

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Hypothesis testing for rare-event simulation: limitations and possibilities

Citation for published version:
Reijsbergen, D, de Boer, P-T & Scheinhardt, W 2016, Hypothesis testing for rare-event simulation: limitations and possibilities. in *7th International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (ISoLA 2016).* Lecture Notes in Computer Science , vol. 9952, Springer, Cham, Corfu, Greece, pp. 16-26, ISoLA 2016, Corfu, Greece, 5/10/16. https://doi.org/10.1007/978-3-319-47166-2_2

Digital Object Identifier (DOI):
10.1007/978-3-319-47166-2_2

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
7th International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (ISoLA 2016)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 29. Mar. 2024

Hypothesis testing for rare-event simulation: limitations and possibilities

Preprint version: see the proceedings of ISoLA 2016 for the full version.

Daniël Reijsbergen[1], Pieter-Tjerk de Boer[2], Werner Scheinhardt[2]

[1]University of Edinburgh, Scotland, UK
[2]University of Twente, Enschede, The Netherlands

Abstract. One of the main applications of probabilistic model checking is to decide whether the probability of a property of interest is above or below a threshold. Using statistical model checking (SMC), this is done using a combination of stochastic simulation and statistical hypothesis testing. When the probability of interest is very small, one may need to resort to rare-event simulation techniques, in particular importance sampling (IS). However, IS simulation does not yield 0/1-outcomes, as assumed by the hypothesis tests commonly used in SMC, but likelihood ratios that are typically close to zero, but which may also take large values.
In this paper we consider two possible ways of combining IS and SMC. One involves a classical IS-scheme from the rare-event simulation literature that yields likelihood ratios with bounded support when applied to a certain (nontrivial) class of models. The other involves a particular hypothesis testing scheme that does not require a-priori knowledge about the samples, only that their variance is estimated well.

1 Introduction

One of the main applications of statistical model checking (SMC) [13, 19] is the use of computer simulation and hypothesis testing to determine whether some probability p in a model is larger or smaller than a given probability threshold p_0. Thus, several suitable hypothesis tests have been developed by various authors to test the hypothesis $p > p_0$ against $p < p_0$, and they are implemented in different tools; for an overview see [16]. These can be combined easily with simulation experiments in which each sample yields a Bernoulli random variable (representing whether the event of interest was observed or not).

In the rare-event context, where the probability p is extremely small, standard simulation is not efficient since observing the target event would require an excessively large number of samples. Techniques to estimate such small probabilities include importance sampling and splitting/restart, both going back to the early days of computing [11]. Recently there has been much interest in applying such techniques in the statistical model checking context, as witnessed by various PhD theses [3, 10, 14] and associated publications. These techniques

have in common[1] that the simulation samples no longer yield Bernoulli random variables, i.e., the outcomes are no longer restricted to $\{0, 1\}$; in fact their distribution may be highly asymmetric. It is a challenging goal to combine rare-event simulation techniques with hypothesis testing schemes in such a way that sound statistical conclusions can be obtained within reasonable simulation time.

In this short paper, we explore the options of extending hypothesis tests to importance sampling. In Section 2, we investigate existing tests and the assumptions they make on the samples. Based on that, we consider two options: upper-bounding the likelihood ratio in Section 3, and using the normal approximation in Section 4. We provide a numerical illustration in Section 5, and brief conclusions in Section 6.

2 Generalizability of existing hypothesis tests

Almost all existing hypothesis tests for statistical model checking fit in a relatively simple framework, cf. [16]. Independent samples $X_i \in \{0, 1\}$ are generated for $i = 1, 2, \ldots$. The test statistic after N samples is $Z_N = \sum_{i=1}^{N} X_i - Np_0$. The test draws a conclusion when (N, Z_N) leaves the so-called critical area; then $Z_N > 0$ is evidence for the hypothesis H_1, which asserts that $p > p_0$. Conversely, $Z_N < 0$ is evidence for the hypothesis H_{-1}, which asserts that $p < p_0$. The shape of the boundaries of the critical area varies from test to test, and is chosen such that confidence levels are upheld; i.e., the probability of errors of the first[2] kind (accepting the wrong hypothesis) and the second kind (finishing undecided, as some tests can do) are upper-bounded by, e.g. 5% for a 95% confidence level. Some tests decide after a fixed number N samples have been drawn (fixed sample size tests, often related to confidence interval calculation), whereas others are sequential, meaning that after every new sample the test decides whether a conclusion can be drawn or more samples are needed. Another difference between tests is how they behave if p is closer to p_0 than some indifference level: class-I tests no longer live up to their confidence guarantee, class-II will tend to terminate undecided, while class-III will insist on drawing more and more samples until a confident conclusion can be drawn. For an overview of tests, and much more detail about their properties, see [16, 20].

So far it was assumed that X_i is an indicator: in each simulation replication the event of interest either does or does not occur, and we are interested in its probability. In case the target event is rare, $X_i = 0$ for all or almost all samples,

[1] For importance sampling this is obvious. For splitting, one common implementation (e.g., [7]) produces each independent sample as the sum of the weights of all target-reaching offspring of an initial particle, so clearly these samples are no longer Bernoulli. Other variants exist which do this differently, with their own complications for hypothesis testing. However, this is outside the scope of this short paper.

[2] Note that "first" and "second" kind are a bit different here than in most hypothesis testing literature, since we have *two* hypotheses to be tested ($p > p_0$ and $p < p_0$), besides the null hypothesis ($p = p_0$). See Section 2.2 in [16].

leading to an unusable estimator. One popular solution for this is importance sampling (see e.g. [11]), where the probability distributions in the model are modified to make the target event more likely, while keeping track of a so-called likelihood ratio by which the events need to be weighed. Effectively then, X_i takes on either the value 0 or the likelihood ratio value, so it is no longer restricted to $\{0, 1\}$. The X_i will typically be very small (representing the rarity of the event), but may take any non-negative real number. Their mean is still p, the probability of interest; however, their variance, which was $p(1-p)$ in the Bernoulli case, may be totally different. Thus, we need to reconsider whether the hypothesis tests are still valid when the X_i are no longer Bernoulli. Table 1 provides an overview. (A similar comparison, but for confidence intervals rather than hypothesis tests, is found in Chapter 2 of [3].)

The table's third column lists conditions on the samples X_i that are used in the derivation or correctness proof of the respective test. As we see, about half of the tests explicitly assume that X_i is an indicator function, which is no longer the case when importance sampling is used. For some of these tests (SPRT and Darling-Robbins), it is crucial that there are indeed only two possible outcomes; for some others, the proof can be generalized to any X_i as long as they are bounded. This is the first option we will explore. In [4], one approach for bounding likelihood ratios was discussed, but it required rather complicated and model-dependent proofs. In Section 3, we show bounded likelihood ratio for a more general class of models and a well-known importance sampling scheme. In fact, our method and model class is similar to [2], where also an upper bound for the likelihood ratios is guaranteed; however, their upper bound is 1, which would lead to very conservative (and thus inefficient) hypothesis tests.

The other tests have normality listed in the third column. This means that these tests rely on the Central Limit Theorem: for sufficiently large N, Z_N becomes approximately normally distributed, so the normal distribution can be used to set decision thresholds such that confidence levels are upheld. This holds for any distribution of the X_i with finite variance, but suffers from the problem of needing to know when N is large enough. The lack of restrictions on X_i makes these tests attractive for use with importance sampling, and we explore this in Section 4, where we will find that only Chow-Robbins can be used.

3 Bounded likelihood ratios in multicomponent systems

As discussed in the previous section, several hypothesis tests can still be applied if the likelihood ratios returned by the IS scheme can be bounded from above. Although it is difficult to construct such bounds in general, there are restricted modelling classes in which this is more straightforward. In this section we discuss the modelling class of multicomponent systems, with a particular focus on

[3] The actual bound on which this is based, is due to Hoeffding [9], but since literature and tools frequently refers to this as Chernoff's, we choose to mention both names here.

the Distributed Database System (DDS). We consider the probability that, after the first component has failed, such a system reaches a system failure state before all components are repaired. This probability is interesting because it appears in expressions for other performance measures such as the system unreliability, unavailability and mean time to failure — in those expressions it is the only quantity that is difficult to estimate. We focus on the specific IS scheme of Balanced Failure Biasing (BFB), a classic IS scheme [18] for highly reliable Markovian systems, although the result of equation (1) holds in more general cases.

The general set-up of a multicomponent system is as follows. The system consists of d component types; let $\mathcal{D} = \{1, \ldots, d\}$. Let $\boldsymbol{x}$ be the state of the Markov chain, where the i-th entry x_i (with $i \in \mathcal{D}$) is the number of failed components of type i. Here, x_i takes values in $0, ..., n_i$, where n_i is the number of components of type i needed to trigger system failure. The initial state is given by $\boldsymbol{x}_0$, which is a d-dimensional zero-valued vector. The failure rate of components of type i is denoted by $\lambda_i(\boldsymbol{x})$, and their repair rate is $\mu_i(\boldsymbol{x})$ $\forall i \in \mathcal{D}, \boldsymbol{x} \in \mathbb{N}^d$. Note that these rates are state-dependent — e.g., in the DDS example, the failure rate of components of type i depends on how many components of type i are

Test	**class**	**conditions on X_i**	**generalisable to non-Bernoulli?**
SPRT	I	$X_i \in \{0, 1\}$	no: assumes hypotheses describe entire outcome distribution.
Gauss-SSP	I	sum of many X_i is approximately normally distributed	no: sample variance under $p = p_0 \pm \delta$ is needed
Gauss-CI	II	sum of many X_i is approximately normally distributed	no: sample variance under $p = p_0$ is needed
Chow-Robbins	II	sum of many X_i is approximately normally distributed	yes: only variance under the true p is needed, which can be estimated during the simulation
Chernoff-Hoeffding[3]-CI	II	$X_i \in \{0, 1\}$	yes: to any bounded X_i.
Azuma	III	$X_i \in \{0, 1\}$	yes: to any bounded X_i.
Darling-Robbins	III	$X_i \in \{0, 1\}$	no: D-R theorem is about entire distributions, not expectations.

Table 1. Overview of existing hypothesis tests for SMC (from [16]) and their requirements w.r.t. the samples X_i

still operational. The *exit rate* of a state $\boldsymbol{x} \in \mathbb{N}^d$ is given by

$$\eta(\boldsymbol{x}) = \sum_{j \in \mathcal{D}} \left(\lambda_j(\boldsymbol{x}) + \mu_j(\boldsymbol{x})\right).$$

Let e_i, $i \in \mathcal{D}$, be a vector of length d filled with $d-1$ zeros and a 1 at position i, and $\boldsymbol{x}_0$ the initial state. The probability of a 'straight' path (see [15]) leading to failure of component type $i \in \mathcal{D}$:

$$\prod_{j=0}^{n_i-1} \frac{\lambda_i(\boldsymbol{x}_0 + je_i)}{\eta(\boldsymbol{x}_0 + je_i)}.$$

Using IS, we simulate under different failure rates $\lambda_i^*(\boldsymbol{x})$ and repair rates $\mu_i^*(\boldsymbol{x})$. Assume (without loss of generality) that the rates are normalized such that the exit rates are the same under the new measure. Then the likelihood ratio of a straight path leading to failure of component type $i \in \mathcal{D}$ is given by

$$\prod_{j=0}^{n_i-1} \frac{\lambda_i(\boldsymbol{x}_0 + je_i)}{\lambda_i^*(\boldsymbol{x}_0 + je_i)}.$$

We define $L^{\max}$ as the largest of these likelihood ratios:

$$L^{\max} = \max_{i \in \mathcal{D}} \prod_{j=0}^{n_i-1} \frac{\lambda_i(\boldsymbol{x}_0 + je_i)}{\lambda_i^*(\boldsymbol{x}_0 + je_i)}.$$

To avoid the rare-event problem, $\lambda_i^*(\boldsymbol{x}_0 + je_i) > \lambda_i(\boldsymbol{x}_0 + je_i)$, so $L^{\max}$ is typically smaller than 1. However, since the exit rates are the same, it must hold that $\mu_i^*(\boldsymbol{x}_0 + je_i) < \mu_i(\boldsymbol{x}_0 + je_i)$, so if a μ-transition takes place the likelihood ratio increases. However, for every μ-transition there must be an accompanying λ-transition that took place earlier, since we started in the state where all components were operational. Let $\iota \in \mathcal{D}$ be the component type in which there's a failure — for every time the μ_ι-transition there has to be a λ_ι-transition to compensate, or else the system cannot end up in a failure state. Also, for the component types i for which the system doesn't fail, the μ_i-transition can only be fired if a spurious (i.e., not contributing to the rare event) λ_i-transition has been fired.

This leads to the following proposition, which is trivial to prove using the above line of reasoning. Let $\boldsymbol{X}'$ be the set of states reachable from the initial state $\boldsymbol{x}_0$. If

$$\frac{\max_{\boldsymbol{x} \in \boldsymbol{X}'} \lambda_i(\boldsymbol{x})}{\min_{\boldsymbol{x} \in \boldsymbol{X}'} \lambda_i^*(\boldsymbol{x})} \frac{\max_{\boldsymbol{x} \in \boldsymbol{X}'} \mu_i(\boldsymbol{x})}{\min_{\boldsymbol{x} \in \boldsymbol{X}'} \mu_i^*(\boldsymbol{x})} \leq 1 \tag{1}$$

then the values of the likelihood ratios are bounded from above by $L^{\max}$.

We will now consider what this means specifically for the application of BFB to the DDS. First, BFB is defined $\forall i \in \mathcal{D}$ as

$$\frac{\lambda_i^*(\boldsymbol{x})}{\eta(\boldsymbol{x})} = \begin{cases} 1/n_f(\boldsymbol{x}) & \text{if } n_r(\boldsymbol{x}) = 0, \\ 0 & \text{if } n_f(\boldsymbol{x}) = 0, \\ (2n_f(\boldsymbol{x}))^{-1} & \text{if failure and } n_r(\boldsymbol{x}) > 0, \end{cases}$$

for failure transitions and

$$\frac{\mu_i^*(\boldsymbol{x})}{\eta(\boldsymbol{x})} = \begin{cases} 0 & \text{if } n_r(\boldsymbol{x}) = 0, \\ 1/n_r(\boldsymbol{x}) & \text{if } n_f(\boldsymbol{x}) = 0, \\ (2n_r(\boldsymbol{x}))^{-1} & \text{if repair and } n_f(\boldsymbol{x}) > 0. \end{cases}$$

for repair transitions. Here, $n_f(\boldsymbol{x})$ is the amount of failures enabled in state $\boldsymbol{x}$, $n_r(\boldsymbol{x})$ is the amount of repairs and $p \in (0,1)$, where $p = \frac{1}{2}$ is a typical choice.

The benchmark parameters of the DDS, as used for example in [17], are as follows. There are $d = 9$ component types — one set of processors, two sets of disk controllers and 6 sets of disks. We have $n_i = 2$ for all $i \in \mathcal{D}$. Let $\lambda = 1/6000$, and x_i the number of operational components of type i. Then the failure rate for type i is $3(2 - x_i)\lambda$ if type i consists of processors or disk controllers, and $(4 - x_i)\lambda$ if type i consists of disks. The repair rate μ_i is 1 if $x_i > 0$. We are interested in the rare event that after the first component has failed, we reach the system failure state before returning to $\boldsymbol{x}_0$.

The quantities involved in (1) are as follows:

- $\max_{\boldsymbol{x} \in \boldsymbol{X}'} \lambda_i(\boldsymbol{x}) = 6 \cdot \frac{1}{6000}$, namely the failure rates of processors and disk controllers if they all are operational;
- $\max_{\boldsymbol{x} \in \boldsymbol{X}'} \mu_i(\boldsymbol{x}) = 1$, in fact $\mu_i(\boldsymbol{x})$ for all component types and all $\boldsymbol{x} \in \boldsymbol{X}'$;
- $\min_{\boldsymbol{x} \in \boldsymbol{X}'} \lambda_i^*(\boldsymbol{x}) \geq \frac{1}{2d} \min_{\boldsymbol{x} \in \boldsymbol{X}'} \eta_i(\boldsymbol{x}) \geq \frac{1}{2d}$, because in all states in $\boldsymbol{X}'$ at least one repair is enabled, meaning that the exit rate must be at least 1;
- $\min_{\boldsymbol{x} \in \boldsymbol{X}'} \mu_i^*(\boldsymbol{x}) \geq \frac{1}{2d} \min_{\boldsymbol{x} \in \boldsymbol{X}'} \eta_i(\boldsymbol{x}) \geq \frac{1}{2d}$ for similar reasons.

Hence, the expression on the left in (1) evaluates to $4d^2/1000 = 0.364 < 1$, so BFB has bounded likelihood ratios in the DDS. Note that this is the maximum contribution of *a single cycle*, not the likelihood ratio on a complete path.

As a side note: if $\boldsymbol{x}_0$ were a valid state, then $\min_{\boldsymbol{x} \in \boldsymbol{X}'} \lambda_i^*(\boldsymbol{x})$ would be very small as $\eta(\boldsymbol{x}_0)$ is very small. However, since we are interested in reaching failure before full repair, we have no so-called high-probability cycles [8].

Regarding $L^{\max}$, this is achieved on the 'straight' paths involving failure of processors or disk controllers. In particular, straightforward computations show that $L^{\max} = \frac{6}{671} \cdot \frac{9}{7} \approx 0.0114967$. Using the approach underlying the Chernoff-Hoeffding bound, we obtain the expression $\alpha = 2e^{-2w^2N^2/(L^{\max})^2}$, for the confidence interval half-width w and confidence level α after having drawn N samples, which leads to:

$$w = \sqrt{\log\left(\frac{2}{\alpha}\right) \cdot \frac{(L^{\max})^2}{2N}}.$$

We will compare this confidence interval to the one obtained using the Central Limit Theorem in Section 5.

4 CLT-based tests for importance sampling

Before we discuss the use of CLT-based tests for importance sampling, we will first spend a few words on the validity of the normality assumption in an Importance Sampling context.

4.1 Correctness of CLT-based tests

All hypothesis tests based on the central limit theorem rely on the assumption that the number of samples N is large enough to warrant the use of the CLT; i.e., that the distribution of Z_N is sufficiently close to normal. This does not just hold for hypothesis tests, but also for establishing confidence intervals around a point estimate. However, in general there is no way of knowing when N is sufficiently large.

The fact that there is no way of being sure that N is large enough, has caused many practitioners to prefer other, more rigorous tests. Indeed, in the Bernoulli case, there are good alternatives as noted earlier. Then again, precisely in the Bernoulli case, one may be able to make slight adjustments to the CLT interval to make it conservative (e.g. [1, 5]).

However, in many cases using the CLT is the only option, and generally accepted as such by practitioners. One such case is using standard (i.e., not importance sampling) simulation to estimate the mean of a non-Bernoulli, and in general not a-priori bounded, random variable, such as a waiting time in a queueing model. At any finite N, one has no assurance that there cannot still later come a very rare, very large X_i that will significantly change the estimate of the mean and variance. The practitioner simply trusts this will not happen, based on his/her understanding of the model.

When using importance sampling with a good change of measure, the distribution of the likelihood ratios will not have a long tail, and the CLT can give a good estimate of the mean and a confidence interval around it. However, a bad change of measure may lead to a distribution of likelihood ratios which does have a long tail, having very large values occurring very rarely, requiring very large N. Among importance sampling practitioners, it is customary to do one's best to make a good change of measure (e.g., one with such nice properties as asymptotic efficiency or bounded normal approximation, cf. [12]), and then apply the CLT to obtain a confidence interval.

We argue that using a CLT-based hypothesis test with importance sampling simulation is not fundamentally different or more "dangerous" than using the CLT to obtain a confidence interval. In either case, one makes a statement about being e.g. 95% sure that the true value is in some interval. So if obtaining confidence intervals from the CLT is deemed reliable in some importance sampling simulation, then hypothesis tests based on the CLT should also be considered reliable.

4.2 Suitability of CLT-based tests for importance sampling

As listed in Table 1, there are several hypothesis tests based on the CLT. Unfortunately, some of those require knowledge of the estimator variance as a function of the probability p of interest. This is used to compute *in advance* how many samples N will be enough to draw the right conclusion with the prescribed confidence level even in the worst (most difficult) case, which typically occurs when p is at or near p_0. In the Bernoulli case, the estimator variance can indeed be

computed for any given p. But in the importance sampling case, this is generally impossible. In fact, the question in that case is meaningless, since for a given p (without further information) there can be many different distributions for X_i, with different variances.

The only test from the table which does not require knowing variance as a function of p, is the Chow-Robbins test. This test is based on a theorem by Chow and Robbins [6] which says that if one wants a confidence interval of predetermined width, one can just keep adding samples and increase N until the CLT indicates that this width has been reached, based on the observed sample variance. This is made into a hypothesis test by simulating long enough so that the half-width of the confidence interval on Z_N is less than ζN, where ζ is an indifference level: if $|p-p_0| < \zeta$, one is willing to accept that the test's probability of terminating conclusively may be less than the specified confidence level (e.g., 95 %).

4.3 Extension of the Chow-Robbins test to class I and III

The Chow-Robbins test as discussed above is a class-II test: it risks terminating inconclusively if p is near p_0. However, the same principle can be used to form a class-I or class-III test, as described briefly below.

For a class-I test, the requirement is that the probability of accepting the wrong hypothesis is less than α if $|p - p_0| \geq \delta$, where $1 - \alpha$ is the confidence level of the test and δ the indifference level. This can be achieved by choosing the confidence interval halfwidth of Z_N to be δN and its level to be $1 - 2\alpha$. One easily verifies that then if $|p - p_0| = \delta$ (the hardest case) the probability of accepting the wrong hypothesis is at most α.

A crude class-III test can be constructed by concatenating class-II tests as follows. The ith (for $i = 1, 2, \dots$) class-II test is given probability of error of first kind $\alpha_i = \alpha/2^i$, indifference level $\zeta_i = \zeta/2^i$, and probability of error of second kind (i.e., taking no decision) $\beta_i = \beta$; here α is the desired probability of wrong conclusion of the resulting class-III test, and β and ζ are parameters to be chosen. Then apply the first test ($i = 1$). If it draws a conclusion, that is the final conclusion. If it finishes undecided, apply test 2, with *new* samples, and so on, until a conclusion is drawn. Clearly, the total probability of drawing a wrong conclusion is upperbounded by $\sum_{i=1}^{N} \alpha_i = \alpha$, as required, and the fact that $\zeta_i \to 0$ makes sure a conclusion is eventually reached.

5 Numerical results

In this section we present numerical results to illustrate the results of Section 3. Since all tests that we are still considering are based on confidence intervals, we show results on confidence interval coverage levels here, rather than results on hypothesis test decision correctness (which would be equivalent).

In particular we present three tables. Table 2 displays sample 95% confidence intervals created using both the CLT and the Chernoff-Hoeffding bound using

N	MC-Gauss	MC-Ch.-Hffd.	BFB-Gauss	BFB-Ch.-Hffd.
10	—	[-4.295E-1, 4.295E-1]	[-3.295E-6, 3.080E-5]	[-4.924E-3, 4.951E-3]
30	—	[-2.480E-1, 2.480E-1]	[-1.406E-6, 4.479E-6]	[-2.849E-3, 2.852E-3]
100	—	[-1.358E-1, 1.358E-1]	[6.295E-5, 7.950E-4]	[-1.132E-3, 1.990E-3]
300	—	[-7.841E-2, 7.841E-2]	[4.278E-4, 9.915E-4]	[-1.918E-4, 1.611E-3]
1000	—	[-4.295E-2, 4.295E-2]	[3.315E-4, 5.787E-4]	[-3.867E-5, 9.488E-4]
3000	[2.754E-5, 2.639E-3]	[-2.346E-2, 2.613E-2]	[4.396E-4, 5.925E-4]	[2.309E-4, 8.011E-4]
10000	[6.184E-5, 9.382E-4]	[-1.308E-2, 1.408E-2]	[4.858E-4, 5.701E-4]	[3.718E-4, 6.841E-4]
30000	[5.068E-4, 1.160E-3]	[-7.008E-3, 8.674E-3]	[4.850E-4, 5.328E-4]	[4.187E-4, 5.990E-4]
100000	[3.357E-4, 6.043E-4]	[-3.825E-3, 4.765E-3]	[4.943E-4, 5.203E-4]	[4.579E-4, 5.567E-4]

Table 2. Sample 95% confidence intervals generated using both the Gaussian approximation and the Chernoff-Hoeffding bound for several values of N. This is done for both standard Monte Carlo (MC) simulation and Balanced Failure Biasing (BFB). The confidence intervals are for estimates for p in the benchmark DDS. For both methods, the results in each row are based on the same sample, but results in the lower columns are not continuations of the previous samples. The Gaussian confidence intervals are asymptotically narrower, but for small values they are prone to being incorrect. The true probability equals 5.0285E-4.

both standard Monte Carlo and Balanced Failure Biasing. Table 3 displays coverage statistics, i.e., simulation estimates of the probability that the confidence interval contains the true probability (this should at least be equal to the confidence level). We compare BFB to similar results for standard Monte Carlo (MC) simulation, which is based on Bernoulli samples.

As we can see in Table 2, both standard MC simulation and BFB produce confidence intervals for small values for N that are unreliable (either because they are completely uninformative or wrong), but for high values of N the confidence intervals are narrower than for those based on the Chernoff-Hoeffding bound. For MC, if no likelihood ratios had the value 1 then we cannot construct a meaningful confidence interval. However, this is possible using the Chernoff-Hoeffding bound. Note that for MC better methods for constructing confidence intervals exist such as the Agresti-Coull interval and the exact binomial (Clopper-Pearson) confidence interval. For BFB with small samples sizes, it is reasonably likely that only very small likelihood ratios are observed, leading to confidence intervals that do not contain the true probability using the CLT. However, if we use the Chernoff-Hoeffding bound the confidence intervals are sufficiently conservative.

In Table 3, we display coverage statistics. In particular, we conduct N_1 simulation experiments, where in each experiment we use N_2 samples to create a confidence interval and then check whether this interval contains the true probability. For the case of MC Gauss, two ways of treating the (rather likely) case where all simulation runs result in 0: it can be counted as giving a confidence interval of $[-\infty, \infty]$ and thus indeed containing the true value, but since $[-\infty, \infty]$ is totally uninformative, from a practical point of view it makes more sense to not count it as a correct confidence interval. Only for very large values of N_2 will the coverage of MC Gauss approach 95%. BFB Gauss's coverage approaches

N_2	MC Gauss 1	MC Gauss 2	MC Ch.-Hffd.	BFB Gauss	BFB Ch.-Hffd.
10	0.0043 ± 0.0013	1.0000 ± 0.0000	1.0000 ± 0.0000	0.4298 ± 0.0097	1.0000 ± 0.0000
30	0.0116 ± 0.0021	1.0000 ± 0.0000	1.0000 ± 0.0000	0.8153 ± 0.0076	1.0000 ± 0.0000
100	0.0463 ± 0.0041	1.0000 ± 0.0000	1.0000 ± 0.0000	0.8907 ± 0.0061	1.0000 ± 0.0000
300	0.1384 ± 0.0068	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9341 ± 0.0049	1.0000 ± 0.0000
1000	0.3886 ± 0.0096	0.9998 ± 0.0003	1.0000 ± 0.0000	0.9458 ± 0.0044	1.0000 ± 0.0000
3000	0.7830 ± 0.0081	0.9991 ± 0.0006	1.0000 ± 0.0000	0.9502 ± 0.0043	1.0000 ± 0.0000
10000	0.8742 ± 0.0065	0.8777 ± 0.0064	1.0000 ± 0.0000	0.9470 ± 0.0044	1.0000 ± 0.0000

Table 3. Coverage results for the DDS benchmark setting. $N_1 = 10000$. In MC Gauss 1, a sample with only zeroes is counted as producing an incorrect interval, while in MC Gauss 2, it is counted as producing a correct (but non-informative) interval of $[-\infty, \infty]$.

95% much earlier. As we can see, the Chernoff-Hoeffding-based results are much more reliable than the Gauss-based results.

Note that the good performance of the methods based on the CLT for high N_2 justified their use as discussed in Section 4. Of course, it depends on the application when N_2 is high 'enough', whereas the Chernoff-Hoeffding-based methods are safe regardless of the choice of N_2. On the other hand, the Chernoff-Hoeffding-based methods clearly are rather conservative and thus such a test would take more simulation effort than strictly needed to come to a conclusion with the requisite confidence level.

6 Conclusions

In this short paper we have considered the options for hypothesis tests for importance sampling Two approaches seem promising: tests which work if the likelihood ratio is upper bounded, and tests based on the Chow-Robbins theorem if the normal approximation is known to be applicable (i.e., the number of samples high enough). For the former we have shown that for a particular class of models the well-known BFB heuristic indeed has an upper bound on the likelihood ratio. Two obvious lines for future work are (i) finding more general ways of constructing changes of measure with provably bounded likelihood ratio, and (ii) finding ways of establishing whether the normal approximation is indeed applicable.

Acknowledgments

This work is partially supported by the EU projects SENSATION, 318490, and QUANTICOL, 600708.

References

1. A. Agresti and B.A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

2. C. Alexopoulos and B.C. Shultes. Estimating reliability measures for highly-dependable markov systems, using balanced likelihood ratios. *IEEE Transactions on Reliability*, 50(3):265–280, Sep 2001.
3. B. Barbot. *Acceleration for statistical model checking.* PhD thesis, École normale supérieure de Cachan, 2014.
4. B. Barbot, S. Haddad, and C. Picaronny. Coupling and importance sampling for statistical model checking. In *Tools and Algorithms for the Construction and Analysis of Systems*, pages 331–346. LNCS Volume 7214, Springer, 2012.
5. L.D. Brown, T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–117, 2001.
6. 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.
7. T. Dean and P. Dupuis. Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic Processes and their Applications*, 119:562–587, 2009.
8. A. Goyal, P. Shahabuddin, P. Heidelberger, V.F. Nicola, and P.W. Glynn. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers*, 41(1):36–51, 1992.
9. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.
10. C. Jegourel. *Rare Event Simulation for Statistical Model Checking.* PhD thesis, Université de Rennes 1, 2014.
11. H. Kahn and T.E. Harris. Estimation of particle transmission by random sampling. In *Monte Carlo method; proceedings of a symposium held June 29, 30, and July 1, 1949*, volume 12 of *Nat. Bur. Standards Appl. Math. Series*, pages 27–30, 1951.
12. P. L'Ecuyer, J. Blanchet, B. Tuffin, and P. Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 20(1):6, 2010.
13. A. Legay, B. Delahaye, and S. Bensalem. Statistical model checking: an overview. In *Runtime Verification*, pages 122–135. Springer, 2010.
14. D.P. Reijsbergen. *Efficient simulation techniques for stochastic model checking.* PhD thesis, University of Twente, Enschede, December 2013.
15. D.P. Reijsbergen, P.T. de Boer, W. Scheinhardt, and B.R. Haverkort. Fast simulation for slow paths in Markov models. In *Proceedings of RESIM 2012*, pages 36–38, 2012.
16. D.P. Reijsbergen, P.T. de Boer, W.R.W. Scheinhardt, and B.R. Haverkort. On hypothesis testing for statistical model checking. *International Journal on Software Tools for Technology Transfer*, 17(4):377–395, 2015.
17. W.H. Sanders and L.M. Malhis. Dependability evaluation using composed SAN-based reward models. *Journal of parallel and distributed computing*, 15(3):238–254, 1992.
18. P. Shahabuddin. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, pages 333–352, 1994.
19. H.L.S. Younes. Error control for probabilistic model checking. In *Verification, Model Checking, and Abstract Interpretation*, pages 142–156. Springer, 2006.
20. Companion website to our paper [16].
http://wwwhome.ewi.utwente.nl/∼ptdeboer/hyptest-for-smc/.