Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

STATISTICAL ANALYSIS AND COMPARISON OF SIMULATION MODELS OF HIGHLY DEPENDABLE SYSTEMS - AN EXPERIMENTAL STUDY

Peter Buchholz
Dennis Müller

Informatik IV
TU Dortmund
D-44221 Dortmund, GERMANY

ABSTRACT

The validation of dependability or performance requirements is often done experimentally using simulation experiments. In several applications, the experiments have a binary output which describes whether a requirement is met or not. In highly dependable systems the probability of missing a requirement is 10^{-6} or below which implies that statistically significant results have to be computed for binomial distributions with a small probability. In this paper we compare different methods to statistically evaluate simulation experiments with highly dependable systems. Some of the available methods are extended slightly to handle small probabilities and large samples sizes. Different problems like the computation of one or two sided confidence intervals, the comparison of different systems and the ranking of systems are considered.

1 INTRODUCTION

Contemporary systems have to fulfill very strict requirements concerning performance and dependability. In many application areas highly dependable systems are used such that requirements have to hold with a probability of 99.9999% which implies that probabilities of system failures are in $O(10^{-6})$. To assure that a system meets such requirements, a detailed model based analysis becomes necessary. We consider in this paper cases where simulation experiments are applied to analyze whether a system meets some quantitative requirements that hold with a very high probability. We further assume that the outcome of a single experiment is binary, i.e., the requirement is met or not. Since the output is stochastic, a statistical analysis is necessary to obtain reliable results with a predefined significance probability. Consequently, confidence intervals have to be computed with a range that is determined by the probability of missing the requirement, e.g., one wants to compute a confidence interval with a width of 10% around the estimated probability of a system failure which implies that the width of the confidence interval is in $O(10^{-(k+1)})$ for a probability of $O(10^{-k})$. For $k \geq 6$ a huge number of experiments has to be performed. However, due to the increase in computing power and the availability of efficient simulation systems such experiments can be made. Other problems related to the evaluation of a single system are the comparison of two or more systems and the selection of the system with the highest probability of meeting a requirement from a set of systems. The later two problems are denoted as (multiple) statistical comparisons and ranking and selection.

The mentioned problems are well known for a long time in stochastic simulation such that various methods for statistical evaluation exist and are even published in simulation textbooks (Banks, Nelson, and Nicol 2004, Law and Kelton 2000), but topics like ranking and selection are still important research topics (Swisher, Jacobson, and Yücesan 2003). However, although the problems we consider in this paper are in principle known in simulation, available evaluation methods usually make assumptions that do not hold in our case. The common assumption is that observations are independent and normally distributed. In our case the outcome of the experiments follows a binomial distribution with a very small probability. Thus, even if according to the central limit theorem the binomial distribution converges towards a normal distribution for the number of observations n going to infinity, in practical cases, where the number of experiments is finite, the use of the available methods based on the normal distribution for binomial distributions results often in too optimistic results which overestimate the true significance probabilities. This is in principle known and has been analyzed in statistics (Chang et al. 2008), in particular for the computation of confidence intervals (Brown, Cai, and DasGupta 2001). However, the approaches from statistics do not consider small probabilities in the range of 10^{-6} or large sample sizes and they also do not consider the use of the statistical techniques for simulation output analysis. In this paper we address the following five problems:

1. The computation of one and two sided confidence intervals for the estimated probability of missing the performance/dependability requirement in Section 3.
2. The comparison of the probability of missing the performance/dependability measure with a standard in Section 4.
3. The comparison of different systems in Section 5.
4. The screening of a set of systems to select inferior systems in Section 6.
5. The selection of the best system from a set of systems in Section 7.

In each section we present different methods which are available in the literature, extend some of the methods slightly, and perform experiments to analyze the quality of the results. Before we consider the concrete problems, some basic notations and definitions are introduced in the first section. The paper ends with a short summary of the results we derived from the experiments.

2 BASIC NOTATIONS AND DEFINITIONS

We consider simulation or real experiments with a binary output, where the output is 0 if the systems meets the requirement and where it is 1 if the system does not meet the requirement which is denoted as a *miss*. Let $p \ll 1$ be the probability that the outcome of the experiment is 1. For contemporary technical systems p in the range of $10^{-6} - 10^{-8}$ is a common value. We assume that n experiments are performed and denote by s the number of experiments with outcome 1. Then

$$\hat{p} = \frac{s}{n} \tag{1}$$

is an estimator for p. The number of misses s is the realization of a random variable S with binomial distribution such that

$$\binom{n}{s} p^s (1-p)^{n-s} \tag{2}$$

is the probability of observing s *misses*, if the probability of a miss is p. The variance of S equals $p(1-p)$. We use the notation $S \sim Bin(n,p)$ to express that S has a binomial distribution with parameters n and p.

It is well known that for large n the distribution of S may be approximated by a normal distribution with mean p and standard deviation $\sqrt{p(1-p)}$. We denote by $z_{1-\alpha}$ the $1-\alpha$-quantile of the standard normal distribution. If the binomial distribution is approximated by a normal distribution, then the large set of statistical methods based on normal distributions can be applied. However, it is also known (Brown et al. 2001, Chang et al. 2008) that the convergence of the binomial distribution towards the normal distribution is very slow if p is much smaller than 0.5 which is in our applications clearly the case. Several heuristics are available to decide whether the normal approximation is adequate. Some of the heuristics taken from Brown, Cai, and DasGupta (2001), Leemis and Trivedi (1996) are $np \geq 5$, or $np(1-p) \geq 5$, or $n\hat{p} \geq 5$, or $\hat{p} \pm 3\sqrt{\hat{p}(1-\hat{p})/n}$ does not contain 0. As already shown in Brown, Cai, and DasGupta (2001) and also by some of our examples below, these heuristics are useless in extreme cases where p is small.

One approach to make observations more normally distributed is to batch them before statistical methods are applied. Let x_i the result of the i-th experiment and let b be the batch size, then $\bar{n} = \lfloor n/b \rfloor$ values are computed such that

$$y_j = \frac{\sum_{i=(j-1)\cdot b}^{j\cdot b-1} x_i}{b} \text{ and } \hat{p} = \frac{\sum_{j=1}^{\bar{n}} y_j}{\bar{n}} \ . \tag{3}$$

However, the y_j are no longer binomially distributed. The variance of the batched observations is then given by

$$\hat{\sigma}^2 = \frac{1}{\bar{n}-1} \sum_{j=1}^{\bar{n}} (y_j - \hat{p})^2 \ . \tag{4}$$

For increasing batch sizes the y_j become more normally distributed.

If we compare systems (e.g., to find the system with the smallest probability of missing a requirement), then we assume that K systems are available and denote the related quantities as p_k, $\hat{p}_k$, n_k $\bar{n}_k$, $\hat{\sigma}_k^2$, s_k, $y_{j,k}$ and $x_{i,k}$, respectively.

The statistical methods we present in the subsequent sections are analyzed analytically, if this is possible, however, often an experimental evaluation is necessary. Experiments are performed as simple simulation experiments using a binomial

distribution with probability p which is very small. To analyze the coverage of the different statistical methods we replicate the experiments m times and measure the percentage of experiments where the result contains the *true* value. The *true* value is computed if the confidence interval includes the probability of a miss or if the best configuration is selected. If α is the required significance probability, then approximately $(1-\alpha)m$ experiments should result in the *true* value. To evaluate the coverage we perform the m experiments several times using different random number streams and compute from these replications the mean coverage and the 90% confidence interval of the coverage. The length of a single experiment depends on the concrete goal and is partially dynamically selected. However, since p is in the range of 10^{-6} about 10^8 experiments are necessary to observe an average of 100 misses.

3 COMPUTATION OF CONFIDENCE INTERVALS

We first consider the analysis of a single system. The goal is to compute a confidence interval for $\hat{p}$. We distinguish between the two sided and the one sided case, i.e.,

$$Prob[p \in [l_\alpha, u_\alpha]] \geq 1-\alpha \text{ and } Prob[p \leq u'_\alpha] \geq 1-\alpha \ . \tag{5}$$

3.1 Interval Estimators

The common confidence intervals that are known from simulation textbooks (Banks, Nelson, and Nicol 2004, Law and Kelton 2000) equal

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{\bar{n}}} \text{ and } \hat{p} + z_{1-\alpha}\sqrt{\frac{\hat{\sigma}^2}{\bar{n}}} \tag{6}$$

for the one and two sided case where $\bar{n} = n$ and $\hat{\sigma}^2 = \hat{p}(1-\hat{p})$ if $b = 1$. If $\bar{n}$ is small, the quantile of the normal distribution is substituted by the quantile of a t distribution with $\bar{n}-1$ degrees of freedom which is denoted as $t_{\bar{n}-1,1-\alpha/2}$. For larger values of $\bar{n}$ the difference between t- and normal distribution is negligible.

In most textbooks it is mentioned that the coverage of the above confidence intervals (i.e., the true probability that p is in the interval) can be much smaller than $1-\alpha$, if binomial distributions are analyzed and an unlucky combination of the parameters n and p is chosen.

In the statistical literature a large number of alternative methods is proposed to compute more reliable confidence intervals for binomial output, e.g., (Agresti and Coull 1998, Brown, Cai, and DasGupta 2001, Wang 2006). We present here some methods that have shown in the mentioned papers to be much better than the standard confidence interval (6), which is denoted as Wald's interval in statistics. The first alternative due to Agresti and Coull (1998) uses a simple redirection of the standard interval by defining

$$\tilde{s} = s + z^2_{1-\alpha/2}/2 \ , \ \tilde{n} = n + z^2_{1-\alpha/2} \text{ and } \tilde{p} = \tilde{s}/\tilde{n} \tag{7}$$

for the two sided case and the corresponding values with $z^2_{1-\alpha}$ for the one sided case. Afterwards, $\tilde{p}$ and $\tilde{n}$ rather than $\hat{p}$ and $\bar{n}$ are used in (6). For a detailed derivation of the redirection of the values which is based on the Wilson score interval we refer to the literature (Agresti and Caffo 2000, Brown, Cai, and DasGupta 2001).

The so called Clopper-Pearson interval (Brown, Cai, and DasGupta 2001) uses directly the binomial distribution and computes the interval $[l,u]$ or the upper bound u' as

$$l = \min_{\theta\in[0,1]} P[Bin(n,\theta)] \geq s] = \alpha/2, \ u = \max_{\theta\in[0,1]} P[Bin(n,\theta)] \leq s] = \alpha/2 \text{ and } u' = \max_{\theta\in[0,1]} P[Bin(n,\theta)] \leq s] = \alpha \ . \tag{8}$$

This interval is sometimes denoted as the exact interval. However, as described in Brown, Cai, and DasGupta (2001), Wang (2006) several other "exact" intervals exist. The direct computation of the values in (8) is cumbersome but it is well known that the binomial distribution is related to the beta function (Brown et al. 2001, Press et al. 2002). Let $B(a,b)$ be the beta function, $B_p(a,b), 0 \leq p \leq 1$ be the incomplete beta function and $I_p(x,y)$ the regularized beta function that are defined as

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, \ B_p(x,y) = \int_0^p t^{x-1}(1-t)^{y-1}dt \text{ and } I_p(x,y) = \frac{B_p(x,y)}{B(x,y)} \ . \tag{9}$$

The cumulative distribution function of the binomial distribution can then be represented as

$$\sum_{i=x}^{n} \binom{n}{i} p^i (1-p)^{n-i} = I_p(x, n-x+1) \tag{10}$$

with $0 \leq p \leq 1$ and $0 < x \leq n$. For $x = 0$ the sum becomes 1 for all p. Note that the regularized beta function is monotonic in p, which simplifies the determination of minimal and maximal values. The values of the beta distribution can be computed numerically, the corresponding procedures are given in (Press et al. 2002) such that

$$\alpha/2 = I_{l_\alpha}(s, n-s+1) \text{ and } (1-\alpha/2) = I_{u_\alpha}(s+1, n-s) \tag{11}$$

have to be solved for l_α and u_α, respectively. Similarly,

$$1-\alpha = I_{u_\alpha}(s+1, n-s) \tag{12}$$

has to be solved in the one sided case. Let $\beta_\alpha(a,b)$ be the α-quantile of the regularized beta function with parameters (a,b).

Since $I_p(x,y)$ is only defined for $x,y > 0$ the remaining cases have to be defined separately. If $s = 0$, then no miss has been observed in the experiments. In this case the lower bound is zero (i.e., $l = 0.0$). The upper bound u has to be chosen such that the significance level of the confidence interval becomes α. We obtain

$$1-\alpha = \sum_{i=1}^{n} \binom{n}{i} u_\alpha^i (1-u_\alpha)^{n-i} = 1-(1-u_\alpha)^n \;\Rightarrow\; u'_\alpha = u_\alpha = 1 - \sqrt[n]{\alpha}\ . \tag{13}$$

The case $s = n$ is not interesting here since we consider small miss probabilities but may be handled similarly.

For the determination of the Clopper-Pearson interval numerical computations have to be performed. With today's computing power the boundaries can be computed efficiently and in a numerically stable way. However, the Clopper-Pearson interval is known to be conservative such that in the statistical literature often other intervals are recommended (Agresti and Coull 1998, Brown, Cai, and DasGupta 2001, Leemis and Trivedi 1996, Wang 2006). One recommended alternative from Brown, Cai, and DasGupta (2001) is Jefferey's interval which is defined by the following boundaries.

$$l = \beta_{\alpha/2}(s+0.5, n-s+0.5),\ u = \beta_{1-\alpha/2}(s+0.5, n-s+0.5) \text{ and } \ u' = \beta_{1-\alpha}(s+0.5, n-s+0.5)\ . \tag{14}$$

For $s = 0$ the interval is computed via (13). As shown in Brown, Cai, and DasGupta (2001) (14) is always contained in (8).

3.2 Experimental Results

For a fixed number of samples and a batch size of one, coverage and width of the confidence intervals can be computed analytically by a complete enumeration of the number of misses. For every number of misses s the probability can be computed via (2) and the different confidence intervals can also be computed such that coverage and average width can be computed without experiments. If p is small, then a complete enumeration is not necessary because the probability of observing s misses converges rapidly towards 0 for an increasing s. If the batch size is larger than one, an enumeration is too costly and experiments are performed to determine coverage and average width.

Table 1 shows the results for $p = 10^{-6}$, 10^7 samples and $\alpha = 0.05$. The standard interval (6) has the smallest width but it has only a coverage of 0.926 which is less then the required coverage of 0.95. Batching does not really improve the situation. We tried several batch sizes and the best was 10^6 such that only 10 observations are evaluated in a single experiment to compute the confidence interval. A larger batch size does not increase the coverage but it increases the width. A smaller batch size results in a lower coverage. However, even for batch size 10^6, the required coverage of 0.95 lies outside the confidence interval for the coverage and the width of the confidence intervals is larger than in all other cases. From the other three intervals, the Jefferey interval has a coverage of 0.944 which is below 0.95. (7) and (8) have a coverage above 0.95 and also have a similar width. The examples show nicely that even for a large number of observations, it is preferable to apply the more sophisticated methods rather than the textbook approach. If we consider only 10^6 observations, then the coverage of the standard interval goes down to 0.63, whereas the other methods have a coverage of 0.98.

It is known that for small sample sizes the coverage of all methods is irregular and depends on p and the sample size (Brown, Cai, and DasGupta 2001). Figure 1 shows that the same holds for small p and fairly large sample sizes. We computed the coverage of the different methods for $p = 10^{-6}$ and vary the sample size from 10 millions to 50 millions

Table 1: Quality of the confidence intervals for 10^7 samples

method	avg. width	coverage
Eq. (6) ($b = 10^6$)	$1.379 \cdot 10^{-6}$ $(\pm 1.66 \cdot 10^{-8})$	0.940 (± 0.0061)
Eq. (6) ($b = 1$)	$1.223 \cdot 10^{-6}$	0.926
Eq. (7)	$1.341 \cdot 10^{-6}$	0.963
Eq. (14)	$1.244 \cdot 10^{-6}$	0.944
Eq. (8)	$1.344 \cdot 10^{-6}$	0.975

in steps of 1 million. It can be be seen that the standard interval has the worst coverage which is always below 0.95, the exact approach (8) is always above 0.95 and the remaining two approaches are most times but not always above 0.95. The general structure of our results is similar to the small sample size case with larger values of p (Brown, Cai, and DasGupta 2001). However, apart from the observation that results become better with an increasing sample size, not much can be said about specific cases. One can see that for each method there are lucky and unlucky sample sizes and the behavior is non monotonic. Unfortunately, lucky cases may become unlucky if p is modified slightly.

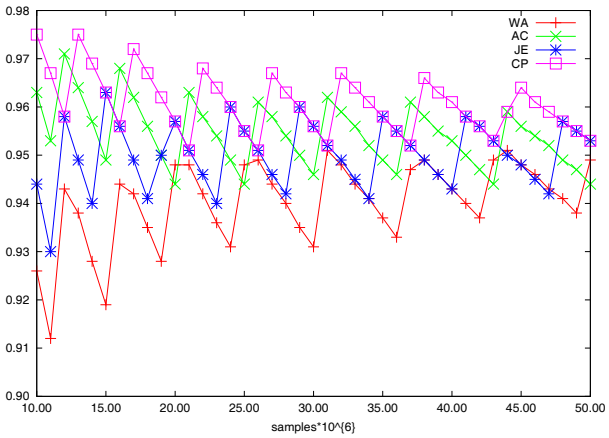

Figure 1: Coverage of the methods for $p = 10^{-6}$

In a second series of experiments we compute confidence intervals with a width of 20% of the estimated mean and set $\alpha = 0.1$. Experiments are stopped when the confidence intervals are small enough. Confidence intervals are computed dynamically during the experiment (Banks, Nelson, and Nicol 2004, pp. 414). This setting cannot be evaluated analytically.

The results are shown in Table 2. For these experiments we consider the average number of samples which is needed to compute the confidence intervals of the required width. All methods reach the required coverage of 0.9. The "exact" method (8) is slightly more conservative than the other approaches and, consequently, needs about 5% more samples than the other methods which are all very similar. The effort for computing the confidence intervals (14) and (8) is slightly higher than in the cases, where the quantile of the t or normal distribution is used because the beta function has to be evaluated numerically. However, a single run computing the confidence interval with a width of 20% of $\hat{p}$ needs about 9.7 second for (6), (7) and about 11.3 seconds for (14), (8). If real simulation experiments are applied, which need much more time than the simple generation of random numbers, then the difference for computing the confidence intervals becomes negligible.

4 COMPARISON WITH A STANDARD

In the following we consider the case where we compare the miss probability of one or more systems with a standard. The approach may be used to compute the probability that a system fullfils the requirement to have a miss probability of less than a predefined threshold q.

Table 2: Quality of the confidence intervals with a width of less than 20% of $\hat{p}$

method	samples		coverage	
Eq. (6)($b=1$)	$2.710 \cdot 10^{+8}$	$(\pm 2.47 \cdot 10^{+5})$	0.907	(± 0.0039)
Eq. (7)	$2.701 \cdot 10^{+8}$	$(\pm 2.83 \cdot 10^{+5})$	0.908	(± 0.0036)
Eq. (14)	$2.711 \cdot 10^{+8}$	$(\pm 2.47 \cdot 10^{+5})$	0.907	(± 0.0033)
Eq. (8)	$2.811 \cdot 10^{+8}$	$(\pm 2.47 \cdot 10^{+5})$	0.912	(± 0.0109)

4.1 Statistical Methods

In principle, we can use the approaches from the previous section to compute one sided confidence intervals. We denote by u'_α the upper bound for the miss probability computed with one of the methods presented above. If $u'_\alpha \leq q$, then the probability of missing the threshold is less than α. To compute a better bound for the probability, one can compute

$$\gamma = \arg \min_{\alpha \in (0,1)} \left(u'_\alpha \leq q \right) \tag{15}$$

using one of the approaches presented in the previous section to compute one sided confidence intervals. $1-\gamma$ is then an estimate for the probability of having miss probability less or equal to q.

The approach can also be applied to compute probabilities for K systems resulting in probabilities γ_k ($k = 1, \ldots, K$). If all experiments are independent, then the probability that all systems have miss probability less than or equal to q is denoted as $\gamma_{1:K}$ and can be computed as

$$1 - \gamma_{1:k} = \prod_{k=1}^{K} (1 - \gamma_k) \ . \tag{16}$$

4.2 Experimental Results

Since in (7), the confidence interval depends on α which is computed as a result here, it is impractical to use the method of Agresti and Coul for this setting. Furthermore, our experiments showed that also the use of Jefferey's method is not recommended since it is not better than the other methods (6) and (8).

We set $q = 10^{-6}$ and examine three cases for p, namely $p = 9 \cdot 10^{-7}$, $p = 10^{-6}$ and $p = 1.1 \cdot 10^{-6}$. In the first case, the probability that $\hat{p} < q$ should converge towards 1 for an increasing n. In the second and third case, it should converge towards 0.5 and 0.0, respectively. Table 3 summarizes the results which can be determined analytically. The results show that both methods, the standard approach and the exact approach behave very similar for this problem. The exact approach is slightly better for $p < q$ and the standard approach is better for $p > q$.

Table 3: Probability that p is smaller than 10^{-6}

	$P[p < 10^{-6}]$					
	Eq. (6)			Eq. (8)		
	samples			samples		
p	10^7	10^9	10^9	10^7	10^9	10^9
$0.9 \cdot 10^{-6}$	0.621	0.775	0.990	0.635	0.777	0.990
10^{-6}	0.530	0.509	0.503	0.550	0.514	0.505
$1.1 \cdot 10^{-6}$	0.443	0.254	0.015	0.457	0.256	0.015

Figure 2 shows the probability of $P[p < 10^{-6}]$ for p between $0.5 \cdot 10^{-6}$ and $1.5 \cdot 10^{-6}$ using 10^7 and 10^8 samples, respectively. The difference between the standard and the exact method is negligible and almost not visible in the figure. The results show that 10^7 experiments are not sufficient to validate or disprove $p < 10^{-6}$, if p is not much smaller or larger than 10^{-6}, whereas the results for 10^8 experiments are much better and allow a clear decision as long as p is not close to the threshold.

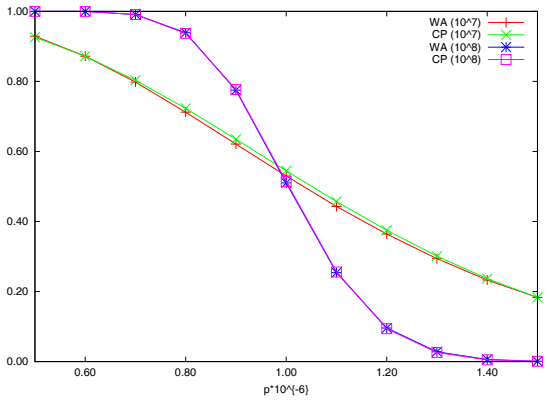

Figure 2: $P[p < 10^{-6}]$ for different values of p and 10^7 or 10^8 samples

5 MULTIPLE COMPARISONS

Let p_1 and p_2 be the probabilities for two different systems to miss a performance or dependability requirement. The goal is to compute a confidence interval for $p_1 - p_2$.

5.1 Methods to Compute Confidence Intervals for Comparisons

The comparison of two or more systems is a commonly known problem which is treated in many textbooks on stochastic simulation (Banks, Nelson, and Nicol 2004, Law and Kelton 2000). We consider first the case of two systems and assume that experiments are independent. Since the variance equals $p_i(1-p_i)$, we have to use for the computation of confidence intervals for $\hat{p}_1 - \hat{p}_2$ an approach for unequal variances. The standard textbook approach for normal distributed data results in the following confidence interval (Banks, Nelson, and Nicol 2004, Law and Kelton 2000).

$$(\hat{p}_1 - \hat{p}_2) \pm t_{\nu,1-\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{\bar{n}_1} + \frac{\hat{\sigma}_2^2}{\bar{n}_2}} \text{ where } \nu = \frac{\left(\hat{\sigma}_1^2/\bar{n}_1 + \hat{\sigma}_2^2/\bar{n}_2\right)^2}{\left(\left(\hat{\sigma}_1^2/\bar{n}_1\right)^2/(\bar{n}_1-1) + \left(\hat{\sigma}_2^2/\bar{n}_2\right)^2/(\bar{n}_2-1)\right)} \tag{17}$$

is an approximation for the degrees of freedom used in the t distribution which is usually rounded to the nearest integer. The formula can be applied to the original data such that $\hat{\sigma}_i^2 = \hat{p}_i(1-\hat{p}_i)$ or using batches like for the computation of confidence intervals.

The computation of confidence intervals for the parameters of two binomial distributions is also a topic in statistics (Santner et al. 2007). In Agresti and Caffo (2000) the authors propose to use the idea of (7) also for two binomial distributions. Thus, rather than n_k and p_k

$$\tilde{s}_k = s_k + z_{1-\alpha/2}^2/4,\ \tilde{n}_k = n + z_{1-\alpha/2}^2/2 \text{ and } \tilde{p}_k = \tilde{s}_k/\tilde{n}_k\ . \tag{18}$$

are used. Then (17) becomes

$$(\tilde{p}_1 - \tilde{p}_2) \pm t_{\nu',1-\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2}} \tag{19}$$

where ν' results from (17) with $\bar{n}_k = \tilde{n}_k$ and $\hat{\sigma}_k^2 = \tilde{p}_k(1-\tilde{p}_k)$ $(k = 1,2)$.

The following method has been developed by Newcombe (1998) as a result of comparing 11 methods for computing confidence intervals for the difference between two binomial parameters. The interval is given by

$$(\hat{p}_1 - \hat{p}_2) - \sqrt{(\hat{p}_1 - x_1)^2 + (y_2 - \hat{p}_2)^2},\ (\hat{p}_1 - \hat{p}_2) + \sqrt{(\hat{p}_2 - x_2)^2 + (y_1 - \hat{p}_1)^2} \tag{20}$$

where

$$x_k = \frac{2s_k + z^2_{1-\alpha/2} - z_{1-\alpha/2}\sqrt{z^2_{1-\alpha/2} + 4s_k(1-s_k/n_k)}}{2(n_k + z^2_{1-\alpha/2})} \ , \ y_k = \frac{2s_k + z^2_{1-\alpha/2} + z_{1-\alpha/2}\sqrt{z^2_{1-\alpha/2} + 4s_k(1-s_k/n_k)}}{2(n_k + z^2_{1-\alpha/2})} \ . \tag{21}$$

There are several other methods to compute confidence intervals for the difference of two parameters. In particular methods based on the inversion of some test statistics have been developed and are included in several statics program packages (Santner et al. 2007). However, these methods require numerical computations to evaluate test statistics over the search space with $n_1 n_2$ elements. In the small-sample case which is considered in the mentioned papers and supported by the software packages, the computation is manageable. If n_1 and n_2 are large, then the required computations become prohibitive such that the methods in their current form cannot be applied and are not considered here.

It is straightforward to modify all approaches such that instead of a two sided, one sided confidence intervals are computed.

If K different systems should be compared, then the comparisons are no longer independent. Assume that C comparisons have to be made and the probability that all confidence intervals include the true differences should be at least $1-\alpha$, then according to the Bonferroni inequality (Banks, Nelson, and Nicol 2004, p.468) the following relation holds.

$$(1-\alpha) \geq (1 - \sum_{c=1}^{C} \alpha_c) \tag{22}$$

where α_c is the significance probability of comparison c. Usually this gives a conservative bound, in particular if C is large. If all systems are compared, we have $C = K(K-1)/2$. The above equation implies that for a single confidence interval $\alpha_c = \alpha/C$ has to be chosen.

5.2 Experimental Results

Table 4 contains the results for two different configurations where 20 and 100 millions of samples are drawn. For 100 million samples, the coverage of all methods is very close to 95% the required significance probability. In this case also the width of the confidence intervals is very similar. For 20 million samples the standard method does not reach the required coverage whereas for the other two methods the coverage is in all experiments above 95%. Since (18) usually gives a slightly tighter confidence interval than (20), this method is recommended.

6 SCREENING PROCEDURES

The decision which of a set of K systems to choose can be formulated in two different ways. First, one can select a subset of $H \leq K$ systems which contains with probability of at least $1-\alpha$ the best system or, second, one can select the "best" system with a predefined probability. In this section we analyze methods for the first formulation which is denoted as subset selection or screening in simulation (Swisher, Jacobson, and Yücesan 2003). The second formulation is analyzed afterwards.

Before we consider the statistical methods, we describe the general setting. We consider K different systems with miss probabilities p_k and assume without loss of generality $p_1 \leq p_2 \leq \ldots \leq p_K$. Furthermore it is not possible to select the best system, when probabilities are arbitrarily close. Thus, it is usually assumed that $\delta > 0$ is available such that differences that are smaller than δ are insignificant. In our setting the goal is to find a system k such that $|p_k - p_1| \leq \delta$. Of course, if $|p_2 - p_1| > \delta$, then the first system is the only system that should be selected. δ is denoted an indifferent zone parameter.

6.1 Statistical Methods

A screening procedure which is often used in simulation has been proposed in (Bechthofer, Santner, and Goldsman 1995) and used for subset selection in simulation by (Boesel, Nelson, and Kim 2003, Nelson, Swann, Goldsman, and Song 2001). We assume that α is chosen from $(0,(K-1)/K)$.

The idea of the screening procedure is to perform a pairwise comparison of systems and eliminate all systems which have shown to be inferior to at least one system which remains in the subset. Define the following coefficients which are similar to the confidence interval in (17),

$$W_{ij} = \sqrt{\frac{t^2_{n_i-1,(1-\alpha/2)^{1/k}} \hat{\sigma}_i^2}{n_i} + \frac{t^2_{n_j-1,(1-\alpha/2)^{1/k}} \hat{\sigma}_j^2}{n_j}} \ . \tag{23}$$

Table 4: Confidence intervals for $p_1 - p_2$ with $\alpha = 0.05$

n_1	n_2	method	avg. width	coverage	n_1	n_2	method	avg. width	coverage
$p_1 = p_2 = 10^{-6}$					$p_1 = 5 \cdot 10^{-7}$ and $p_2 = 1.5 \cdot 10^{-6}$				
$5 \cdot 10^6$	$1.5 \cdot 10^7$	(17)	$1.994 \cdot 10^{-6}$	0.930	$5 \cdot 10^6$	$1.5 \cdot 10^7$	(17)	$2.231 \cdot 10^{-6}$	0.932
		(18)	$2.156 \cdot 10^{-6}$	0.964			(18)	$2.377 \cdot 10^{-6}$	0.960
		(20)	$2.247 \cdot 10^{-6}$	0.958			(20)	$2.243 \cdot 10^{-6}$	0.955
10^7	10^7	(17)	$1.742 \cdot 10^{-6}$	0.948	10^7	10^7	(17)	$1.742 \cdot 10^{-6}$	0.946
		(18)	$1.826 \cdot 10^{-6}$	0.959			(18)	$1.826 \cdot 10^{-6}$	0.957
		(20)	$1.905 \cdot 10^{-6}$	0.958			(20)	$1.901 \cdot 10^{-6}$	0.958
$1.5 \cdot 10^7$	$5 \cdot 10^6$	(17)	$1.994 \cdot 10^{-6}$	0.930	$1.5 \cdot 10^7$	$5 \cdot 10^6$	(17)	$1.728 \cdot 10^{-6}$	0.947
		(18)	$2.156 \cdot 10^{-6}$	0.964			(18)	$1.913 \cdot 10^{-6}$	0.972
		(20)	$2.247 \cdot 10^{-6}$	0.958			(20)	$2.046 \cdot 10^{-6}$	0.963
$2.5 \cdot 10^7$	$7.5 \cdot 10^7$	(17)	$9.03 \cdot 10^{-7}$	0.947	$2.5 \cdot 10^7$	$7.5 \cdot 10^7$	(17)	$1.009 \cdot 10^{-6}$	0.946
		(18)	$9.17 \cdot 10^{-7}$	0.953			(18)	$1.022 \cdot 10^{-6}$	0.952
		(20)	$9.25 \cdot 10^{-7}$	0.951			(20)	$1.027 \cdot 10^{-6}$	0.951
$5 \cdot 10^7$	$5 \cdot 10^7$	(17)	$7.83 \cdot 10^{-7}$	0.950	$5 \cdot 10^7$	$5 \cdot 10^7$	(17)	$7.83 \cdot 10^{-7}$	0.949
		(18)	$7.91 \cdot 10^{-7}$	0.952			(18)	$7.91 \cdot 10^{-7}$	0.952
		(20)	$7.98 \cdot 10^{-7}$	0.952			(20)	$7.98 \cdot 10^{-7}$	0.951
$7.5 \cdot 10^7$	$2.5 \cdot 10^7$	(17)	$9.03 \cdot 10^{-7}$	0.947	$7.5 \cdot 10^7$	$2.5 \cdot 10^7$	(17)	$7.82 \cdot 10^{-7}$	0.949
		(18)	$9.17 \cdot 10^{-7}$	0.953			(18)	$7.99 \cdot 10^{-7}$	0.954
		(20)	$9.25 \cdot 10^{-7}$	0.951			(20)	$8.12 \cdot 10^{-7}$	0.953

The subset $\mathscr{H} \subset \{1,\ldots,K\}$ which contains the system with the smallest miss probability is defined as

$$\mathscr{H} = \left\{h | h \in \{1,\ldots,K\} \wedge \forall k \neq h : \hat{p}_h \leq \hat{p}_k + (W_{hk} - \delta)^+\right\} \tag{24}$$

where $(a)^+ = \max(0,a)$. The proof that $\mathscr{H}$ contains for normally distributed observations the "best" configuration with a probability of at least $1-\alpha$ can be found in the online companion of (Boesel, Nelson, and Kim 2003). Substitution of the maximum by the minimum yields the required result. Although we argued above that the indifferent zone parameter should not be zero for the selection of the best system, for screening also $\delta = 0$ may be used. In this case $\mathscr{H}$ has to contain the system or one of the systems with the smallest miss probability. If $\delta > 0$, then $\mathscr{H}$ has to contain one system k with $|p_k - p_1| \leq \delta$.

The procedure is valid for normally distributed observations and only asymptotically valid for binomial distributions. To make the data more normally distributed, one can use batching and compute the variance via (3) and (4). Alternatively, one can extend the approach by applying the modified values $\tilde{s}_k$, $\tilde{n}_k$ and $\tilde{p}_k$ from (18) such that the coefficient W_{ij} become

$$W_{ij} = \sqrt{\frac{t^2_{n_i-1,(1-\alpha/2)^{1/k}} \tilde{p}_i(1-\tilde{p}_i)}{\tilde{n}_i} + \frac{t^2_{n_j-1,(1-\alpha/2)^{1/k}} \tilde{p}_j(1-\tilde{p}_j}{\tilde{n}_j}} \; . \tag{25}$$

These coefficients are afterwards used to compute $\mathscr{H}$.

6.2 Experimental Results

We compare the behavior of the screening procedure with (23) or (25), respectively. The goal of the experiments is to analyze the coverage and the size of the selected subsets. Coverage for these examples means the percentage of experiments where the selected subset contains a system k with $|p_k - p_1| \leq \delta$. Experiments are performed for $K = 5, 10, 20$, $\delta = 0, 5 \cdot 10^{-8}, 10^{-7}$ and $\alpha = 0.05$. We consider as an example $p_1 = 9 \cdot 10^{-7}$, $p_2 = 10^{-6}$, $p_3 = 1.1 \cdot 10^{-6}$ and $p_4,\ldots,p_K$ uniformly distributed between p_2 and p_3. For (23) we use $n = 5$ and batch sizes of $10^3, 10^4, \ldots, 10^8$ and for (25) $n = 5 \cdot 10^4, 5 \cdot 10^4, \ldots, 5 \cdot 10^9$. Consequently, for both variants the same number of evaluations is performed. For each setting 200 replications are made

which assures that the confidence intervals for the coverage and the size of the subset are so small that they are not visible in the following graphs.

Figure 3 shows the size of $\mathscr{H}$ for the different experiments. As expected $|\mathscr{H}|$ can be reduced by increasing the number of evaluations and δ. It can also be observed, that less than $5 \cdot 10^7$ evaluations are insufficient for a successful selection since no systems are removed from the initial set. On the other hand $5 \cdot 10^8$ evaluations lead to a correct selection of the best system in most cases. (25) reduces $\mathscr{H}$ significantly faster than (23). (23) and (25) are both relatively conservative and therefore the coverage (i.e. the ratio of correct selections) is in both cases larger than 95% (98% and 96%). It is interesting to note that the coverage reaches the smallest value for a medium number of samples. This seems to be due to the fact that in these cases very few misses are observed and a single miss can be responsible to remove a system from the set $\mathscr{H}$.

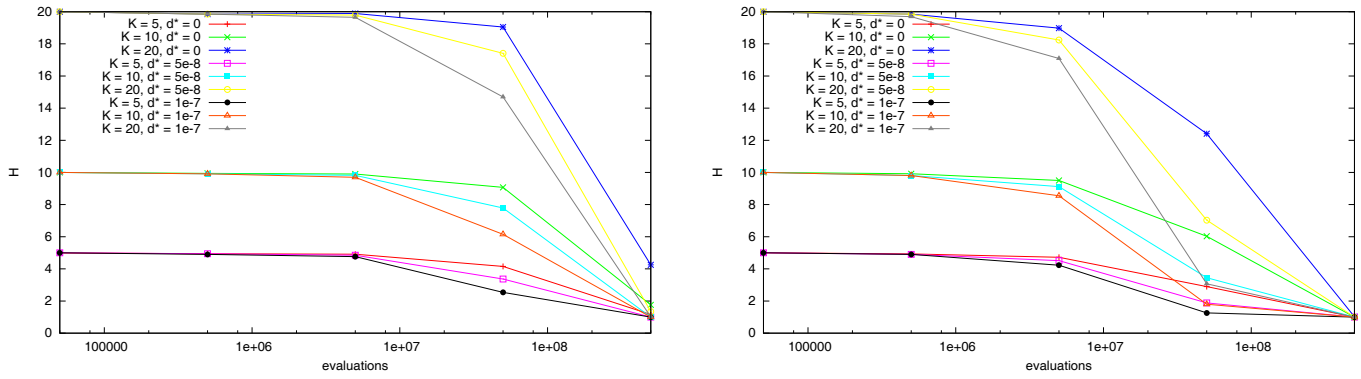

Figure 3: Size of $\mathscr{H}$ of (23) (left side) and (25) (right side)

7 RANKING AND SELECTION

In contrast to a screening procedure where clearly inferior systems are screened out, the goal of a ranking and selection procedure is to find an ordering of systems or to select the "best" system. We restrict our attention to the problem of selecting the system with the smallest probability. The selection procedure should select some system k with $|p_k - p_1| \leq \delta$ with a probability of at least $1-\alpha$. Ranking and selection procedures are most times based on the analysis of the so called *least favorable configuration*. Then a sample size n is determined such that the selection of a system with the smallest number of misses results with probability $1-\alpha$ in the selection of a system k with $|p_k - p_1| < \delta$. The least favorable configuration is in our case $p_1 < p_2 = \ldots = p_{K-1} = p_K = p_1 + \delta$ and $p_1 \approx 0.5 - \delta$ (Buzaianu and Chen 2005, Mulekar and Young 1993, Sobel and Huyett 1957). The ranking and selection procedure for binomial populations from (Sobel and Huyett 1957) computes n with respect to δ and α such that the selection of a system with the smallest s_k will result in a correct selection with probability $1-\alpha$. The correct values for n depending on α and δ can be taken from tables in the literature or have to be computed or approximated. However, for our applications δ has to be very small in the range of 10^{-7} or below. The corresponding values for n are not available yet but can be generated from the following equation which computes the probability that for the first system $s = 0,1,\ldots,n$ misses are observed and for the systems $2,\ldots,K$ at least s misses are observed. In case of an equal number of misses, the best configuration is selected randomly.

$$1-\alpha \geq \sum_{s=0}^{n} \binom{n}{s} (p-\delta)^s (1-p+\delta)^{n-s} \left(\sum_{k=0}^{K-1} \binom{K-1}{k} \frac{1}{k+1} \left(\binom{n}{s} p^s (1-p)^{n-s} \right)^k \left(\sum_{t=s+1}^{n} \binom{n}{t} p^t (1-p)^{n-t} \right)^{K-k-1} \right) \tag{26}$$

The bound for the correct selection holds for known $p_1 = p - \delta$ and $p_k \geq p$. It says that if n samples are drawn for each system and one of the systems with the smallest number of misses is chosen randomly, then it will be the first system with a probability of at least $1-\alpha$. However, the value of p is unknown and the use of $p \approx 0.5$ which would result in the largest α is too conservative if p is known to be small and the values are hard to compute for large n as necessary when δ is in the range of 10^{-7}. We present some example results below.

An alternative to the use of ranking procedures for binomial distributions are those for normal distributions which are only asymptotically valid. Several of those procedures are available (Swisher, Jacobson, and Yücesan 2003). Most of these

approaches are based on the two stage selection procedure by (Rinott 1978). However, this approach is based on the normal distribution assumption such that the selection probabilities is only asymptotically valid for binomial distributions.

7.1 Experimental Results

We evaluate (26) for $\delta = 10^{-7}$ to obtain the number of samples to select the best configuration with the predefined significance probability. Figure 4 shows the value $1-\alpha$ for $K = 5$ or 10, $n = 10^8$ or 10^9 and different values of p. For $K = 5$ and 10^9 samples per system, we select with probability of at least 0.952 the best system, if we choose the system with the smallest samples mean and assume that $p = 10^{-6}$. If we do the same with $K = 10$, then the probability goes down to 0.914. 10^8 samples are not sufficient, since we obtain only probabilities of 0.483 and 0.329, if we select the system with the smallest sample mean out of 5 or 10 alternatives. The situation is even more worse, if we consider the course of $1-\alpha$ when we increase p. The probability of making the right choice decreases quickly and for $p = 10^{-3}$ we have even for 10^9 samples only probabilities of 0.224 and 0.116 to select the best system out of 5 or 10 systems. This is not much better than a random selection and clearly shows that the use of the least favorable configuration as recommended in statistics is not usable. A priori information about the size of p_1 is necessary and may be gained by doing some experiments with one of the systems to compute an upper bound for the miss probability according to some significance probability α' using the methods presented in section 3. The resulting upper bound can then be used in (26) to compute the probability of a correct selection. If $1-\alpha$ is the probability of a correct selection and we do not reuse the results that have been used to estimate the upper bound, then the overall probability of a correct selection becomes $(1-\alpha)(1-\alpha')$. Unfortunately, our results indicate that fairly exact results have to be available for the miss probability to obtain higher probabilities of choosing the correct system because the selection probability is very sensitive for the choice of p.

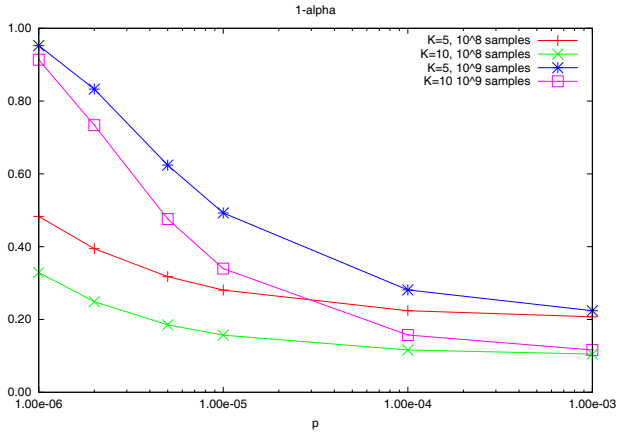

Figure 4: Probability of the right selection for varying values of p

8 CONCLUSIONS AND SUMMARY OF RESULTS

The paper presents methods for the statistical analysis of experiments with highly dependable systems that have a small probability of missing a performance or dependability requirement. It is shown that the use of standard evaluation methods known form stochastic simulation is often crucial since the methods tend to overestimate the significance of the results such that confidence intervals become too small or systems are selected with a smaller probability than expected. On the other hand statistical methods for binomial distributions are usually not designed for very small probabilities and large sample sizes. However, experiments in this paper indicate that these methods can often be adopted for small probabilities and large samples sizes and can also be used to improve the quality of screening or ranking and selection procedures, if they are combined appropriately with methods for normally distributed data. These combined methods usually outperform standard methods for normally distributed data applied to batched observations.

Experiments in the paper are restricted to the independent cases. I.e., all results are stochastically independent. It is known that for simulation the use of common random numbers often results in much more efficient methods when different systems are compared or ranked (Law and Kelton 2000). However, it is also known that common random numbers are often not applicable if different systems are described by different simulation models. Nevertheless, the use of common random number in combination with the different selection methods will be a future research topic.

REFERENCES

Agresti, A., and B. A. Caffo. 2000. Simple and effective confidence intervals for proportions and differences of the proportions result from adding two successes and two failures. *The American Statistican* 54:280–288.

Agresti, A., and B. A. Coull. 1998. Approximate is better than "exact" for interval estimation of binomial properties. *The American Statistican* 52:119–126.

Banks, J., B. L. Nelson, and D. Nicol. 2004. *Discrete-event system simulation*. Prentice Hall.

Bechthofer, R. E., T. J. Santner, and D. Goldsman. 1995. *Design and analysis for statistical selection, screening and multiple comparisons*. Wiley.

Boesel, J., B. L. Nelson, and S. Kim. 2003. Using ranking and selection to "clean up" after simulation optimization. *Operations Research* 51 (5): 814–825.

Brown, L. D., T. T. Cai, and A. DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical Science* 16 (2): 101–133.

Buzaianu, E., and P. Chen. 2005. On selecting among treatments with binomial outcome. *Communications in Statistics - Theory and Methods* 34 (6): 1247–1264.

Chang, C. H., J. J. Lin, N. Pal, and M. C. Ching. 2008. A note on improved approximation of the binomial distribution by the skew-normal distribution. *The American Statistican* 68 (2): 167–170.

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling and analysis*. Wiley.

Leemis, L., and K. S. Trivedi. 1996. A comparison of approximate interval estimators for the bernoulli parameter. *The American Statistican* 50:63–68.

Mulekar, M. S., and L. J. Young. 1993. A fixed sample size selection procedure for negative binomial populations. *Metrika* 40:25–35.

Nelson, B. L., J. Swann, D. Goldsman, and W. Song. 2001. Simple procedures for selecting the best simulated system when the number of alternatives is large. *Operations Research* 49:950–963.

Newcombe, R. 1998. Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine* 17:873–890.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 2002. *Numerical recipes in C++: The art of scientific computing, 2nd ed.* Cambridge University Press.

Rinott, Y. 1978. On two-stage selection procedures and related probability inequalities. *Communications in Statistics - Theory and Methods* A7:799–811.

Santner, T. J., V. Pradhan, P. Senchaudhuri, C. R. Mehta, and A. Tamhane. 2007. Small-sample size comparisons of confidence intervals for the difference of two independent binomial proportions. *Computational Statistics and Data Analysis* 51:5791–5799.

Sobel, M., and M. J. Huyett. 1957. Selecting the best one of several binomial populations. *Bell System Technical Journal* 36:537–576.

Swisher, J. R., S. H. Jacobson, and E. Yücesan. 2003. Discrete-event simulation optimization using ranking, selection, and multiple comparison procedures: A survey. *ACM Transactions on Modeling Computer Simulation* 13 (2): 134–154.

Wang, W. 2006. Smallest confidence intervals for one binomial proportion. *Journal of Statistical Planning and Interference* 136 (12): 4293–4306.

AUTHOR BIOGRAPHIES

PETER BUCHHOLZ received the Diploma degree (1987), the Doctoral degree (1991) and the Habilitation degree (1996) all from the TU Dortmund, where he is currently a professor for modeling and simulation. His current research interests are efficient techniques for the analysis of stochastic models, formal methods for the analysis of discrete event systems, the development of modeling tools, as well as performance and dependability analysis of computer and communication systems. His e-mail address is <peter.buchholz@udo.edu>.

DENNIS MÜLLER received the Diploma degree from the TU Dortmund in 2004, where he is currently working as junior scientist. His current research interests are focussed on efficient search heuristics for discrete event systems in combination with metamodels and ranking and selection. His e-mail address is <dennis.mueller@udo.edu>.