clear distinction between a null and alternative hypothesis, as there is no asymmetry in our desire to reject any one of the two claims. Accordingly, we specify *two* alternative hypotheses, each of which we would like to accept if it were true:

$$\begin{aligned} H_{+1} &: p > p_0, \\ H_{-1} &: p < p_0. \end{aligned} \tag{2}$$

Additionally, we have the null hypothesis

$$H_0 : p = p_0.$$

Note that the null hypothesis *cannot* be shown to be correct, as its negation $p \neq p_0$ cannot be disproved statistically. The reason is that no matter how many samples we draw and no matter how much evidence we see for $p = p_0$, there will always be some small ϵ such that we cannot reject the claim that $p = p_0 + \epsilon$. However, H_0 *can* be shown to be *in*correct.

The procedure to test which of the alternative hypotheses is true is as follows: after having drawn N samples, we let $S_N(\mathbf{X})$ be the test statistic given by the sum of X_1 up to X_N, i.e.,

$$S_N(\mathbf{X}) = \sum_{i=1}^{N} X_i,$$

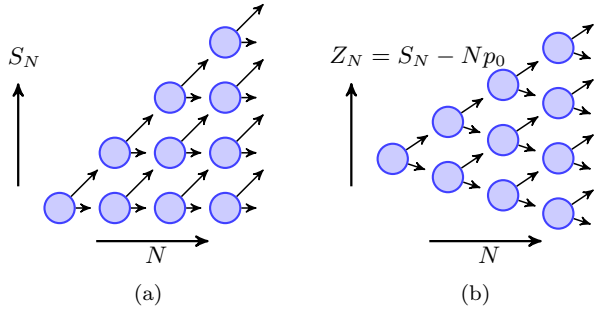

Figure 1: Markov chain representations of the processes S_N and Z_N. In each state, both processes jump up with probability p and down with probability $1-p$. The two processes have the same structure; the only difference is a normalisation of the variable on the vertical axis.

and omit the argument $\mathbf{X}$ for brevity. We can then view the evolution of S_N as the evolution of a discrete-time Markov chain on state space $\mathbb{N}^2$, with the number of drawn samples on the x-axis and the value of the test statistic on the y-axis, where in each step we take a jump to the right or top-right (as can be seen in Figure 1a).

While we are drawing samples, the expected behaviour of the process S_N is that it drifts away from the x-axis. The true parameter p determines the speed of this drift. Remember that our main interest is to test whether $p - p_0$ is positive or negative. Hence, we focus on the *shifted* test statistic

$$Z_N \triangleq S_N - Np_0.$$

The process Z_n is essentially a random walk that always jumps up by $1-p_0$ with probability p, or down by p_0 with probability $1-p$. Its evolution is depicted in Figure 1b. The speed at which Z_N drifts away from the x-axis is completely determined by $p - p_0$. If $Z_N \gg 0$ then this is strong evidence for H_{+1}, while if $Z_N \ll 0$ then this is strong evidence for H_{-1}.

We then specify *four* test decision areas which are subsets of $\mathbb{R}^2$. Three of them are called *critical*, which means that we draw a conclusion as soon as they are entered by Z_N. The first critical area $\mathcal{U}$ is the area such that as soon as Z_N enters $\mathcal{U}$, we accept H_{+1}. The second critical area $\mathcal{L}$ does the same for H_{-1}. As soon as Z_N enters the critical area $\mathcal{I}$, we stop the test without accepting any hypothesis. We accordingly say that the test was *inconclusive*. All that is outside these three areas makes up the *non-critical* area $\mathcal{NC}$.

The tests that we consider in this paper are completely determined by the shape of these areas. There are two main types of tests: *fixed sample size tests*, where the decision is taken after an a-priori determined number of samples, and *sequential tests*, where the decision whether or not to continue sampling is made on the basis of the samples so far; mixtures of both types are also possible. Typical examples of both types are illustrated

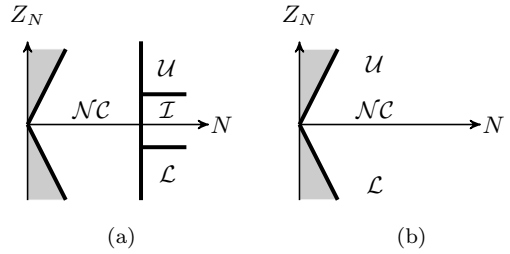

Figure 2: Graphical representation of the test decision areas $\mathcal{L}$, $\mathcal{U}$, $\mathcal{I}$ and $\mathcal{NC}$. Left: a typical fixed sample size test. Right: a typical sequential test. Grey areas represent areas in which Z_N cannot go.

in Figure 2. The relevant parts of these figures are the boundaries between the area $\mathcal{NC}$ and the three other areas. After all, we only continue the testing procedure while we are in $\mathcal{NC}$; we stop when a relevant boundary is crossed. For example, the exact shape of the border between $\mathcal{U}$ and $\mathcal{I}$ in Figure 2a is irrelevant because we stop when we enter either.

In a fixed sample size test, as illustrated in Figure 2a, the borders between $\mathcal{NC}$ and the other areas is a single straight line because for a fixed sample size test, the point at which we stop is always at the same value for N on the x-axis. In a sequential test like in Figure 2b, the important thing is that $\mathcal{NC}$ continues indefinitely, since we keep sampling until we draw a conclusion. Note that this also implies that typical sequential tests in principle do not have an area $\mathcal{I}$.[5] Hence, the structure of the sequential tests is entirely determined by two borders: the $\mathcal{L}$-$\mathcal{NC}$ boundary, denoted by $l(N)$, and the $\mathcal{U}$-$\mathcal{NC}$ boundary, denoted by $u(N)$. Most of the discussion of the sequential tests will therefore focus on the shape of these functions. For fixed sample size tests on the other hand, we merely need to determine two numbers, u^* and l^*, which depend on the chosen sample size, but are not functions of the number of samples N drawn at present.

Given $\mathcal{L}, \mathcal{U}$ and $\mathcal{I}$, we want to bound the probability that these areas are entered given that a hypothesis is valid. To formalise this, for $i \in \{-1, +1\}$, let A_i be the event that we reject H_0 in favour of H_i, and let A_0 be the event that we do not reject H_0, meaning that the test remains inconclusive. More specifically we have

$$\begin{aligned} A_{+1} &= \{\text{reach } \mathcal{U} \text{ before } \mathcal{L} \text{ or } \mathcal{I}\},\\ A_{-1} &= \{\text{reach } \mathcal{L} \text{ before } \mathcal{U} \text{ or } \mathcal{I}\},\\ A_0 &= \{\text{reach } \mathcal{I} \text{ or stay in } \mathcal{NC}\},\\ \neg A_{+1} &= A_{-1} \cup A_0,\\ \neg A_{-1} &= A_{+1} \cup A_0. \end{aligned}$$

[5] In practice, we might set a time-out parameter τ, thus letting all states that are to the right of τ be part of $\mathcal{I}$.

Then we typically impose the following two conditions on the two errors of the first kind ('*false positives*'):

$$\mathbb{P}(A_{+1} \mid \neg H_{+1}) \leq \alpha_1, \tag{3}$$
$$\mathbb{P}(A_{-1} \mid \neg H_{-1}) \leq \alpha_2. \tag{4}$$

These probabilities deal with drawing a *wrong* conclusion. We will usually bound these probabilities by replacing the condition $\neg H_{+1}$ (or $\neg H_{-1}$) by the worst case, which is H_0. A more detailed explanation of this is given in Section 3.1.

Also we like to impose conditions on the two errors of the second kind ('*false negatives*'):

$$\mathbb{P}(\neg A_{+1} \mid H_{+1}) \leq \beta_1, \tag{5}$$
$$\mathbb{P}(\neg A_{-1} \mid H_{-1}) \leq \beta_2. \tag{6}$$

These probabilities deal with drawing *no* (or a *wrong*) conclusion. For tests that always draw a conclusion (i.e., where A_0 never happens), these probabilities coincide with the ones in (3) and (4), assuming that H_0 is never exactly true. For tests that may end inconclusively, the probabilities in (5) and (6) are usually only slightly larger than the probability of A_0 (given H_{+1} or H_{-1}, respectively) since, e.g. under H_{+1} the event A_{-1} is much less likely than A_0. This is the reason we will, instead of (5) and (6), often use the *power*, which is a function of the real value of p, and is defined[6] as

$$\mathbb{P}(A_{-1} \cup A_{+1}) = 1 - \mathbb{P}(A_0).$$

Throughout this paper, we will choose $\alpha = \alpha_1 = \alpha_2$ and $\beta = \beta_1 = \beta_2$ for simplicity. In principle, the total probability of error of the first kind is $\alpha_1 + \alpha_2$, since if H_0 is true, accepting either H_{+1} or H_{-1} constitutes an error. But if H_{+1} or H_{-1} is true, the probability of error of the first kind is only α_2 or α_1, respectively. We argue that we should focus on the latter case. This is not to say that H_0 *cannot* hold in practice. However, if it does, then statistical model checking cannot be used to show it holds, as argued earlier. Thus, an investigator who wants to know whether H_0 is true should use a different model checking technique. Furthermore, an investigator who does *not* care about H_0 probably does not mind either H_{+1} or H_{-1} being accepted in that case. Throughout this paper, we will therefore assume $\alpha = \alpha_1 = \alpha_2$. An investigator who *does* care about H_0 could replace α by $\alpha/2$ in all tests[7] at the cost of increasing the computational effort.

[6] This corresponds to the definition of power in [11] and the general statistical literature, but note that in our case rejecting H_0 does not necessarily mean that we draw the correct conclusion since we have *two* alternative hypotheses.

[7] except the Sequential Probability Ratio Test (SPRT) of Section 3.4 and the Gauss-SSP test of Section 3.5 (whose strong assumptions preclude the validity of H_0).

2.3 Main criteria and classification of tests

Given a selection of tests specified using the framework of Section 2.2, it is up to the investigator to decide which test she finds the most appealing. We use three main criteria by which to judge the appeal of these tests:

1. the *correctness:* we call a test correct if its probability of not drawing the correct conclusion is guaranteed to be smaller than α, where $1-\alpha$ is the confidence level; mathematically, this means (3) and (4) hold;
2. the *power:* recall the definition of power from Section 2.2 as the probability that the test will eventually draw a conclusion, i.e., $1-\mathbb{P}(A_0)$;
3. the *efficiency:* the number of samples needed (in expectation) before a conclusion can be drawn.

As these three criteria are partly contradictory, each test will be affected adversely on at least one criterion when p is close to p_0. We introduce three classes of tests, based on which criterion is affected (most):

I. Tests whose probability of drawing a *wrong conclusion* exceeds α when $|p-p_0|$ is small.
II. Tests whose probability of drawing *no conclusion* (or a wrong conclusion) exceeds β when $|p-p_0|$ is small.
III. Tests that are always correct and always draw a conclusion, at the cost of drawing an *extremely large number of samples* before reaching a conclusion when $|p-p_0|$ is small.

Note that this classification in itself is independent of the *type* of test as described in the previous subsection, namely fixed sample size or sequential. However, it is worth mentioning at this point that a fixed sample size test that satisfies criterion 1 can never also satisfy criterion 2, at least not for all possible values p close to p_0. Such a test will therefore always be in class II. In other words, tests in class III, that satisfy both criteria 1 and 2, are necessarily sequential tests.

For each class, we introduce an extra input parameter, which influences how poor the performance will be when $|p-p_0|$ is small. For classes I and II, the extra parameter is a threshold on $|p-p_0|$, below which the investigator no longer cares about the test's correctness or the power, respectively. We call these parameters the *correctness-indifference level* δ for class-I tests, and the *power-indifference level* ζ for class II. Class-III tests do not need such a threshold parameter, since their correctness and power do not suffer when $|p-p_0|$ is small; however, they may use a *guess* called γ, representing the investigator's expectation of $|p-p_0|$, to minimise the runtime for that case.

We emphasise that, although the three parameters δ, ζ, and γ are all related to the difference between p and p_0, their meaning is different. The choice of δ or ζ (in class I/II tests) depends on the interest of the investigator (namely, in what case she no longer cares either about the correctness or the probability of receiving a meaningful answer), while the choice of γ depends on her expectation of the true p, and only influences the running time, but never the correctness or power.

All of the above is summarised in Table 1, which also shows the tests we will consider in Section 3, including their classes and types.

3 Overview of the Tests

In this section, we discuss several hypothesis tests that an investigator can choose to use. In particular, we focus on how they fit into the framework of Section 2.2. For a quick overview we refer to Table 1 (a more detailed overview follows later in Table 5). How these tests can be expressed in terms of the parameters of Section 2.3 is the subject of Section 4. The first five tests (which belong to classes I and II), have been implemented in existing model checking tools, or are described in the model checking literature, while the others (belonging to class III), to the best of our knowledge, are not.

The outline of this section is as follows. Starting with class II tests, we begin in Section 3.1 with a discussion of a hypothesis testing procedure that uses a confidence interval based on the Gaussian approximation and a sample size that is fixed beforehand. In Section 3.2 we focus on a similar method based on the Chernoff-Hoeffding bound. In Section 3.3, we discuss the Chow-Robbins test, which is based on confidence intervals that are sequential in the sense that we continue sampling until the width of the confidence interval has reached a given value. Turning to class I tests, we discuss the Sequential Probability Ratio Test in Section 3.4, followed by its 'fixed sample size variant', the Gauss-SSP test, in Section 3.5. In Sections 3.6 and 3.7 we discuss the two tests in class III, namely the Azuma test and the Darling-Robbins test, respectively. These two tests have not been implemented in model checking tools so far. Finally, in Section 3.8 we briefly discuss some noteworthy tests that have been proposed but (to the best of our knowledge) never implemented.

3.1 Binomial and Gaussian Confidence Intervals

The idea behind the test described in this section is the *confidence interval* based on an *a priori* fixed sample size N. Formally, a $(1-\alpha)$-confidence interval is an interval $[l,u]$ that is constructed using a procedure that, with probability $1-\alpha$, produces intervals containing the true probability p. As we argued in the introduction, a confidence interval can be used for a hypothesis test by checking if p_0 is inside the interval.

The critical regions for this test have the form displayed in Figure 2a. Since the number of samples drawn is fixed to be N, the non-critical region $\mathcal{NC}$ consists of all points (n,z) for which $n<N$. The other regions can

	Class I	Class II	Class III
Risk when $p \approx p_0$	correctness: wrong conclusion, i.e., error of first kind (& efficiency)	power: no conclusion, i.e., error of second kind (& efficiency)	efficiency: large running time
Parameter	correctness-indifference level δ	power-indifference level ζ	guess γ
Fixed sample size tests	Gauss-SSP	Gauss-CI Chernoff-CI	
Mixed tests		Chow-Robbins	
Sequential tests	SPRT		Azuma Darling

Table 1: Overview of test classes

be characterised by two values, namely l^*, which is the border between $\mathcal{L}$ and $\mathcal{I}$, and u^* which is the border between $\mathcal{I}$ and $\mathcal{U}$. According to (3), we must choose u^* such that when H_0 or H_{-1} is true, the probability that $Z_N > u^*$ is smaller than α. As we already mentioned in Section 2.2, it is sufficient to check this under the worst case assumption, i.e., whether

$$\mathbb{P}(Z_N > u^* | H_0) < \alpha \tag{7}$$

holds. The reason is that under H_{-1} (i.e., for any true $p \leq p_0$), high values of Z_N are even less likely than under H_0 (when $p = p_0$), so that $\mathbb{P}(Z_N > u^* | \neg H_{+1}) \leq \mathbb{P}(Z_N > u^* | H_0)$. Hence (7) implies (3). Analogously, we base l^* only on H_0 and not on H_{+1}.

If N is large enough, we can use the CLT to argue that the distribution of Z_N can be well approximated by a normal distribution. Let Φ be the standard normal cumulative distribution function and $\mathrm{Var}(Z_N) = \mathrm{Var}(Z_N | H_0) = Np_0(1-p_0)$, then it follows from basic statistical analysis (see [29] for details) that

$$l^* = \Phi^{-1}(\alpha) \sqrt{\mathrm{Var}(Z_N)}, \tag{8}$$

$$u^* = \Phi^{-1}(1-\alpha) \sqrt{\mathrm{Var}(Z_N)} = -l^*. \tag{9}$$

Note that the procedure above is *not* exactly the same as constructing a confidence interval and checking whether p_0 is inside the interval. The one difference is that under H_0, we can assume that the variance of both S_N and Z_N is given by $Np_0(1-p_0)$, while for a regular confidence interval this would be estimated using the realisation of S_N, i.e., $\mathrm{Var}(Z_N) = S_N(1 - S_N/N)$. This difference is only noticeable when $|p - p_0|$ is large.

In this paper we call the test described above the '*Gauss-CI*' test because of its relationship with the Gaussian confidence interval obtained using the CLT. Alternatively, confidence intervals can be based on the exact binomial distribution; they are called '*Clopper-Pearson*' intervals in the scientific literature. A third alternative exists in the form of the '*Agresti-Coull*' confidence intervals, which are between the binomial and Gaussian confidence intervals in terms of the degree of approximation – such intervals have been implemented in the tool MRMC. Hypothesis tests can also be based on such confidence intervals, but since the difference with Gaussian intervals is only noticeable at very small N, we will not separately consider such tests in this paper.

The choice of N is non-trivial. It impacts both the efficiency (obviously) and the power. In Section 4.1, we demonstrate how to determine N such that for a given power-indifference level ζ the power of the test is guaranteed to be at least $1 - \beta$.

3.2 Confidence Intervals using the Chernoff-Hoeffding Bound

The test described in this section is a fixed sample size test based on a different type of confidence interval. Its basis is the Chernoff-Hoeffding bound [20], which states the following: for any sequence $X_1, X_2, \ldots, X_N$ of independent random variables with $\mathbb{P}(0 \leq X_i \leq 1) = 1$, it holds for all $t > 0$ that

$$\mathbb{P}(|\bar{X} - \mathbb{E}(\bar{X})| > t) \leq 2e^{-2Nt^2}, \tag{10}$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. A test that is analogous to the Gauss-CI test of Section 3.1 is then as follows. The investigator chooses a significance parameter α and a so-called 'approximation parameter' ϵ. She then draws

$$N = \frac{1}{2\epsilon^2} \log\left(\tfrac{2}{\alpha}\right) \tag{11}$$

samples. We can then rewrite (10) to

$$\mathbb{P}(|\bar{X} - p_0| \geq \epsilon) \leq \alpha. \tag{12}$$

The test is then as follows: we draw N samples and check if $|\bar{X} - p_0| > \epsilon$. If so, we reject the null hypothesis, otherwise the test is inconclusive. If we reject the null hypothesis, we accept H_{+1} if $\bar{X} > p_0$ and we accept H_{-1} otherwise. The test satisfies (3) and (4) because under the null hypothesis $\mathbb{E}(\bar{X}) = p_0$, so that (12) is really an upper bound for the probability of rejecting the null hypothesis when it is valid. A test of this form is implemented in the tool PRISM. Since we assume that H_0 does not hold, as we argued in Section 2.2, we replace

$\frac{2}{\alpha}$ in (11) by $\frac{1}{\alpha}$ when we compute N for the tables in Section 5.

As with the Gauss-CI test, the shape of the critical regions is as displayed in Figure 2a. Apart from α and p_0, the main parameter that determines the location of the critical region boundaries is ϵ. Since ϵ does not have a clear interpretation in terms of the output guarantees described in Section 2.3, we discuss in Section 4.2 how to calculate it from a power-indifference level ζ instead. As with the Gauss-CI test, ζ will turn out to impact both the power and efficiency of the test.

3.3 Chow-Robbins Test

The test described in this section is similar to the test described in Section 3.1, but the difference is that we continue drawing samples until the width of the confidence interval for $\hat{p}_N = S_N/N$ has reached some given value, denoted by 2ϵ, at confidence level $1-2\alpha$. Then H_{+1} can be accepted if this confidence interval is entirely above p_0, H_{-1} if it is entirely below p_0, and the test is inconclusive otherwise.

After having drawn N samples, the width of this confidence interval (at confidence level $1-2\alpha$) equals $2\Phi^{-1}(\alpha)\sqrt{\mathrm{Var}(\hat{p}_N)}$, where $\mathrm{Var}(\hat{p}_N) = \hat{p}_N(1-\hat{p}_N)/N$. This width is maximal when $\hat{p} = \frac{1}{2}$ and is smaller when $\hat{p}_N$ is closer to 0 or 1. Hence, this test can reach a conclusion more quickly than the Gauss-CI test when p is further away from $\frac{1}{2}$ than p_0, and takes longer otherwise. We call this test the '*Chow-Robbins test*' after the authors of [10], who showed that a confidence interval created this way asymptotically satisfies the requirements on the errors of the first kind.

The critical areas of this test do not look like those depicted in Figure 2. It is between a fixed sample size test and a sequential test: even though the sample size is clearly not fixed, the sample size is upper bounded as there is a maximal N for which the confidence interval reaches the specified width even if the variance of $\hat{p}_N$ is maximal (i.e., when $\hat{p}_N = \frac{1}{2}$). The exact shape of the critical regions is discussed further in Section 5.1.

What is left is choosing the half-width of the confidence interval, denoted by ϵ because of analogy with ϵ in the Chernoff-CI test of Section 3.2. The parameter ϵ impacts both the power and the efficiency. In Section 4.3, we show how to choose ϵ based on the power-indifference level ζ.

3.4 Sequential Probability Ratio Test

The Sequential Probability Ratio Test (SPRT) for statistical model checking was introduced by Younes[8] in [42], based on ideas that go back to [37]. In [37], Wald tries to sequentially test which of the following two hypotheses is true,

$$\begin{aligned} H_{+1} &: p \geq p_{+1}, \\ H_{-1} &: p \leq p_{-1} \end{aligned} \tag{13}$$

for values $p_{-1} < p_{+1}$. He argues that a suitable test statistic is the so-called hypotheses' likelihood ratio:

$$T_N \triangleq \frac{p_{+1}^{S_N}(1-p_{+1})^{N-S_N}}{p_{-1}^{S_N}(1-p_{-1})^{N-S_N}}.$$

Clearly, small values of T_N speak in favour of H_{-1} while large values speak for H_{+1}. The idea is then to construct boundaries l' and u' such that when T_N crosses either of these boundaries we accept the corresponding hypothesis. We then have to bound, for given boundaries $l' < u'$, the probability of crossing l' given H_{+1} and the probability of crossing u' given H_{-1}. Wald showed how to achieve such a bound. In particular, for $l' = \alpha_1/(1-\alpha_2)$ and $u' = (1-\alpha_1)/\alpha_2$ one knows that the probability of accepting H_{-1} while H_{+1} is true is smaller than α_2 while the probability of accepting H_{+1} while H_{-1} is true is smaller than α_1.

To evaluate the validity of $\mathcal{P}_{>p_0}(\phi)$, we have the hypotheses of (2), which are similar to those of (13) with $p_{+1} = p_{-1} = p_0$. Unfortunately, in this case the value T_N is always 1. The idea proposed in [42] is to choose an *indifference* level δ such that we can safely assume that the true value for p is not inside the interval $[p_0-\delta, p_0+\delta]$. Then we can set $p_{-1} = p_0-\delta$ and $p_{+1} = p_0+\delta$ and carry out the above procedure. To be precise, the hypotheses in this setting[9] are given by

$$\begin{aligned} H'_{+1} &: p > p_0 + \delta, \text{ and} \\ H'_{-1} &: p < p_0 - \delta. \end{aligned} \tag{14}$$

To see how this test fits into the framework of Section 2.2, first note that instead of the test statistic T_N we could also use

$$\log T_N \triangleq q_1 S_N + q_2 N,$$

where

$$q_1 = \log\left(\frac{p_{+1}\cdot(1-p_{-1})}{(1-p_{+1})\cdot p_{-1}}\right),\ q_2 = \log\left(\frac{1-p_{+1}}{1-p_{-1}}\right).$$

Hence, an equivalent formulation is to use the process $Z_N = S_N - Np_0$ of Figure 2 as a test statistic, with boundaries

$$l(N) = \frac{1}{q_1}(\log l' - q_2 N) - Np_0, \tag{15}$$

$$u(N) = \frac{1}{q_1}(\log u' - q_2 N) - Np_0. \tag{16}$$

[8] We use slightly different terminology than the authors of [42]. They use H_0 for the H_{+1} of (13), they use H_{+1} for the H_{-1} of (13), and they use H_2 to denote $p \in [p_{-1}, p_{+1}]$. Furthermore, where they speak of a type 1 error and a type 2 error in the case of the SPRT, we speak of two errors of the first kind.

[9] A more general approach would be to set $H'_{+1} : p > p_0 + \delta_1$ and $H'_{-1} : p < p_0 - \delta_2$ with δ_1 and δ_2 not necessarily equal. We choose a symmetric indifference region for the sake of simplicity.

These are linear functions in N. So, whereas the boundaries of (8) and (9) are proportional to $\sqrt{N}$, the boundaries of (15) and (16) increase linearly. One can verify that when $p_0 = \frac{1}{2}$ or in the limit $\delta \downarrow 0$ the boundaries are constants.

The bounds on the error probabilities are only valid if p does not lie in $[p_0 - \delta, p_0 + \delta]$; consequently, δ impacts the correctness of the test. Furthermore, the efficiency is affected. Since δ is the only parameter, and has clear interpretations in terms of the output guarantees of Section 2.3, the parameter choice for this test is not discussed further in Section 4.

3.5 Gauss-SSP test

The test discussed in this section goes back to [15], was discussed in [38] and [35] and has been implemented in the tool VeStA and its offshoots. It can be seen as a fixed sample size version of the SPRT. As with the SPRT, we assume that p is outside the interval $[p_0 - \delta, p_0 + \delta]$ and, hence, consider the hypotheses of (14). The idea is then to draw N samples, with N fixed beforehand, and accept H'_{+1} if $Z_N \geq 0$ and to accept H'_{-1} otherwise. The sample size N is computed such that the requirements on the two errors of the first kind are met. To make this precise: we can write the first error of the first kind (given in the general setting by (3)) as follows:

$$\begin{aligned}\mathbb{P}(A_{+1} \mid H'_{-1}) &\leq \mathbb{P}(Z_N \geq 0 \mid p = p_0 - \delta) \\ &= \mathbb{P}\left(Y_N \geq \frac{N\delta}{\sqrt{\mathrm{Var}(Z_N)}} \;\middle|\; p = p_0 - \delta \right),\end{aligned}$$

where $\mathrm{Var}(Z_N) = N(p_0 - \delta)(1 - p_0 + \delta)$ if $p = p_0 - \delta$ and

$$Y_N = \frac{Z_N + N\delta}{\sqrt{\mathrm{Var}(Z_N)}}.$$

is a normalised version of Z_N. One obtains a similar expression for the second error of the first kind. In [35] the exact binomial distribution of Z_N is used to find an upper bound for these probabilities. In this paper, we use the fact that Y_N is approximately normally distributed for large N, which leads to the following requirements on N in order to bound the errors of the first kind:

$$N \geq \left(\frac{\Phi^{-1}(1 - \alpha_1)}{\delta} \right)^2 (p_0 - \delta)(1 - p_0 + \delta)$$

$$N \geq \left(\frac{\Phi^{-1}(\alpha_2)}{\delta} \right)^2 (p_0 + \delta)(1 - p_0 - \delta),$$

where Φ (as in Section 3.1) denotes the Gaussian cumulative distribution function.

As with the SPRT, the indifference parameter δ impacts both the correctness and efficiency of the test and because its interpretation is clear, this test is not discussed further in Section 4.

We call this test the 'Gauss-SSP' test; SSP stands for Single Sampling Plan as it was called in [43]. An SSP test variant that uses the Chernoff-Hoeffding bound instead of the Gaussian approximation is discussed in [18] (the authors call the method based on this test '*approximate model checking*'). This test can be called Chernoff-SSP, and compares to the Gauss-SSP in a way that is similar to how the Chernoff-CI test compares to the Gauss-CI; it will not be discussed here further.

3.6 Azuma Test

The test of this section is the first of two class-III tests to be discussed in this paper. These tests are different from the tests discussed previously in the sense that their input parameters only determine the efficiency of the test. So far, no class-III tests have been implemented in the model checking tools. These tests have the shape of the typical sequential test depicted in Figure 2b: they are characterised by functions $u(N)$ and $l(N)$ denoting the boundaries between $\mathcal{U}$ and $\mathcal{L}$ respectively and $\mathcal{NC}$. We assume that the tests are symmetric (i.e., $u(N) = -l(N)$), which means that $u(N)$ remains to be chosen such that (3–6) are satisfied.

The function $u(N)$ must asymptotically grow faster than $\sqrt{N}$, otherwise errors of the first kind will be too likely for small $|p - p_0|$. An informal argument is that the standard deviation of the process Z_N grows proportionally to $\sqrt{N}$, so that even under H_0, given an infinite amount of time such boundaries will eventually be crossed with probability 1. This is discussed in greater detail in [29,30]. Also, $u(N)$ must grow slower than linearly in N, otherwise errors of the second kind will be too likely for small $|p - p_0|$. The argument here is that even under one of the alternative hypotheses, the drift of Z_N is only linear, so that for $|p - p_0|$ small enough the function $u(N)$ will diverge linearly from the expected trajectory. As a result, the probability of ever crossing a linearly increasing $u(N)$, and thus taking a decision, is too small when $|p - p_0|$ is tiny.

The first shape of $u(N)$ that we consider is the form

$$u(N) = a(N + k)^b, \text{ with } b \in (\tfrac{1}{2}, 1).$$

For this case, both the correctness of the test and a lower bound on the power are proven in [29,30] using a bounding result that is comparable to, and inspired by, Ross' "generalized Azuma inequality" in Section 6.5 of [32] (which also explains the name of the test). In particular, (3), (4), (5) and (6) are all satisfied, with

$$\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = e^{-8(3b-2)a^2k^{2b-1}}. \qquad (17)$$

The input parameters of the test are a and k; we discuss in Section 4.4 how to choose these parameters, based on a guess γ which only affects the efficiency.

3.7 Darling-Robbins Test

In this section, we consider a test similar to the one described in the previous section but with a different form of $u(N)$. It is based on [11] (Theorem 3), in which the following statement is proven for the test of Section 2.2 for general $\mathcal{U}$-$\mathcal{NC}$ boundary $u(N)$ and $\mathcal{L}$-$\mathcal{NC}$ boundary $-u(N)$: if one can find an $\epsilon > 0$ such that

$$\sum_{n=1}^{\infty} e^{-\frac{u^2(n)}{n+1}} \leq \epsilon \tag{18}$$

then the probability of error is bounded from above by $2\sqrt{2}\epsilon$. If we assume that H_0 does not hold, then the probability of error can be upper bounded by $\sqrt{2}\epsilon$. The idea is then to carry out the test of Section 2.2, with $u(N)$ chosen such that (18) can be used to show that (3–6) hold.

The bound (18) can in principle also be applied to the test from the previous section, with $u(N) = a(N+k)^b$, but turns out to be much looser than the bound of (17). On the other hand, the proof in [29] of (17) requires analytical steps that do not work for boundaries that are of order $N^{2/3}$ or tighter, such as $N\log(N)$. So in order to evaluate such tighter boundaries, only (18) is available. Using this rather loose bound will negatively affect the efficiency of the resulting method.

In this paper we apply the test based on (18), which we call the '*Darling-Robbins*' or '*Darling*' test for brevity, only to boundaries of the form

$$u(N) = \sqrt{a(N+1)\log(N+k)}.$$

As with the Azuma test, the remaining input parameters of this are a and k; how they can be chosen based on a guess γ which only affects the efficiency is discussed in Section 4.4.

3.8 Other Tests

In this section, we mention some other hypothesis tests that could also be applied in the context of statistical model checking. None of these tests has been implemented in the major model checking tools, and we will not discuss them in the rest of this paper.

The first is the Bayesian SPRT which was proposed for statistical model checking in [23] and which is based on ideas going back to [21]. In Bayesian statistics, the true parameter p is itself seen as the realisation of a random variable, of which a prior distribution must be given, which affects not only the efficiency of the test but also its correctness. For a more detailed discussion of the Bayesian SPRT in our framework, see [29].

A second test is the one proposed in [26] and which is mentioned, among others, in [41]. The input of this test is a constant c which represents the relative cost of drawing a sample compared to the cost of accepting an invalid hypothesis. The critical areas are then constructed such that the expected cost is minimised in a Bayesian setting.

Finally, in [39] a variant of the SPRT is proposed that includes an inconclusive area $\mathcal{I}$ (thus, in our terminology, it essentially turns the SPRT from a class I test into a class II version). In fact, the test entails that two SPRT tests are performed *simultaneously* (i.e., based on the *same* sample path of Z_N), namely one testing $p \geq p_0 + \delta$ against $p \leq p_0$, and one testing $p \geq p_0$ against $p \leq p_0 - \delta$. At first sight this test seems to fit in the framework of Section 2.2, with a somewhat remarkable shape of $\mathcal{NC}$ (see Figure 1b. in [39]), but one needs to be careful here: since the sample path is not stopped when one of the subtests draws a conclusion, one should not just look where the process Z_N eventually ends up, but also take into account its whole sample path. Thus, when implemented correctly, the test does not formally fit in the framework of Section 2.2.

4 Choice of parameters

In Section 3, we discussed a range of tests in terms of the framework of Section 2.2; in particular, we focused on the general shape of the critical areas. We found that for each test, an additional parameter was still needed to be able to determine the exact shape of the critical areas. For the tests in class I (i.e., the SPRT and Gauss-SSP test of Sections 3.4 and 3.5), this was the indifference level δ. This parameter has a clear interpretation as discussed in Section 2.3; consequently, these tests do not further appear in this section. For the other tests we discuss how to choose their parameters such that they have clear interpretations in terms of the output guarantees of Section 2.3.

The class II tests are treated in Sections 4.1, 4.2 and 4.3, where we discuss how to parametrise the Gauss-CI, Chernoff-CI and Chow-Robbins tests respectively using the power-indifference level ζ. This replaces their previous parametrisation in terms of either the sample size N or the confidence interval width ϵ, of which the latter has a clear interpretation for making quantitative statements (i.e., confidence intervals) but less so for hypothesis testing. In Section 4.4, we discuss the parametrisation of the class III tests (Azuma and Darling) in terms of the guess γ.

4.1 Choice of parameters for the Gauss-CI test

In Section 3.1, we derived the expressions (8) and (9) for the critical region boundaries, with α, p_0 and N left as parameters. While the interpretation of α and p_0 is clear, the choice of N for the Gauss-CI test is non-trivial as it settles the trade-off between the power and the efficiency. If p is very different from p_0, a small value for N suffices — choosing N too large then leads to extra

inefficiency. Alternatively, if p is close to p_0 a large value for N is needed — choosing N too small then leads to a decrease in power. For making quantitative statements, the goal is often to choose N such that the width of the confidence interval has a certain value. But since we focus on hypothesis testing, we want a procedure for choosing N such that (5) and (6) are satisfied.

If $p - p_0$ were known to be equal to some given value $\zeta > 0$, then the minimal choice of N for which (5) and (6) are still satisfied can be calculated. For large N, $\hat{p}_N \triangleq S_N/N$ can be well approximated by a normally distributed random variable with mean $p_0 + \zeta$ and variance $\sigma^2 \triangleq (p_0 + \zeta)\,(1 - p_0 - \zeta)/N$. Writing $\xi = \Phi^{-1}(1-\alpha)$ and $\sigma^2_{H_0} \triangleq (p_0)\,(1-p_0)/N$, the probability of *not* being able to accept H_{+1} after drawing N samples is given by

$$\begin{aligned}&\mathbb{P}\left(\hat{p}_N \leq p_0 + \xi\sigma_{H_0}\right)\\ &= \mathbb{P}\left(\frac{\hat{p}_N - p_0 - \zeta}{\sigma} \leq \frac{\xi\sigma_{H_0} - \zeta}{\sigma}\right)\\ &= \Phi\left(\frac{\xi\sigma_{H_0} - \zeta}{\sigma}\right) = \Phi\left(\frac{\xi\sqrt{p_0(1-p_0)} - \zeta\sqrt{N}}{\sqrt{(p_0+\zeta)(1-p_0-\zeta)}}\right).\end{aligned} \tag{19}$$

Setting this equation equal to β and solving for N yields the following expression:

$$N_G = \left(\frac{\xi\sqrt{p_0(1-p_0)} - \Phi^{-1}(\beta)\sqrt{(p_0+\zeta)(1-p_0-\zeta)}}{\zeta}\right)^2.$$

An analogous procedure can be carried out for $p = p_0 - \zeta$, which means that we have two expressions for N. Taking the maximum of the two guarantees that if this many samples are drawn, (5) and (6) hold when $|p - p_0| > \zeta$.

4.2 Choice of parameters for Chernoff-CI test

For the Chernoff-CI test of Section 3.2, the remaining parameter is ϵ, which is related to the width of a confidence interval. Since this has an impact on the power, we use it to establish an upper bound on the error probability of the second kind. Assume, without loss of generality, that H_{+1} holds, so that $p = p_0 + \Delta$ for some $\Delta > 0$; outside the power-indifference region, we have $\Delta \geq \zeta$. Note that we can use (11) to write ϵ as ϵ_N, i.e., as a function of N. For an error of the second kind to occur it must hold that after N samples we have that $\bar{X} - p_0 < \epsilon_N$. We can use a form of the Chernoff-Hoeffding bound [20] and the fact that $\mathbb{E}(p_0 - \bar{X}) = -\Delta$ to establish

$$\begin{aligned}\mathbb{P}(\bar{X} - p_0 < \epsilon_N) &= \mathbb{P}(p_0 - \bar{X} > -\epsilon_N)\\ &= \mathbb{P}(p_0 - \bar{X} + \Delta > \Delta - \epsilon_N)\\ &\leq e^{-2N(\Delta-\epsilon_N)^2}.\end{aligned}$$

Setting $\beta = e^{-2N(\Delta-\epsilon_N)^2}$ means that (5) is valid. It has two solutions for N, one of which gives positive $\Delta - \epsilon_N$ (which is a requirement of the Chernoff-Hoeffding bound). Setting $\Delta = \zeta$ in this solution, we find the worst-case number of samples needed outside the power-indifference region:

$$N_C = \frac{2\sqrt{\log(\beta)\log(\alpha)} - \log(\alpha\beta)}{2\zeta^2}. \tag{20}$$

A similar argument can be made for $\Delta < 0$ and (6), which leads to the same value for N_C for both error probabilities of the second kind.

		$p_0 = 0.5$		$p_0 = 0.2$	
α	ζ	N_C	N_G	N_C	N_G
0.05	0.1	600	259	600	189
	0.025	9587	4199	9587	2785
	0.01	59915	26265	59915	17056
0.025	0.1	738	372	738	273
	0.025	11805	6035	11805	4012
	0.01	73778	37752	73778	24540

Table 2: Chernoff (N_C) and Gaussian (N_G) sample sizes, $\beta = \alpha$.

Table 2 compares the sample size for the Chernoff-CI test N_C and the sample size for the Gauss-CI test N_G, for the same values of α and β, calculated using (19). One thing to note is that the Chernoff-CI test's sample size does not depend on p_0, while the Gauss-CI test's sample size does. Another thing is that the sample size for the Chernoff-CI test always seems to be larger than for the Gauss-CI test.

4.3 Choice of parameters for Chow-Robbins test

For the Chow-Robbins test of Section 3.3, the only parameter left to choose is the (half-)width of the confidence interval ϵ, such that the error probability of the second kind is bounded as desired.

To obtain a value for ϵ we start by observing that $\hat{p}_N$ is approximately normally distributed with mean $p_0 + \zeta$ and standard deviation[10] $\sigma = \epsilon/\Phi^{-1}(1-\alpha) = -\epsilon/\Phi^{-1}(\alpha)$. The test will not accept H_{+1} if $\hat{p}_N \leq p_0 + \epsilon$, leading to (19) with $\xi\sigma_{H_0} = \epsilon$ substituted. Setting this to β, one finds

$$\epsilon = \frac{\zeta}{1 + \Phi^{-1}(\beta)/\Phi^{-1}(\alpha)}.$$

4.4 Choice of parameters for Azuma and Darling tests

Since the Azuma and Darling tests are closely related, we discuss their parameter choices together. These two tests have non-critical area upper boundaries $u(N; a, k)$

[10] This follows from the stopping criterion of the Chow-Robbins test, but it is only an approximation, not just because the number of samples N is finite, but also because N in Chow-Robbins' stopping criterion depends on the samples themselves, which violates an assumption of the central limit theorem.

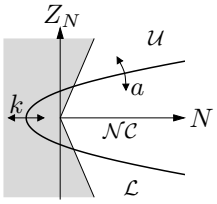

Figure 3: Impact of the parameters a and k on the shape of the critical regions of the Azuma test.

γ	Azuma a	Azuma k	Darling a	Darling k
10^{-1}	0.3274	$1.95{\cdot}10^{2}$	1.5973	$6.37{\cdot}10^{2}$
10^{-2}	0.1036	$1.95{\cdot}10^{4}$	1.3897	$5.95{\cdot}10^{4}$
10^{-3}	0.0328	$1.95{\cdot}10^{6}$	1.2934	$5.77{\cdot}10^{6}$
10^{-4}	0.0103	$1.94{\cdot}10^{8}$	1.2371	$5.74{\cdot}10^{8}$
10^{-5}	0.0033	$1.94{\cdot}10^{10}$	1.2000	$5.68{\cdot}10^{10}$
10^{-6}	0.0010	$1.92{\cdot}10^{12}$	1.1734	$5.74{\cdot}10^{12}$

Table 3: Approximately optimal parameter choices for $\alpha = 0.05$. For this table we used $b = \frac{3}{4}$

given by $a(N+k)^b$ and $\sqrt{a(N+1)\log(N+k)}$ respectively. The impact of the parameters a and k on the cone that defines $\mathcal{NC}$ is illustrated in Figure 3 for the Azuma test. The parameter a influences the increase in width of $\mathcal{NC}$ and its influence does not fade relative to N when N grows large. A high value of the parameter k on the other hand makes the area $\mathcal{NC}$ wider for small values of N. For the Darling test, the influence of these two parameters is similar. The Azuma test additionally depends on a parameter b; a high value for b means that the area $\mathcal{NC}$ boundary $u(N)$ will more closely resemble a straight line, which means that it will grow much wider asymptotically.

A high value for k makes it harder to accept an alternative hypothesis in the beginning, but — since a and b can be chosen smaller to maintain the same significance level α — easier to reject as N grows bigger. Since the upper bound on the probability of error is fixed to equal α, we can determine k as a function of a, α and b. For the Azuma test, we easily derive from (17) that

$$k_{\text{Azuma}}(a,\alpha,b) = \left(\frac{\log(\alpha)}{8a^2(2-3b)}\right)^{\frac{1}{2b-1}}.$$

For the Darling test, it is harder to obtain a similar expression from (18) since we have to solve for the lower bound of a summation, but for practical purposes the summation in (18) can be approximated by the integral

$$\int_1^\infty e^{-\frac{u^2(x)}{x+1}}\,dx.$$

We then derive

$$k_{\text{Darling}}(a,\alpha) = \left(\frac{\alpha(a-1)}{\sqrt{2}}\right)^{-\frac{1}{a-1}} - 1.$$

We then minimise the expected number of samples drawn, which we approximate using the intersection of the expected trajectory of Z_N and $u(N)$. This means that we have to solve

$$|p-p_0|N = u(N;a,k(a,\alpha)) \tag{21}$$

for N and then minimise over a. Unfortunately, both in the case of the Azuma and the Darling test, solving (21) for N does not lead to a closed form expression. However, in both cases we can do the minimisation numerically, since the function $u(N)-|p-p_0|N$ has a derivative simple enough to allow for Newton's method to find its roots. We seek the minimum of $N(a)$ for $a \in [0,\infty)$, but for the sake of being able to use straightforward numerical techniques, we search for the minimum of $N(\frac{1}{1+a})$ for $\frac{1}{1+a} \in (0,1]$. Since this is a bounded interval, we can use techniques such as golden section search [9] to find the minimum. For the Darling test we even know that $a > 1$, meaning that we can minimise $N(\frac{1}{a})$ on $(0,1]$.

In Table 3, we show the (approximately) optimal parameters a and k that we found for both tests for several values of γ (recall that this is our guess for $|p-p_0|$). We can see that for the Azuma test, a grows proportional to $\sqrt{\gamma}$, and k inversely proportional to γ^2.

The final remaining value to choose is then the parameter b of the Azuma test. A higher value for b means a tighter bound on the error probability of the first kind, but the area $\mathcal{NC}$ will grow larger asymptotically. The difference in terms of the tightness of the bound can be observed in Table 4, where we display the solutions to equation (21) for the Azuma test with several values of b and the Darling test (with a and k chosen optimally). The impact of a low value for b is twofold: the expected number of needed samples when the guess is correct will be higher, but the test will become less sensitive to the guess γ. Note, however, that even for very low values of b (e.g., 0.67), the Azuma test will still be more sensitive than the Darling test. Since for $b = 0.67$ the Azuma test has a higher expected number of needed samples than the Darling test, while it is still less sensitive, the Azuma test has no advantages over the Darling test so we can say that it performs strictly worse than the Darling test. The choice $b = 0.9$ on the other hand leads to enormous parameter sensitivity. Values of b around $\frac{3}{4}$ seem to strike a nice balance, and in Section 5, where we empirically validate the analysis of this section, we will only consider the Azuma test with this parameter choice.

By going through the above numerical procedure for a wide range of values of α and γ, for $b = 3/4$, and then fitting a function, we have obtained the following approximate solutions:

$$a_{\text{Azuma}} \approx (0.25 - 0.144\alpha^{0.15})\sqrt{\gamma/0.0243}$$

and

$$a_{\text{Darling}} \approx \exp\big(0.4913 - 0.0715x + 0.0988y - 0.00089x^2 + 0.00639y^2 - 0.00361xy\big),$$

$\lvert p-p_0\rvert$	γ	Azuma			Darling
		$b=0.67$	$b=0.75$	$b=0.9$	
10^{-1}	10^{-1}	$\mathbf{7.999{\cdot}10^3}$	$\mathbf{3.891{\cdot}10^2}$	$\mathbf{1.842{\cdot}10^2}$	$\mathbf{1.202{\cdot}10^3}$
	10^{-2}	$3.463{\cdot}10^4$	$1.828{\cdot}10^3$	$9.394{\cdot}10^2$	$1.532{\cdot}10^3$
	10^{-3}	$3.220{\cdot}10^5$	$1.720{\cdot}10^4$	$8.923{\cdot}10^3$	$2.015{\cdot}10^3$
	10^{-4}	$3.199{\cdot}10^6$	$1.708{\cdot}10^5$	$8.801{\cdot}10^4$	$2.496{\cdot}10^3$
10^{-2}	10^{-1}	$4.736{\cdot}10^6$	$1.150{\cdot}10^6$	$9.200{\cdot}10^9$	$1.946{\cdot}10^5$
	10^{-2}	$\mathbf{7.999{\cdot}10^5}$	$\mathbf{3.892{\cdot}10^4}$	$\mathbf{1.842{\cdot}10^4}$	$\mathbf{1.716{\cdot}10^5}$
	10^{-3}	$3.461{\cdot}10^6$	$1.827{\cdot}10^5$	$9.383{\cdot}10^4$	$2.018{\cdot}10^5$
	10^{-4}	$3.222{\cdot}10^7$	$1.718{\cdot}10^6$	$8.844{\cdot}10^5$	$2.495{\cdot}10^5$
10^{-3}	10^{-1}	$5.072{\cdot}10^9$	$1.150{\cdot}10^{10}$	$9.200{\cdot}10^{19}$	$2.735{\cdot}10^7$
	10^{-2}	$4.732{\cdot}10^8$	$1.151{\cdot}10^8$	$9.223{\cdot}10^{11}$	$2.360{\cdot}10^7$
	10^{-3}	$\mathbf{7.999{\cdot}10^7}$	$\mathbf{3.892{\cdot}10^6}$	$\mathbf{1.842{\cdot}10^6}$	$\mathbf{2.218{\cdot}10^7}$
	10^{-4}	$3.463{\cdot}10^8$	$1.825{\cdot}10^7$	$9.304{\cdot}10^6$	$2.500{\cdot}10^7$
10^{-4}	10^{-1}	$5.439{\cdot}10^{12}$	$1.150{\cdot}10^{14}$	$9.200{\cdot}10^{29}$	$3.511{\cdot}10^9$
	10^{-2}	$5.069{\cdot}10^{11}$	$1.151{\cdot}10^{12}$	$9.223{\cdot}10^{21}$	$3.034{\cdot}10^9$
	10^{-3}	$4.735{\cdot}10^{10}$	$1.152{\cdot}10^{10}$	$9.318{\cdot}10^{13}$	$2.815{\cdot}10^9$
	10^{-4}	$\mathbf{7.999{\cdot}10^9}$	$\mathbf{3.892{\cdot}10^8}$	$\mathbf{1.842{\cdot}10^8}$	$\mathbf{2.711}{\cdot}10^9$

Table 4: For each combination $(\gamma, |p-p_0|, \text{test type})$, we display the solution to (21) — i.e., the N for which the expected trajectory leaves $\mathcal{NC}$ — with parameters a and k chosen optimally. Bold values imply that $\gamma = |p-p_0|$, i.e., that the guess is correct. In all cases, $\alpha = 0.05$.

with $x = \log\alpha$ and $y = \log\gamma$. Note that we have not thoroughly quantified the precision of the above approximations. However, they need not be very precise: after all, these calculations are only used to optimise the convergence speed for a guess for $\gamma = |p = p_0|$, and that guess will typically be imprecise by itself; furthermore, any error in the calculation only affects the efficiency of the test, not its correctness. Thus, simple approximations like the above can suffice for use in tools.

5 Results and Comparisons

In this section we compare the performance of the tests discussed in Section 3 — see Table 5 for a summary. We do this in two ways: we will begin in Section 5.1 by comparing the tests in terms of the implied test decision areas as discussed in Section 2.2, and see how these areas behave as a function of the number of samples drawn. In Section 5.2, we will then compare the tests by the three performance measures mentioned in Section 2.3: the correctness, the power and the efficiency. In Section 5.3, we discuss the implementation of the tests in the model checking tools.

5.1 Shape of the Non-Critical Areas (NC)

As was explained in Section 2.2, all of the tests in this paper can be considered in the context of a single framework: a random walk Z_n that always jumps up by $1-p_0$ with probability p or down by p_0 with probability $1-p$. The tests can then be defined in terms of the boundaries of the test decision area $\mathcal{NC}$, as sketched in Figure 2. In Figures 4 and 5, we compare the shapes of these boundaries for all tests introduced before. For tests that can end inconclusively, the boundary of the corresponding decision area $\mathcal{I}$ is drawn as a grey line.

Figure 4 shows the decision boundaries for the symmetrical situtation $p_0 = \frac{1}{2}$. For the parametrisation, the accepted error probabilities of first and second kind α and β are set to 0.05. The indifference parameters δ and ζ are set to 0.025, while the guess γ for $p-p_0$ is 0.1. Note that choosing $\gamma > \delta$ makes sense: it expresses the investigator's guess that p is 0.1 away from p_0 (and that she wishes to optimise the Darling and Azuma tests for that case), but also that she wishes to have reliable results even if p turns out to be only 0.025 away from p_0.

First, consider the Darling and Azuma tests. Although they never terminate inconclusively, they may take very long if p is very close to p_0. Comparing their $\mathcal{NC}$ regions, we see that it is narrower for the Azuma test than for the Darling test for small values of N, but the Azuma boundaries eventually overtake those of the Darling test; this is obvious as functions of the type $N^{\frac{3}{4}}$ are asymptotically wider than those of type $\sqrt{N\log(N)}$.

The SPRT is also a sequential test and may theoretically take indefinitely long. However, its $\mathcal{NC}$ region is narrow, so long runs are unlikely. The price for this is that the SPRT may draw an incorrect conclusion with probability more than α if the true p is not at least δ away from p_0.

Test	Class	Type	Input*	Source	Section	Comments
Gauss-CI	II	fixed N	ζ	CLT	3.1, 4.1	may terminate inconclusively
Chernoff-CI	II	fixed N	ζ	[18], [20]	3.2, 4.2	same guarantees as Gauss-CI, but less efficient
Chow-Robbins	II	mixed	ζ	[10]	3.3, 4.3	same, but more or less efficient depending on p and p_0
SPRT	I	sequential	δ	[37], [42]	3.4	increased risk of drawing wrong conclusion if $\|p-p_0\| < \delta$
Gauss-SSP	I	fixed N	δ	[35]	3.5	same risk as SPRT
Azuma	III	sequential	γ	[29], [30]	3.6, 4.4	error probabilities guaranteed; takes long if $p \approx p_0$
Darling-Robbins	III	sequential	γ	[11]	3.7, 4.4	like Azuma, but efficiency rather insensitive to guess γ

*See 2.3 for details; summary: δ = indifference level for correctness; ζ = indifference level for power; γ = guess only used for efficiency optimisation.

Table 5: Summary of the tests.

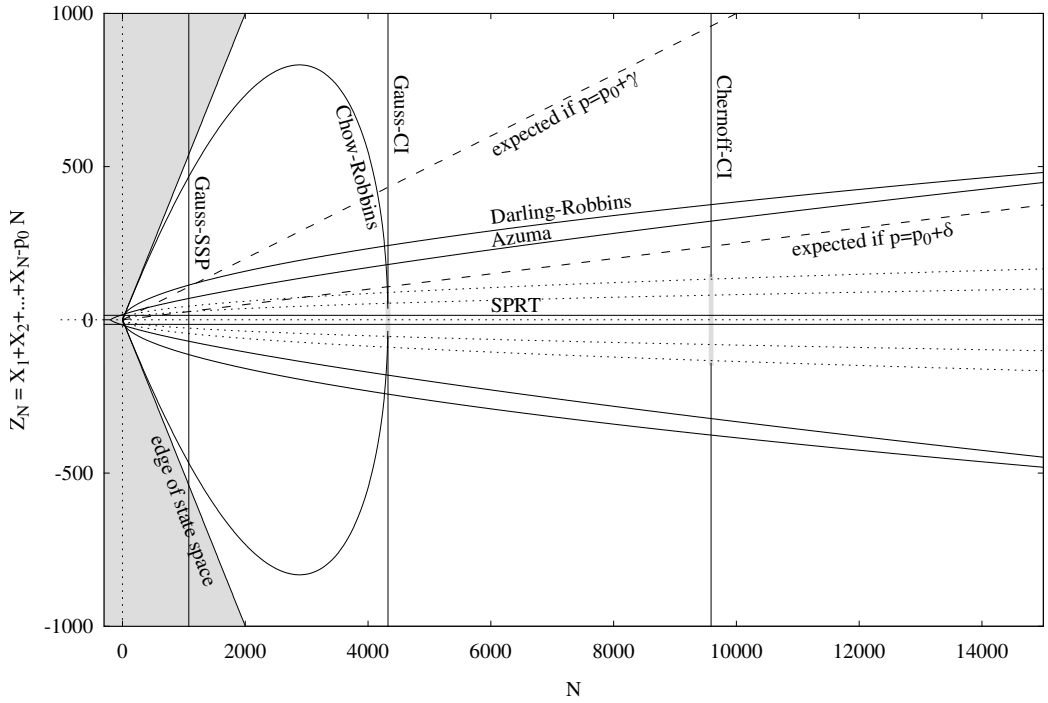

Figure 4: Critical regions, $p_0 = 0.5$, $\delta = \zeta = 0.025$, $\gamma = 0.1$, $\alpha = \beta = 0.05$. Solid lines indicate boundaries of the critical regions; grey lines indicate where the test is inconclusive. Dotted lines indicate thresholds of Gauss-CI and Chernoff-CI for different β. Dashed lines indicates expected sample path for $p = p_0 + \delta$, i.e., edge of indifference region, and for $p = p_0 + \gamma$, i.e., for which the Azuma and Darling tests have been optimally parametrised.

Finally, consider the four (almost) fixed-N tests: Chernoff-CI, Gauss-CI, Chow-Robbins and Gauss-SSP. As was pointed out in Section 4.2, the Chernoff-CI test is based on a looser bound than the others and therefore takes more samples for the same confidence level. The Gauss-CI and Chow-Robbins use the same bound and only differ in how they determine at what N to terminate. For the Gauss-CI test, this N is determined in advance, based on obtaining a sufficiently narrow confidence interval under the null hypothesis ($p = p_0$). On the other hand, Chow-Robbins stops as soon as the confidence interval is narrow enough based on the actual samples. If p is close to 0 or 1, this may occur much sooner. The Gauss-SSP test is similar to the Gauss-CI test in that its stopping time is determined in advance. However, it stops earlier because, like the SPRT, it takes the risk of drawing the wrong conclusion with probability more than α if $|p - p_0| < \delta$, while the Gauss-CI test in that case mostly risks terminating inconclusively.

As an aid in understanding, a dashed line shows the expected sample path if $p = p_0 + \delta$, i.e., at the border of where the investigator is indifferent about the outcome. This line crosses the Gauss-CI, Chernoff-CI and Chow-Robbins boundaries well away from the grey (inconclusive) parts, thus showing that these tests are indeed likely to conclude conclusively (namely 95%), and extremely unlikely to draw the wrong conclusion. The SPRT area is rather narrow; the dashed line leaves it

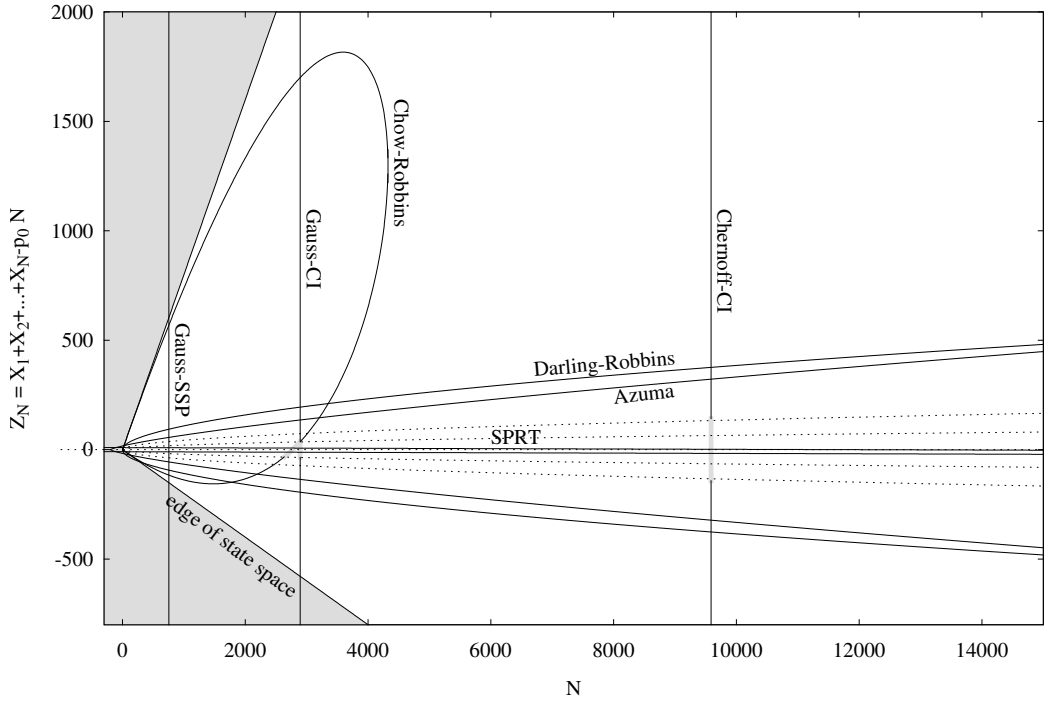

Figure 5: As figure 4, but for $p_0 = 0.2$.

soon, at a point where it is still relatively near to the lower edge of this area, illustrating the 5% risk of drawing the wrong conclusion. The Gauss-SSP test runs the same risk, due to its early termination. The Azuma and Darling boundaries are intercepted beyond the edge of the figure. These tests take rather long in this case because we chose to parametrise them optimally for a larger difference between p and p_0, namely $\gamma = 0.1$ rather than $\delta = 0.025$, which is illustrated by the other dashed line.

Figure 5 is similar to Figure 4 but with p_0 set to 0.2. We mention the main differences. First, although this is barely visible, the boundaries of the SPRT are not constant, but they decrease in N (see (15) and (16)). Second, the Gauss-CI test's area $\mathcal{NC}$ is less broad due to the smaller variance under the null hypothesis. Third, the Chow-Robbins test may now take longer than the Gauss-CI test, because the Chow-Robbins test continues until the confidence interval has reached a prescribed width, which takes longest if $p = 0.5$. The Gauss-CI test stops earlier in that case, because p is so far away from its value under the null hypothesis that a much wider confidence interval still allows for a confident decision.

Figures 4 and 5 are only two examples of the figures that can be generated interactively on the website [1] mentioned earlier.

5.2 Simulation Results

In this section, we compare the tests discussed in this paper by empirically evaluating their performance for a range of underlying parameter values.[11] Since we only compare different statistical tests, we do not need to consider the simulation aspect of statistical model checking. Accordingly, we let our computer program directly draw samples from a Bernoulli distribution with (known) parameter p. With p chosen, the remaining parameter to be chosen is δ, ζ or γ. In all cases $\alpha = \beta = 0.05$.

For each test we estimate the following metrics:

1. ρ, the probability that a test accepts the right hypothesis, used as a measure for the *confidence* (the higher the better);
2. υ, the probability that a test proves inconclusive, used as a measure for the *power* (the lower the better);
3. η, the expected number of samples drawn before the test is concluded, used as a measure for the *efficiency* (the lower the better).

The procedure is as follows: we conduct each test 1 000 times, let $\hat{\rho}$ be the fraction of correct conclusions, $\hat{\upsilon}$ the fraction of tests that remained inconclusive (where for the sequential tests, we set a 60 second time bound) and $\hat{\eta}$ be the average number of samples drawn. In Tables 6, 7 and 8, we display these estimates plus/minus the half-width of a 95%-CI around the estimate. In Tables 6 and 8 we have set $p_0 = 0.5$; the only difference between these two tables is the choice of $|p - p_0|$, which

[11] All experiments were done using our own Java-code. In [27], it was observed that in VeStA, the Gauss-SSP test seemed to have lower confidence than the SPRT of YMER. Using our own implementation of the tests, we do not observe the same.

equals 0.1 for the former and 0.001 for the latter. For Table 7 we have set $p = 0.2$ and $|p - p_0| = 0.01$. The rows in bold indicate that the input parameter δ, ζ or γ is exactly equal to $|p - p_0|$.

The number of samples needed for the *Gauss-CI* test grows inversely proportional to the square root of ζ. Because the Gauss-CI test is a fixed sample size test, $\hat{\eta}$ has no variance. The main drawback is that if ζ is considerably larger than $|p - p_0|$, the Gauss-CI test will almost never draw a conclusion. This is witnessed by $\hat{v} \gg 0$, seen particularly in Table 8. The bounds on the error probabilities are very tight; in all tables we see that if $\zeta = |p - p_0|$, the probability of drawing the correct conclusion is close to $1 - \beta = 0.95$. Furthermore, in Table 8 we observe that when ζ is chosen much too large, the proportion of incorrect conclusions (i.e., $1 - \hat{\rho} - \hat{v}$) is close to $\alpha = 0.05$.

We see in general that the *Chernoff-CI* test requires more samples than the Gauss-CI test; in Table 7, for which p_0 equals 0.2 instead of 0.5, the difference between the sample sizes of the Gauss-CI and Chernoff-CI tests is larger than in Tables 6 and 8. This is consistent with the discussion of Section 3.2. The bound on the probability of error of the second kind for the Chernoff-CI test appears to be rather loose; when the power-indifference level ζ equals the actual difference $p - p_0$, the estimate for the probability of inconclusive termination $\hat{v}$ is well below $\beta = 0.05$.

That the *Chow-Robbins* test is a mixture of a fixed sample size test and a sequential test can be seen from the low variance of the number of samples drawn. In Table 6, the variance of Z_N under p_0 is considerably higher than under p, so the Chow-Robbins test requires a noticeably smaller sample size on average than the Gauss-CI test. However, the reverse is true in Table 7 and the Chow-Robbins test does slightly worse than the Gauss-CI test as a consequence. Overall, the two tests have similar efficiency.

The *SPRT* is the most efficient among all tests when δ is picked just right; in each table, its value $\hat{\eta}$ is the lowest among all tests that satisfy correctness. However, we indeed see its performance degrade when its assumptions are violated, i.e., when $|p - p_0|$ turns out to be smaller than δ. In Table 8, the CI for $\hat{\rho}$ contains $\frac{1}{2}$ when δ is large, which is the worst level of ρ that a test can satisfy (after all, if the confidence was even lower one could always use the opposite result of the test and obtain a confidence that is $> \frac{1}{2}$). The average number of samples needed seems to grow inversely proportional to δ.

The *Gauss-SSP* test is similar to the SPRT, albeit slightly less efficient. This was to be expected, see also [39] where the same observation was made.

Both the *Azuma* and *Darling* tests are very conservative: they have a $\hat{\rho}$ of well over 95%. When the guess is (almost) correct, the Azuma test is more efficient than the Darling test. However, if γ is taken to be considerably larger than $|p - p_0|$, the number of samples needed for the Azuma test grows rapidly, while the Darling test remains remarkably insensitive to the model parameters, as can been seen in all tables. The Azuma result $\hat{v} \approx 1$ in Table 8 means that the Azuma test did not draw a conclusion within a 60 second time period.

5.3 Tool Implementation

Table 9 contains a summary of the implementation of the tests of Section 3 in model checking tools. In this section, we discuss each of the tools in some detail.

UPPAAL allows the user to check qualitative as well as quantitative statements (as described in the introduction). Qualitative statements are evaluated using the SPRT. Quantitative statements were evaluated using a sample size determined using the Chernoff-Hoeffding bound; since version 4.1.15, the Chow-Robbins procedure is used to construct a Clopper-Pearson confidence interval. **PRISM** (version 4.1.beta2) implements all four methods in the context of making qualitative statements; Gauss-CI and Chow-Robbins are implemented as versions of the '*ACI*' method, Chernoff as the '*APMC*' method and the SPRT as the '*SPRT*' method. PRISM does not allow the user to directly create confidence intervals for the sake of making quantitative statements; however confidence intervals are created as a by-product of hypothesis tests and can be found in the '*log*' section. **MRMC** (v1.5) [24] only implements the Chow-Robbins test, but, unlike PRISM, also allows this method to be used to evaluate steady-state properties (which we do not discuss in this paper).

COSMOS (v1.0) [7] implements the Chow-Robbins test for quantitative purposes. **PLASMA** (version 1.2.8) [22] implements the SPRT for qualitative statements and the Chernoff test for quantitative statements.

YMER uses the SPRT. Different version of YMER feature different add-ons; e.g., version 3.0.9 includes a numerical solution engine that allows the user to check nested operators, while version 4.0 includes support for unbounded until. The tool **PVeStA**, which is based on the tool VeStA [36], implements the Gauss-SSP test and the Chow-Robbins method. Another variant of VeStA, **MultiVeStA** [33], implements the Chow-Robbins procedure for quantitative purposes. **APMC** (v3.0) [19] implements an SSP test based on the Chernoff bound, cf. end of Section 3.5.

6 Conclusions

We have presented a common framework that allows the hypothesis testing methods—both '*pure*' hypothesis tests and those based on confidence intervals—proposed earlier in the statistical model checking literature to be compared in a mathematically solid, yet intuitive manner. Previously, these methods were often implemented

Test	δ, ζ or γ	$\hat{\rho}$	$\hat{v}$	$\hat{\eta}$
Gauss-CI	**0.1**	**0.953 ± 0.013**	**0.047 ± 0.013**	$\mathbf{2.58 \cdot 10^2}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$2.63 \cdot 10^4$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$2.63 \cdot 10^6$
Chernoff-CI	**0.1**	**0.993 ± 0.005**	**0.007 ± 0.005**	$\mathbf{6.00 \cdot 10^2}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$5.99 \cdot 10^4$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$5.99 \cdot 10^6$
Chow-Robbins	**0.1**	**0.948 ± 0.014**	**0.052 ± 0.014**	$\mathbf{(2.473 \pm 0.004) \cdot 10^2}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$(2.5215 \pm 0.0004) \cdot 10^4$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(2.52178 \pm 0.00004) \cdot 10^6$
SPRT	**0.1**	**0.95 ± 0.014**	**0.0 ± 0.0**	$\mathbf{(3.68 \pm 0.16) \cdot 10^1}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$(3.71 \pm 0.06) \cdot 10^2$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(3.67 \pm 0.02) \cdot 10^3$
Gauss-SSP	**0.1**	**0.943 ± 0.014**	**0.0 ± 0.0**	$\mathbf{6.40 \cdot 10^1}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$6.76 \cdot 10^3$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$6.76 \cdot 10^5$
Azuma	**0.1**	**1.0 ± 0.0**	**0.0 ± 0.0**	$\mathbf{(3.80 \pm 0.01) \cdot 10^2}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$(1.84 \pm 0.01) \cdot 10^3$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(1.72 \pm 0.00) \cdot 10^4$
Darling-Robbins	**0.1**	**1.0 ± 0.0**	**0.0 ± 0.0**	$\mathbf{(1.17 \pm 0.02) \cdot 10^3}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$(1.49 \pm 0.02) \cdot 10^3$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(1.99 \pm 0.03) \cdot 10^3$

Table 6: $p_0 = 0.5$, $p = 0.6$

Test	δ, ζ or γ	$\hat{\rho}$	$\hat{v}$	$\hat{\eta}$
Gauss-CI	0.1	0.111 ± 0.019	0.869 ± 0.021	$1.89 \cdot 10^2$
	0.01	**0.944 ± 0.014**	**0.056 ± 0.014**	$\mathbf{1.71 \cdot 10^4}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$1.68 \cdot 10^6$
Chernoff-CI	0.1	0.012 ± 0.007	0.988 ± 0.007	$6.00 \cdot 10^2$
	0.01	**0.999 ± 0.002**	**0.001 ± 0.002**	$\mathbf{5.99 \cdot 10^4}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$5.99 \cdot 10^6$
Chow-Robbins	0.1	0.081 ± 0.017	0.868 ± 0.021	$(1.91 \pm 0.01) \cdot 10^2$
	0.01	**0.945 ± 0.014**	**0.055 ± 0.014**	$\mathbf{(1.77 \pm 0.00) \cdot 10^4}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(1.75 \pm 0.00) \cdot 10^6$
SPRT	0.1	0.658 ± 0.030	0.0 ± 0.0	$(3.44 \pm 0.17) \cdot 10^1$
	0.01	**0.946 ± 0.014**	**0.0 ± 0.0**	$\mathbf{(2.14 \pm 0.09) \cdot 10^3}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(2.37 \pm 0.04) \cdot 10^4$
Gauss-SSP	0.1	0.577 ± 0.031	0.0 ± 0.0	$5.7 \cdot 10^1$
	0.01	**0.951 ± 0.013**	**0.0 ± 0.0**	$\mathbf{4.50 \cdot 10^3}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$4.34 \cdot 10^5$
Azuma	0.1	1.0 ± 0.0	0.0 ± 0.0	$(1.13 \pm 0.01) \cdot 10^6$
	0.01	**1.0 ± 0.0**	**0.0 ± 0.0**	$\mathbf{(3.86 \pm 0.10) \cdot 10^4}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(1.82 \pm 0.01) \cdot 10^5$
Darling-Robbins	0.1	1.0 ± 0.0	0.0 ± 0.0	$(1.94 \pm 0.02) \cdot 10^5$
	0.01	**1.0 ± 0.0**	**0.0 ± 0.0**	$\mathbf{(1.70 \pm 0.02) \cdot 10^5}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$(2.02 \pm 0.02) \cdot 10^5$

Table 7: $p_0 = 0.2$, $p = 0.21$

in tools completely parallel to one another with little information given about the subtle differences between the methods and their parameters. Our contribution aids general understanding of these methods, reducing the likelihood of incorrect interpretation of the outcomes.

In order for the methods to be meaningfully compared to each other, they have to parametrised. Tools typically ask the user to specify values for parameters that are specific to a method (such as the number of samples), without a clear indication of the consequences for the outcomes. We have expressed all method-specific parameters in terms of quantities that are meaningful to the user, such as the confidence level, the risk of inconclusive termination, and indifference levels.

Having parametrised the methods consistently, we compared them graphically and numerically, highlight-

Test	δ, ζ or γ	$\hat{\rho}$	$\hat{v}$	$\hat{\eta}$
Gauss-CI	0.1	0.055 ± 0.014	0.898 ± 0.019	$2.59{\cdot}10^{2}$
	0.01	0.106 ± 0.019	0.869 ± 0.021	$2.63{\cdot}10^{4}$
	0.001	**0.95 ± 0.014**	**0.05 ± 0.0134**	$\mathbf{2.63{\cdot}10^{6}}$
Chernoff-CI	0.1	0.006 ± 0.005	0.985 ± 0.008	$6.00{\cdot}10^{2}$
	0.01	0.026 ± 0.010	0.973 ± 0.010	$5.99{\cdot}10^{4}$
	0.001	**0.994 ± 0.005**	**0.006 ± 0.005**	$\mathbf{5.99{\cdot}10^{6}}$
Chow-Robbins	0.1	0.043 ± 0.013	0.919 ± 0.017	$(2.581 \pm 0.001){\cdot}10^{2}$
	0.01	0.102 ± 0.019	0.874 ± 0.021	$(2.63 \pm 0.00){\cdot}10^{4}$
	0.001	**0.934 ± 0.015**	**0.066 ± 0.015**	$\mathbf{(2.63 \pm 0.00){\cdot}10^{6}}$
SPRT	0.1	0.482 ± 0.031	0.0 ± 0.0	$(6.80 \pm 0.34){\cdot}10^{1}$
	0.01	0.541 ± 0.031	0.0 ± 0.0	$(5.42 \pm 0.29){\cdot}10^{3}$
	0.001	**0.938 ± 0.015**	**0.0 ± 0.0**	$\mathbf{(3.12 \pm 0.14){\cdot}10^{5}}$
Gauss-SSP	0.1	0.483 ± 0.031	0.0 ± 0.0	$6.40{\cdot}10^{1}$
	0.01	0.59 ± 0.030	0.0 ± 0.0	$6.76{\cdot}10^{3}$
	0.001	**0.962 ± 0.012**	**0.0 ± 0.0**	$\mathbf{6.76{\cdot}10^{5}}$
Azuma	0.1	0.0 ± 0.0	1.0 ± 0.0	$(2.40 \pm 0.02){\cdot}10^{8}$
	0.01	0.0 ± 0.0	1.0 ± 0.0	$(2.39 \pm 0.01){\cdot}10^{8}$
	0.001	**1.0 ± 0.0**	**0.0 ± 0.0**	$\mathbf{(3.78 \pm 0.1){\cdot}10^{6}}$
Darling-Robbins	0.1	1.0 ± 0.0	0.0 ± 0.0	$(2.70 \pm 0.04){\cdot}10^{7}$
	0.01	1.0 ± 0.0	0.0 ± 0.0	$(2.30 \pm 0.03){\cdot}10^{7}$
	0.001	**1.0 ± 0.0**	**0.0 ± 0.0**	$\mathbf{(2.18 \pm 0.03){\cdot}10^{7}}$

Table 8: $p_0 = 0.5$, $p = 0.501$

	UPPAAL (before 4.1.15)	UPPAAL (4.1.15 onward)	PRISM	MRMC	COSMOS	PLASMA	YMER	PVeStA	APMC
Gauss-CI			✓✓						
Chernoff-CI	✓		✓✓			✓			
Chow-Robbins		✓	✓✓	✓✓	✓			✓	
SPRT	✓✓	✓✓	✓✓			✓✓	✓✓		
Gauss-SSP								✓✓	
Chernoff-SSP									✓✓

Table 9: Tool implementation. A ✓✓ means that the procedure is (also) implemented as a hypothesis test, a ✓ means that the procedure is only implemented for making quantitative statements.

ing each method's properties, and demonstrating quantitative performance differences.

Besides all methods known to us and implemented in tools, our comparison has also included two hypothesis testing methods that have not been discussed in the SMC context before. Those two methods (called Azuma and Darling-Robbins in this paper) are sequential methods. They behave fundamentally different from the other methods in cases where the model probability being studied is very close to the threshold under consideration: these methods will neither terminate inconclusively, nor have their confidence level drop.

There is no single best method to be recommended, since this depends on the requirements of the user. The present paper gives an overview both of the methods' characteristics and their performance, summarised in Table 5, and thus can help tool users and authors in making a well-informed choice.

Acknowledgements

This work is partially supported by the Netherlands Organisation for Scientific Research (NWO), project number 612.064.812, and by the EU project QUANTICOL, 600708.

References

1. Companion website to this paper. http://wwwhome.ewi.utwente.nl/∼ptdeboer/hyptest-for-smc/.
2. A. Aziz, K. Sanwal, V. Singhal, and R. Brayton. Model-checking continuous-time Markov chains. *ACM Transactions on Computational Logic (TOCL)*, 1(1):162–170, 2000.
3. A. Aziz, K. Sanwal, V. Singhal, and R.K. Brayton. Verifying continuous-time Markov chains. *Lecture Notes in Computer Science*, 1102:269–276, 1996.
4. C. Baier, B. R. Haverkort, H. Hermanns, and J. P. Katoen. On the logical characterisation of performability

properties. In *Automata, Languages and Programming*, pages 780–792. LNCS Volume 1853, Springer, 2000.
5. C. Baier, B. R. Haverkort, H. Hermanns, and J. P. Katoen. Model-checking algorithms for continuous-time Markov chains. *IEEE Transactions on Software Engineering*, 29(6):524–541, 2003.
6. C. Baier and J. P. Katoen. *Principles of model checking*. MIT press, 2008.
7. P. Ballarini, H. Djafri, M. Duflot, S. Haddad, and N. Pekergin. COSMOS: a statistical model checker for the hybrid automata stochastic logic. In *Proceedings of the Eighth International Conference on the Quantitative Evaluation of Systems (QEST)*, pages 143–144. IEEE, 2011.
8. J. Bengtsson, K. Larsen, F. Larsson, P. Pettersson, and W. Yi. UPPAAL — a tool suite for automatic verification of real-time systems. *Hybrid Systems III*, pages 232–243, 1996.
9. E. Chong and S. Żak. *An Introduction to Optimization*. John Wiley & Sons, 2004.
10. Y. S. Chow and H. Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, 36(2):457–462, 1965.
11. D.A. Darling and H. Robbins. Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences of the United States of America*, 61(3):804–809, 1968.
12. D. El Rabih and N. Pekergin. Statistical model checking using perfect simulation. In *Automated Technology for Verification and Analysis*, pages 120–134. LNCS Volume 5799, Springer, 2009.
13. G.S. Fishman. *Discrete-event simulation: modeling, programming, and analysis*. Springer, 2001.
14. P.W. Glynn. A GSMP formalism for discrete event systems. *Proceedings of the IEEE*, 77(1):14–23, 1989.
15. F.E. Grubbs. On designing single sampling inspection plans. *The Annals of Mathematical Statistics*, pages 242–256, 1949.
16. P.J. Haas and G.S. Shedler. Stochastic Petri net representation of discrete event simulations. *IEEE Transactions on Software Engineering*, 15(4):381–393, 1989.
17. H. Hansson and B. Jonsson. A logic for reasoning about time and reliability. *Formal aspects of computing*, 6(5):512–535, 1994.
18. T. Hérault, R. Lassaigne, F. Magniette, and S. Peyronnet. Approximate probabilistic model checking. *Lecture notes in computer science*, 2937:307–329, 2004.
19. T. Hérault, R. Lassaigne, and S. Peyronnet. APMC 3.0: Approximate verification of discrete and continuous time markov chains. In *Proceedings of the Third International Conference on the Quantitative Evaluation of Systems (QEST)*, pages 129–130. IEEE, 2006.
20. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.
21. H. Jeffreys. *Theory of Probability*. Oxford University Press, 1961.
22. C. Jegourel, A. Legay, and S. Sedwards. A platform for high performance statistical model checking – PLASMA. *Tools and Algorithms for the Construction and Analysis of Systems*, pages 498–503, 2012.
23. S. Jha, E. Clarke, C. Langmead, A. Legay, A. Platzer, and P. Zuliani. A Bayesian approach to model checking biological systems. In *Computational Methods in Systems Biology*, pages 218–234. Springer, 2009.
24. J. P. Katoen, M. Khattri, and I.S. Zapreev. A Markov reward model checker. In *Second International Conference on the Quantitative Evaluation of Systems (QEST)*, pages 243–244. IEEE, 2005.
25. M. Kwiatkowska, G. Norman, and D. Parker. PRISM: Probabilistic symbolic model checker. In *Computer Performance Evaluation: Modelling Techniques and Tools*, pages 113–140. LNCS Volume 2324, Springer, 2002.
26. T.L. Lai. Nearly optimal sequential tests of composite hypotheses. *The Annals of Statistics*, pages 856–886, 1988.
27. A. Legay, B. Delahaye, and S. Bensalem. Statistical model checking: an overview. In *Runtime Verification*, pages 122–135. Springer, 2010.
28. K. Matthes. Zur Theorie der Bedienungsprozesse. In *Proceedings of the Third Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 513–528. Publishing House of the Czechoslovak Academy of Sciences, 1962.
29. D. Reijsbergen. *Efficient simulation techniques for stochastic model checking*. PhD thesis, University of Twente, Enschede, December 2013.
30. D. Reijsbergen, P.T. de Boer, and W. Scheinhardt. A sequential hypothesis test based on a generalized azuma inequality. Forthcoming.
31. B.D. Ripley. *Stochastic Simulation*. John Wiley & Sons, 1987.
32. S.M. Ross. *Stochastic Processes*. John Wiley & Sons, 1996.
33. S. Sebastio and A. Vandin. MultiVeStA: Statistical model checking for discrete event simulators. In *Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, 2013.
34. K. Sen, M. Viswanathan, and G. Agha. Statistical model checking of black-box probabilistic systems. In *Computer Aided Verification*, pages 202–215. LNCS Volume 3114, Springer, 2004.
35. K. Sen, M. Viswanathan, and G. Agha. On statistical model checking of stochastic systems. In *Computer Aided Verification*, pages 266–280. LNCS Volume 3576, Springer, 2005.
36. K. Sen, M. Viswanathan, and G. Agha. VeStA: A statistical model-checker and analyzer for probabilistic systems. In *Proceedings of the Second International Conference on the Quantitative Evaluation of Systems (QEST)*, pages 251–252. IEEE, 2005.
37. A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.
38. H.L.S. Younes. *Verification and planning for stochastic processes with asynchronous events*. PhD thesis, Carnegie Mellon, 2005.
39. H.L.S. Younes. Error control for probabilistic model checking. In *Verification, Model Checking, and Abstract Interpretation*, pages 142–156. Springer, 2006.
40. H.L.S. Younes, E. Clarke, and P. Zuliani. Statistical verification of probabilistic properties with unbounded until. *Formal Methods: Foundations and Applications*, pages 144–160, 2011.

41. H.L.S. Younes, M. Kwiatkowska, G. Norman, and D. Parker. Numerical vs. statistical probabilistic model checking. *International Journal on Software Tools for Technology Transfer (STTT)*, 8(3):216–228, 2006.
42. H.L.S. Younes and R.G. Simmons. Probabilistic verification of discrete event systems using acceptance sampling. In *Computer Aided Verification*, pages 223–235. LNCS Volume 2404, Springer, 2002.
43. H.L.S. Younes and R.G. Simmons. Statistical probabilistic model checking with a focus on time-bounded properties. *Information and Computation*, 204(9):1368–1409, 2006.