Journal of Experimental & Theoretical Artificial Intelligence
Vol. X, No. X, June 2016, 1–25

ARTICLE

Improving experimental methods on success rates in Evolutionary Computation

David F. Barrero[a]*, María D. R-Moreno[a] and David Camacho[b]

[a] *Universidad de Alcalá. Computer Engineering Department. Alcalá de Henares, Spain*;
[b] *Universidad Autónoma de Madrid. Escuela Politécnica Superior. Madrid, Spain*

(*May 2012*)

Due to the complexity of theoretical approaches in Evolutionary Computation (EC), research has being largely performed on experimental basis. One popular measure used by the EC community is the success rate (SR), which is used alone or as part of more complex measures such as Koza's computational effort in Genetic Programming. A common practice in EC is to report just a punctual estimation of the SR, without additional information about its associated uncertainty. We aim to motivate EC researchers to adopt more rigorous practices when working with SRs. In particular, we introduce the importance of correctly reporting this measure and highlight its binomial nature. Unfortunately, this fact is usually overlooked in the literature. Considering the binomiality of the SR opens the whole corpus of binomial statistics to EA research and practice. In particular, we focus on studying several methods to compute SR confidence intervals, the factors that determine their quality in terms of coverage probability and interval length. Due to its practical interest, we also briefly discuss the number of required runs to build confidence intervals with a certain quality, providing a sound method to set the number of runs, one of the most important experimental settings in EC. Evidence suggests that Wilson is, on average, a reliable and simple method to bound an estimation of SR with confidence intervals, while the standard method, which is quite popular because of its conceptual simplicity, should be avoided in any case. However, other methods can also be of interest under certain circumstances. We encourage to report the number of trials and successes, as well as the interval, to ease further comparability of the results.

Keywords: Experimentation; Success rate; Confidence intervals; Genetic Programming; Evolutionary Computation; Performance measures

1. Introduction

Experimentation is a basic tool in science. In many fields, when a theoretical approach is too complex, experimentation becomes the only way to manage that complexity. Although there are notable attempts to address Evolutionary Computation (EC) from a theoretical approach (Holland 1992, Poli et al. 2010), its research still relies heavily on empirical approaches (Bartz-Beielstein 2006, Birattari 2009, Bartz-Beielstein et al. 2010) where experimentation plays the central role. The target of any empirical study is to answer a research question by designing and running a set of experiments (Barr et al. 1995, Rardin and Uzsoy 2001). Some measures are gathered in these experiments and analyzed using statistical methods (Wineberg and Christensen 2010a,b, Derrac et al. 2011) to provide some light to the research question.

*Corresponding author. Email: david@aut.uah.es

Several measures have been used in EC research (Barreto et al. 2010). The selection of one measure instead of another one depends on the object of study, the algorithm and the goals of the experiment designer (Zitzler et al. 2000, Cruz et al. 2011). However, there are some common measures that are heavily used such as mean best fitness or mean average fitness (Eiben and Smith 2009). One of the most common ones is the success probability or success rate (SR). Due to the stochastic nature of Evolutionary Algorithms (EAs), when an EA is run it might, or might not, reach a solution that solves the problem it was designed for. SR is defined as the probability of an EA to find such solution, which is determined by imposing a success predicate, for instance, when an individual achieves a certain quality. In other words, it is not possible to use SR if there is not a criteria to identify a good enough solution (Eiben and Jelasity 2002, Bartz-Beielstein 2003). Some authors have proposed measures similar to SR, for instance, (Chen et al. 2012, 2010) proposed the solvable rate, which has a theoretical background.

SR should be used with caution. As Luke and Panait (Luke and Panait 2002) noticed, fitness might not be correlated with SR, and consequently it should not be used as a measure of the population quality. Nonetheless, finding literature that reports SR as a quality measure of the population is not too hard. In any case, SR provides an insight to the capabilities of the algorithm to find a solution. Sometimes SR is not the measure of interest, but rather it is part of a complex measure such as the *computational effort* (Koza 1992) in Genetic Programming (GP). In this case the accuracy of the complex measure depends on the quality of the estimation of SR.

One characteristic of SR as has been defined above is that it is defined as a scalar, but the value of the scalar cannot be known in the general case. It has to be estimated. Angeline (Angeline 1996) was the first person to observe this fact when working with computational effort, and suggested that a measure about a stochastic process should take into account its random nature. The same can be said of SR, a single point is not enough to characterize the stochastic nature of this metric, and some additional information about its statistical properties should also be reported, for instance, confidence intervals (CIs).

In the context of GP and computational effort, the first author that used CIs was Keijzer. He first noticed (Keijzer et al. 2001) that when the success probability was low, the computational effort is highly volatile. Additionally, the CI width associated to the estimator is nearly as large as the estimated value. This observation motivated Christensen (Christensen and Oppacher 2002) to study the reliability of Koza's computational effort. Walker (Walker et al. 2007a,b) studied several methods to construct computational effort CIs and concluded that Wilson's method is the most competitive one. Niehaus (Niehaus and Banzhaf 2003) studied systematically different factors that affect the precision of the measurement of computational effort. So, the reliability of computational effort has been well studied, however, these measures depends on the measure of SR, which is a more basic problem that, to the authors' knowledge, has not been addressed before in the context of EA.

Experimental methods and reporting practices in EC have been improving in the last years. Probably, the increasing interest and number of publications written about these topics is partially responsible of this fact. But despite this, there are still some weak reporting practices when dealing with the SR that could be improved. In this paper we try to call the attention of EC researchers and practitioners about the statistical nature of the SR and review some statistical tools that can be used to provide more rigorous empirical methods.

The contributions and significance of this paper are two-folded. First, we show that the number of success runs in an EA (and therefore the SR) can be modeled

using a binomial distribution, providing theoretical and empirical support as well in the context of EC. Second, we review and characterize CIs computation methods to report SR, validating theoretical models of their accuracy with data obtained from EAs. With this result, it is possible to determine the fiability of the SR estimation and exploit this knowledge to improve the experimental design in order to get a more reliable estimation of the SR, leading to more robust experimental practice. In particular, we show how to set the number of runs needed to get an estimation of the SR with a given maximum error. A preliminary version of this work was reported by Barrero et al. (Barrero et al. 2010). In this paper we provide additional empirical evidence, theoretical evidence, a much detailed description of the performance of binomial confidence intervals and a brief discussion about the number of runs that an EA researcher/practitioner should run to generate CIs with a certain quality.

The paper is structured as follows. First we introduce the basic problem of estimating a probability such as SR in EC. Then we study the statistical distribution that models SR and we continue with a description of several methods described by Statistics literature to calculate confidence intervals of a binomial distribution. Section 5 studies the performance of several CIs methods in relation to the number of samples of the experiment and SR. In section 6, we compare binomial CIs applied to some classical GP problems with CIs applied to theoretical binomial distribution. Section 7 briefly describes a method to estimate the number of runs needed to build CIs of a given quality. Finally, some conclusions are presented and future work lines are outlined.

2. Measuring a probability

From the perspective of SR an EA experiment is just a Bernoulli trial: an EA run is just an experiment whose outcome is a binary random variable X that can take two values, let's call them 'success' or 'failure'. From the EC perspective, the terms 'success' and 'failure' should be understood in a wide sense; the result of a classification rule that, given a population at a given time, classifies the algorithm execution as one of these two labels. For instance, in a regression problem to get a solution within a an error ϵ. The rest of our discussion relies in one assumption: the execution of the experiments are independent. This definition excludes some EAs whose restarts are not independent, for instance, an adaptative restarting strategy.

SR is defined as the probability $P(X = `success') = p$, and is described by the Bernoulli distribution. The most common case in EC is that p is unknown, which is precisely the parameter we want to estimate. The procedure to do it is well known, the EA execution is repeated n times yielding a number k of successes and $n - k$ failures, then a probability p is computed as $p = \frac{k}{n}$. Actually, we have described a Bernoulli process, a sequence $X_i, i = 1, 2, ..., n$ of random variables that are the outcome of a sequence of independent Bernoulli trials, and therefore they are described by a Bernoulli distribution. Despite its simplicity, a number of trivial and non-trivial issues arise when this experiment is analyzed in more detail.

Consider the next naïve experiment. We aim to empirically measure the probability of obtaining head when a coin is tossed, i.e., we have to estimate the value of p, $\hat{p}$. Of course, if the coin is equilibrated and the experiment is well implemented, that probability is $1/2$. But we want to study it empirically, so, we try the experiment 3 times and count the number of successes. In this case there are only four possible outcomes, $k \in \{0, 1, 2, 3\}$, and thus there are only four probabilities that can be estimated, $\hat{p} \in \{0/3, 1/3, 2/3, 3/3\}$, which are not close to $1/2$. We need to approximate p using the law or large numbers, the higher is the number of n, the

Table 1. Simulation of the estimation of the probability of getting heads when 1000 coins are tossed.

Experiment	Successes	$\hat{p}_i$
1	483	0.483
2	531	0.531
3	594	0.594
4	521	0.521
5	513	0.513
Total	2642	$\frac{2642}{5000} = 0.5284$

lower is the error.

Five simulations of the experiment described above with $n = 1000$ is shown in Table 1. It can be seen that even with a large number of experiments (1000 experiments), it is not possible to provide an exact estimation of p, each one of the five experiments yields different values of $\hat{p}$. Even if we average the probability of the five experiments -in this case $n = 5000$-, $\hat{p} = 0.5284$. Thus, providing a fixed value for p without any other information is a partial view of the estimator, and one hardly can do sound claims in base of this estimation (Angeline 1996). A reference is needed about how far or close $\hat{p}$ is expected to be from p. In order to obtain this information we previously have to study the statistical properties of the estimation of a probability, which is a well known problem in Statistics.

3. Distribution of SR

Regardless of the particular nature of the EA under study, the estimation of the success probability of an EA consists in running the experiment n times, use a heuristic to identify whether a particular run has been successful, and then count the number of successful runs in generation $i \in \mathrm{N}^+$, $k(i)$. Finally, the estimation is calculated as $\hat{p}(i) = k(i)/n$.

We are usually interested in $p(i)$ when the experiment has finished. So, for clarity and without loss of generality, if the algorithm has been run for G generations, we define SR as $SR = p(G)$. How $p(i)$ depends on time is a different topic that, for the specific case of GP, is addressed by Barrero et al. (Barrero et al. 2011).

If we assume that the experiments are independent, which is not a very restrictive assumption, measuring $\hat{p}$ is equivalent to estimating the number of successes k in n independent experiments. It is well known in statistics that the number k of successes is a random variable described by a binomial distribution, and thus the probability of getting k successes in n trials is given by:

$$Bin(k, n) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $p = k/n$, $C(n,k) = \frac{n!}{k!(n-k)!}$ and $k \in \{0, 1, 2, ..., n\}$.

It is straightforward to deduce the binomial distribution function. Given n experiments, there will be k successes and $n-k$ failures, if the success probability is p then, by definition, the probability of failure is $1-p$, the probability of getting k successes is p^k and the probability of getting $n-k$ failures is $(1-p)^{(n-k)}$. Therefore the probability of getting p^k and $n-k$ failures is $p^k(1-p)^{(n-k)}$. Moreover, the order in which successes appear is not a matter, they can appear in any combination of successes and failures, and there are $C(n,k)$ combinations, so we conclude that the

probability of getting k successes in n experiments when the success probability is p is given by $C(n,k)p^k(1-p)^{(n-k)}$, which is the binomial probability mass function.

So it can be deduced that the probability of getting k successes from n runs in an EA experiment is described by a binomial distribution. A binomial depends on two parameters, k and n, and thus the properties of the estimator of p is independent of the domain and the type of EA used. We can completely characterize the estimator if the number of runs and number of successes are known, which is the common situation in EC. More importantly, the properties of the estimator do not depend on the algorithm internals, following that this is of general application to any EA.

3.1 *Empirical study*

In order to get empirical evidence to support our claim we have selected four GP problems: artificial ant with the Santa Fe Trail, 6-multiplexer, 5-parity and a symbolic regression problem with no ephemeral random constants (ERCs). These are classical problems proposed by Koza (Koza 1992) and are widely used by GP literature. We have run the experiments with a standard tree-based GP algorithm using ECJ v18 and its default parameter settings[1]. The main parameters used in the GP executions are shown in Table 2.

Experimentation without any trick would require a huge number of runs, so, we used bootstrapping (Cohen 1995). A large number of $100,000$ runs were executed (this number is reduced to $5,000$ for the 5-parity problem due to computational resource limitations), and its result stored in a dataset. These datasets are used later to bootstrap $\hat{p}$ with different values of n. Since p is not known, we have approximated it with a precise estimation $\hat{p}_{best}$, which used the whole datasets. This precise estimation was used as the real p for comparison purposes. Table 3 shows $\hat{p}_{best}$, k, n and confidence intervals for $\alpha = 0.05$ and $\alpha = 0.01$ calculated with different methods introduced in the section section.

We first aim to compare graphically experimental results and the binomial distribution $Bin(\hat{p}_{best}, n)$. The procedure is the following one. First, we simulate $2,000$ experiments bootstrapping $2,000$ values of k. Each one of these values is calculated resampling with replacement n runs contained in the dataset and counting the number of successful runs. This procedure is repeated for each $n \in \{30, 50, 100, 250, 500, 1000\}$. After that, there will be $2,000$ simulated experiments with n runs each one, and a total number of $2,000$ values of k. These values of k were represented in an histogram using n as a factor.

The histograms of the four problems under study are depicted in Figure 1. The black points in the figure shows the binomial distribution $Bin(\hat{p}_{best}, n)$. It can be seen that the empirical number of successes follows closely the theoretical binomial distribution in the four problems and all values of n, even for the small ones, which is an evidence of binomiality. In addition, we can observe that as n increases, both the theoretical binomial and the empirical distribution of $\hat{p}$ are closer to the normal, which is consistent with the central limit theorem.

In order to provide additional graphic evidence to support our claim, Figure 2 shows a quantile plot of the four problems considered in this study. Quantile plots represent the number of successes of $2,000$ bootstrapped values of k with $n = 1000$, as was described above, against the theoretical number of successes obtained from the binomial distribution $Bin(\hat{p}_{best}, 1000)$. The plots show a linear relationship,

[1]Following the example of Daida (Daida et al. 1997) all the datasets, scripts and configuration files needed to repeat the experiments shown in this article are freely available on *http://atc1.aut.uah.es/~david/jetai2016/*. No modifications were done to the ECJ default configuration, with the exception of the Artificial Ant whose time steps were changed from 400 to 600.

Table 2. Tableau for the problems under study: artificial Ant with the Santa Fe Trail, 6-multiplexer, even 5-parity and symbolic regression without ERC.

Parameter	Artificial ant	6-multiplex.	5-parity	Regression
Population	500	500	4,000	500
Generations	50	50	50	50
Terminal Set	Left, Right, Move, IfFoodAhead	A0, A1, A2, D0, D1, D2, D3, D4, D5	D0, D1, D2, D3, D4	X
Function set	Progn2, Progn3, Progn4	And, Or, Not, If	And, Or, Nand, Nor	Add, Mul, Sub, Div, Sin, Cos, Exp, Log
Success predicate	$fitness = 0$	$fitness = 0$	$fitness = 0$	$fitness \leq 0.001$
Initial depth	5	5	5	5
Max. depth	17	17	17	17
Selection	Tournament (size=7)	Tournament (size=7)	Tournament (size=7)	Tournament (size=7)
Crossover	0.9	0.9	0.9	0.9
Reproduction	0.1	0.1	0.1	0.1
Elitism size	0	0	0	0
Terminals	0.1	0.1	0.1	0.1
Non terminals	0.9	0.9	0.9	0.9
Observations	Timesteps=600		Even parity	No ERC $y = x^4 + x^3 + x^2 + x$ $x \in [-1, 1]$

which suggests the correctness of the binomial assumption. Quantile plots for other values of n were depicted (not shown) with the same result.

We performed a fit-of-goodness study with a Pearson's χ^2 test. This test evaluates whether a set of samples comes from a population with a given distribution, a $Bin(\hat{p}_{best}, n)$ in this case. Since the measure of Pearson's χ^2 test is random due to the resampling, all the experiments have been repeated 100 times and the p-value averaged. The test was performed for $n \in \{15, 30, 50, 100, 250, 500, 1000\}$.

The results of the experiments for the four study cases under study can be found in in Table 4. It shows the mean p-value with its standard deviation and their difference. We set the rejection criteria of the null hypothesis (population comes from a $Bin(\hat{p}_{best}, n)$ distribution) as $p - value - sd < \alpha$, i.e., the difference between the p-value and its standard deviation was higher than a certain significance level, let us say $\alpha = 0.05$. Looking at the results shown in Table 4 we can observe that almost all the p-values are around 0.23, but it tends to get lower values when n is higher. Similarly, standard deviations get higher as n increases. Two facts can explain this behavior. First, the range of values that the random variable $Bin(\hat{p}_{best}, n)$ is wider when n is high, so it is logical that the dispersion of the p-value was proportional to n. Secondly, effect size might have a role in the explanation of the results. We should keep in mind that $\hat{p}_{best}$ is just an estimation of the real probability associated to the GP problem, this discrepancy is more apparent when n is high, so it is logical that the p-value got lower values.

Looking at Table 4 we only find evidence to reject null hypothesis in four cases,

Table 3. Best estimation of success probability ($\hat{p}_{best}$), number of successes (k) and number of runs (n) for the four GP problems under study, and their confidence interval (CI) calculated with the standard (Std), Agresti-Coull (AC) and Wilson (Wil) methods using confidence levels $\alpha = 0.05$ and $\alpha = 0.01$.

	Artificial ant	6-multiplexer	5-parity	Regression
$\hat{p}_{best}$	0.13168	0.95629	0.061	0.29462
k	13,168	95,629	305	29,462
n	100,000	100,000	5,000	100,000
CI $Std_{\alpha=0.05}$	[0.129584, 0.133776]	[0.955023, 0.957557]	[0.054366, 0.067634]	[0.291794, 0.297445]
CI $Std_{\alpha=0.01}$	[0.128926, 0.134434]	[0.954625, 0.957955]	[0.052282, 0.069718]	[0.290907, 0.298333]
CI $AC_{\alpha=0.05}$	[0.129598, 0.133790]	[0.955005, 0.957540]	[0.054688, 0.067985]	[0.291802, 0.297453]
CI $AC_{\alpha=0.01}$	[0.128950, 0.134460]	[0.954594, 0.957926]	[0.052831, 0.070333]	[0.290920, 0.298347]
CI $Wil_{\alpha=0.05}$	[0.129598, 0.133790]	[0.955005, 0.957540]	[0.054697, 0.067977]	[0.291802, 0.297453]
CI $Wil_{\alpha=0.01}$	[0.128950, 0.134459]	[0.954594, 0.957925]	[0.052850, 0.070314]	[0.290920, 0.298347]

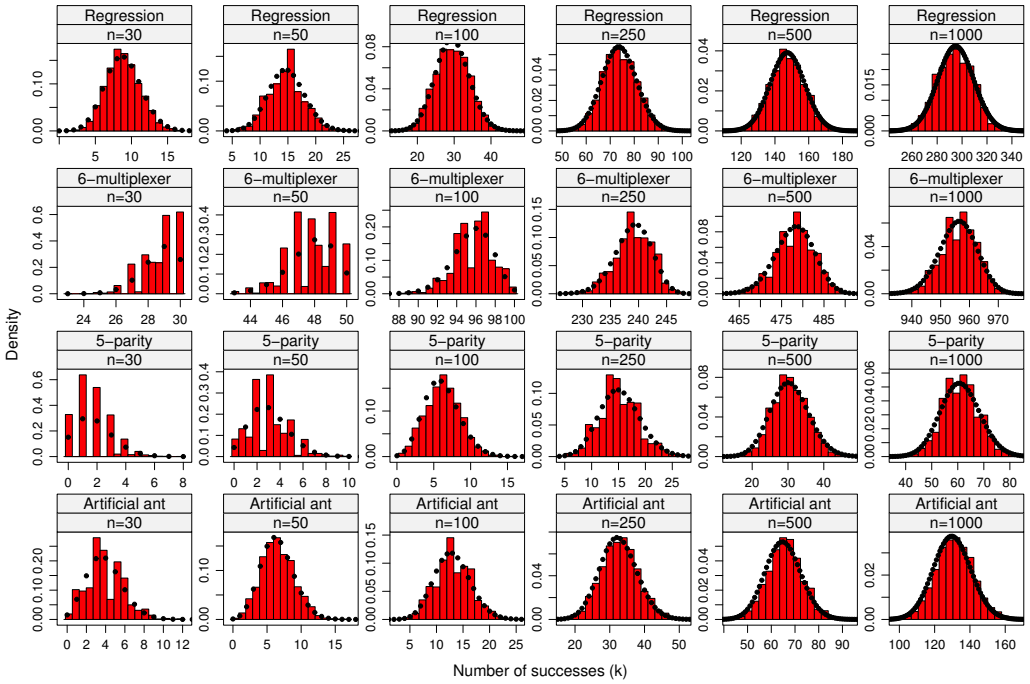


Figure 1. Histograms of $\hat{p}_{best}$ for different simulated sample sizes for Santa Fe Trail ($\hat{p}_{best} = 0.13168$), 5-parity ($\hat{p}_{best} = 0.061$), 6-multiplexer ($\hat{p}_{best} = 0.95629$) and regression ($\hat{p}_{best} = 0.29462$). The binomial density distribution Bin($\hat{p}_{best}, n$) is shown overlapped with black points.

all of them with high values of n. The results of the testing in the rest of the cases does not provide enough evidence to lead us to reject our initial hypothesis.

In conclusion, there are strong theoretical reasons to claim that success probability in EAs is a random variable that can be modeled with a binomial distribution. All the experiments carried out in four classical GP problems supports our claim for GP, histograms, quantile plots and Pearson's χ^2 test for fit support the bino-

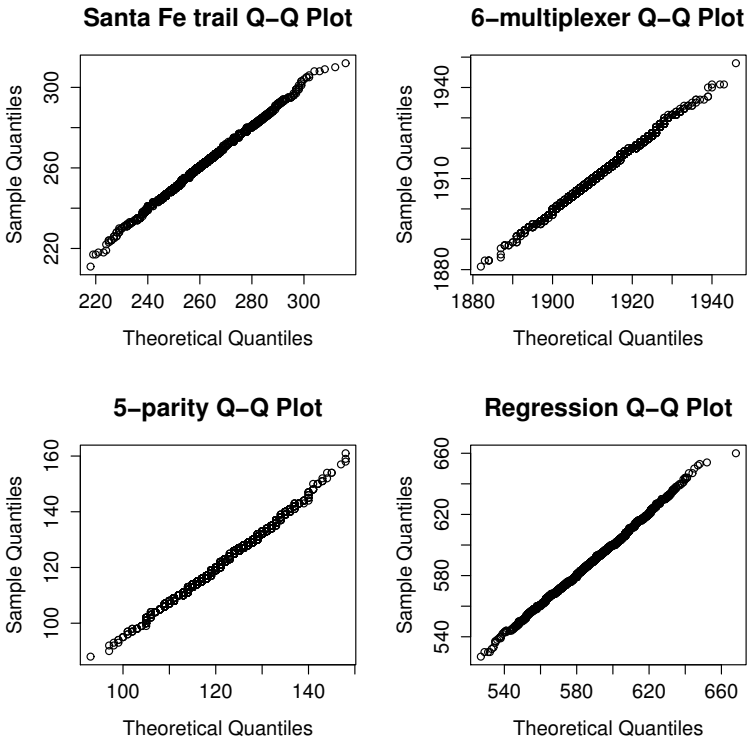


Figure 2. Quantile plot of four classical GP problems (Santa Fe Trail, 5-parity, 6-multiplexer and regression with no ERC) against a binomial distribution $Bin(\hat{p}_{best}, 1000)$ ($\hat{p}_{best}$ is, respectively, 0.13, 0.06, 0.96 and 0.29, see Table 3).

miality of the number of successful runs in an EA experiment. Therefore, it seems to be reasonable to assume binomiality until section 6, where this issue is resumed and additional evidence provided. One of the most notable consequences of the binomial nature of SR is that the statistical methods developed for binomial can be applied to SR in the context of EA. One of these methods is confidence intervals.

4. Confidence intervals

Using a binomial distribution to model the SR of EAs entails several benefits, one of them is that all the extense literature about binomials can be applied. In particular, the problem of estimating the SR of an EA can be generalized to the problem of estimating the parameters of a binomial distribution, which has been a subject of intense research in Statistics. Any estimator has a certain associated uncertainty, so, reporting only the value of the estimator provides only a part of the story. It is necessary to provide additional information about that uncertainty. A powerful tool to characterize it is CIs. Our goal is to get a basic understanding of how to use binomial CIs in the context of EAs, with a focus on their properties.

CIs for binomial distribution is a well studied problem due to its wide range of practical applications, so, it is not surprising that there are many methods to calculate binomial CIs (Brown et al. 2001), and rigorous comparisons have been published (Vollset 1993, Newcombe 1998, Brown et al. 2001, 2002, Ross 2003, Pires and Amado 2008). A binomial distribution is fully described by two parameters: the number of trials (n), and the number of successes (k). Alternately, the success probability p can also be used, which can be directly calculated from n and k simply as $p = k/n$. It is interesting from the perspective of EC because it decouples

Table 4. Pearson's χ^2 goodness-of-fit against a binomial distribution with $\alpha = 0.05$. 1,000 p-values were calculated, each one with 200 simulated experiments. Average p-values ($\overline{p-val}$), standard deviation of p-values ($\overline{sd}$) and their difference ($diff = \overline{p-val} - \overline{sd}$) are shown. Data that drives to reject the null hypothesis ($\overline{p-val} - \overline{sd} < 0.05$) is marked with bold letters.

	Santa Fe			6-Multiplexer		
N	$\overline{p-val}$	$\overline{sd}$	diff	$\overline{p-val}$	$\overline{sd}$	diff
15	0.2275	0.0032	0.2243	0.2206	0.0789	0.1417
30	0.2303	0.0243	0.206	0.2242	0.0060	0.2182
50	0.2374	0.0197	0.2177	0.2293	0.0053	0.224
100	0.2355	0.0535	0.182	0.2342	0.0285	0.2057
250	0.2397	0.1155	0.1242	0.2420	0.0249	0.2171
500	0.1885	0.1479	**0.0406**	0.2326	0.0756	0.157
1000	0.1279	0.1813	**-0.0534**	0.2109	0.1301	0.0808
	5-Parity			Regression		
N	$\overline{p-val}$	$\overline{sd}$	diff	$\overline{p-val}$	$\overline{sd}$	diff
15	0.2211	0.0042	0.2169	0.2331	0.0152	0.2179
30	0.2279	0.0041	0.2238	0.2425	0.0203	0.2222
50	0.2327	0.0048	0.2279	0.2453	0.0382	0.2071
100	0.2383	0.0125	0.2258	0.2316	0.0809	0.1507
250	0.2300	0.0631	0.1669	0.2132	0.1535	0.0597
500	0.2348	0.1044	0.1304	0.1303	0.1629	**-0.0326**
1000	0.2041	0.1006	0.1035	0.0407	0.1006	**-0.0599**

its study from the particular EA used. Only these two parameters are needed in order to fully describe the statistic properties of the CI, regardless of the internal dynamics of the EA and its particularities. One of the parameters in the binomial distribution, n, is usually known by the EA practitioner, while the SR, p, is usually unknown and thus it is the parameter that we are usually interested to estimate.

4.1 *Description of the CIs methods under study*

There are numerous binomial CIs calculation methods, and including all in this study would be unrealistic, so, we have selected those ones that we consider more representative due to its wide use or its presence in the literature. We have selected four methods: standard, 'exact', Agresti-Coull and Wilson. A brief introduction to these methods follows.

Standard interval. Also known as *asymptotic method*, *normal approximation* or *Wald interval*. It is the best known, oldest (Laplace 1812) and extended method, even the name represents how extensive the usage is. It is well known that a binomial $Bin(p, n)$, when np is large enough (usually $np > 30$), approximates a normal distribution $N(np, np(1-p))$ (see Figure 1). Therefore if the binomial approaches a normal, it is possible to generate intervals with the same method used with the normal distribution (Wald 1943). Although this method has been widely reported to suffer several flaws (Newcombe 1998, Brown et al. 2001, Walker et al. 2007a,b), it is widely used due to its simplicity and its presence in basic Statistics books. The standard interval is given by

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{1}$$

where $z_{\alpha/2}$ is the upper-$\alpha/2$ critical point from $N(0,1)$ and whose values can be found tabulated in statistical tables as well as statistical packages. One drawback that the standard interval presents is that this interval cannot be calculated when p is 0 or 1.

Clopper-Pearson or 'exact' interval. This interval is described as 'exact', with quotes, because it is deduced from the binomial distribution. Ironically, despite its name, its discrete nature makes this method unnecessarily conservative, and therefore far from being exact. The limits $[L, U]$ of the 'exact' interval (Clopper and Pearson 1934) are given by the solution to p of the equations $P(bin(n, p_U) \leq X) \geq \frac{\alpha}{2}$ and $P(bin(n, p_L) \geq X) \geq \frac{\alpha}{2}$, which yields the following equations:

$$\sum_{k=0}^{k} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \frac{\alpha}{2}$$
$$\sum_{k=x}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \frac{\alpha}{2}$$

The solution of these equations is not trivial and can be expressed using the beta distribution as follows.

$$\begin{aligned} L_{CP}(k) &= B(\alpha/2; k, n-k+1) \\ U_{CP}(k) &= B(1-\alpha/2; k+1, n-k) \end{aligned} \tag{2}$$

where $B(\alpha; a, b)$ stands for the α quantile of a $Beta(a, b)$ distribution. Sometimes the 'exact' interval is expressed as a function of the F-distribution, due to its relationship with the beta distribution:

$$L_{CP}(k) = \left[1 + \frac{n-k+1}{kF_{2k,2(n-k+1),1-\alpha/2}}\right]^{-1}$$
$$U_{CP}(k) = \left[1 + \frac{n-k}{(k+1)F_{2(k+1),2(n-k),\alpha/2}}\right]^{-1}$$

where $F_{a,b,c}$ represents the $1-c$ quantile from the F distribution with degrees of freedom a and b.

Agresti-Coull interval. Also known as *adjusted Wald*, a term introduced by the original paper of Agresti and Coull (Agresti and Coull 1998). This is a modification of the standard interval where some pseudo-observations are added to (1). In this way, instead of calculating the standard interval using n and p computed as $p = k/n$, Agresti-Coull uses $\tilde{p}$ and $\tilde{n}$ calculated as

$$\tilde{p} = \frac{(k + \frac{1}{2}z_{\alpha/2}^2)}{(n + z_{\alpha/2}^2)} \tag{3}$$

and

$$\tilde{n} = (n + z_{\alpha/2}^2) \tag{4}$$

then, the standard interval is calculated as in (1), but using $\tilde{p}$ and $\tilde{n}$ instead of p

and n,

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \tag{5}$$

It should be pointed out that for a common case where $\alpha = 0.05$, then $z^2_{\alpha/2} \approx 2$ and thus $\tilde{p}$ and $\tilde{n}$ can be simplified to $\tilde{p} = (k+2)/(n+4)$ and $\tilde{n} = (n+4)$. Consequently it is equivalent to adding two failures and two successes. In this way the probability remains unchanged, and the calculus of the CI is the same than the standard intervals, but their properties are significantly improved.

It is interesting to note that the center of the interval is not given by $\hat{p} = \frac{k}{n}$, as usual, but rather by $\hat{p} = (k + \frac{1}{2}z^2_{\alpha/2})(n + z^2_{\alpha/2})^{-1}$, which is not placed in the center of the interval. Nevertheless, as n is increased (and indirectly also k), the center of the interval tends to be closer to $\hat{p} = \frac{k}{n}$, so increasing the number of experiments generates more symmetric intervals.

Wilson interval. Also known as the *score* method. Wilson interval (Wilson 1927) is derived from a normal approximation as the solutions to the equations $(\tilde{p} - p_0)/\sqrt{p_0(1-\tilde{p})/n} = \pm z^2_{\alpha/2}$ which is given by

$$CI_W = \frac{k + \frac{1}{2}z^2_{\alpha/2}}{n + z^2_{\alpha/2}} \pm \frac{z^2_{\alpha/2}\sqrt{n}}{n + z^2_{\alpha/2}}\sqrt{p(1-p) + \frac{z^2_{\alpha/2}}{4n}} \tag{6}$$

The center of the Wilson interval has the same form as Agresti-Coull, so we can point out the same considerations about it. Actually, when $\alpha = 0.05$ Wilson intervals are quite similar to Agresti-Coull.

4.2 *Discussion about CI methods*

Many authors have studied the performance of CI methods using rigorous statistical approaches (Vollset 1993, Newcombe 1998, Brown et al. 2001, 2002, Ross 2003, Pires and Amado 2008). Brown (Brown et al. 2001) recommends, for small n (40 or less), Wilson or Jeffreys (a variation of Bayes intervals, not covered here) methods, while for large n values (more than 40) he recommends also Agresti-Coull. Similarly, Piegorsch (Piegorsch 2004) remarks that while Wilson and Jeffreys perform better when $n < 40$, the rest of methods are similar for higher values of n.

Some GP studies have been focused in the more specific problem of estimating the computational effort, (Niehaus and Banzhaf 2003, Walker et al. 2007a,b), all of them have noticed the poor performance of normal approximation, and recommend the use of Wilson. These studies apply several CIs methods to computational effort, nevertheless they use a pure experimental approach, without a theoretical or statistical justification to support the methods used. It is not considered that some of the CIs studied, such as Wilson, are supposed to be used with binomial distributions.

We aim to study the performance of the most significant binomial CI methods from a systematic, general and problem independent point of view, and how its performance depends on p and n. Once the behavior of the CIs was understood in terms of p and n, it is easy to extrapolate the results to a EC experimental context.

5. Confidence intervals performance

This section, inspired by (Brown et al. 2001), analyzes the performance of some CI methods. We are interested in showing the relationship between the two parameters of a binomial distribution and how they influence the performance of the CI. We use two related metrics to measure the performance of the CI methods, the coverage probability and the interval width.

On the one hand *coverage probability* (or CP) is defined as the probability of a CI to contain p, more formally, $CP = P(L \leq p \leq U)$. It is worth noting that increasing the CP of an interval is trivial, just by increasing its width. Furthermore, the coverage of the CI $[0, 1]$ is always 1 because p must be contained in that interval by definition. On the other hand CI width (or *CIW*) is defined as the difference between U and L, $CIW = U - L$. Of course, a tight interval is better than a wide interval, given that both have the same CP.

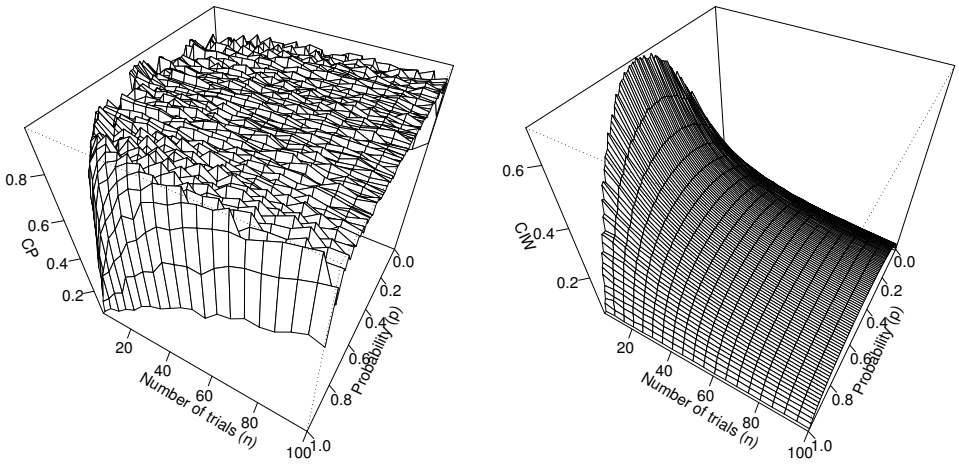
CP and CIW are closely related. There is a trade-off between CP and CIW, so CI methods have to find a balance between them. A good CI is not that one with a CP next to 1, but rather a tight interval with a CP close to the nominal coverage $1 - \alpha$, i.e., $P(L \leq p \leq U) \approx 1 - \alpha$. If an interval achieves a CP higher than the nominal one is at a cost of a wider interval. In terms of EA experimentation, such a conservative method would lead, for example, to a higher difficulty to detect significant differences between the SR of two algorithms. Understanding the properties of CP and CIW might lead to designing better experiments related to SR and composed measures such as computational effort.

5.1 *CIs performance overview*

The performance of an interval depends on the binomial parameters, but these parameters are not independent. In order to get an initial glimpse to this question we need a huge set of EA experiments as well as a strict control on the SR, which is quite difficult to achieve in real EAs. So, instead of running EAs, we have simulated them.

Since parameters n and p fully describe the binomial CI of an EA, we do not have to run the algorithm. From the SR point of view, the result of an EA experiment is binary. Therefore we have characterized pseudo-EA runs with a set of labels, "success" or "failure". For each probability $p \in \{i/2000 \forall i = 1, ..., 2000\}$, we generated a dataset with 100.000 labels "success" with probability p and "failure" with probability $1 - p$. Once the dataset was created, we took n pseudo-experiments $2,000$ times, with $n = 5, 6, ..., 100$, and calculated $\hat{p}_i$ for each set of experiments. In summary, we bootstrapped $\hat{p}$ using $2,000$ resamples for several combinations of p and n. The CI with $\alpha = 0.05$ was calculated using each method under study and each combination of p and n. Finally, CP and CIW were calculated. In this way we simulated the execution of EAs varying number of runs and probabilities, yielding a total of $95 \times 2,000 \times 2,000 = 380,000,000$ simulated EA runs.

The relationship between CIW, p and n can be seen in Figure 3 (b), where the CIW of standard CIs with $\alpha = 0.05$ is depicted. The shape of the figure is the same for the rest of the methods under study, so we only show the diagram of one method. CIW is symmetric for the plane $p = 0.5$ due to the fact that these methods are equivariant, their limits for $(n - k)/n$ are complements of those for k/n (Newcombe 1998). The plane $p = 0.5$ defines the symmetry of the figure as well as the maximum of CIW. CIs are wider when SR is close to 0.5, alternatively the closer is p to its boundaries 0 and 1, the tighter is the interval. It is explained by the constraints that the boundaries introduces to the calculation of the interval.



(a) (b)

Figure 3. (a) CP and (b) CIW for various sample sizes (n) and SR (p) for standard CI with confidence level $\alpha = 0.05$. Similar shape and fluctuations are found in the rest of the methods.

Looking at the behavior of CIW with n, we can observe that CIW is monotonically decreasing, whether the number of runs is increased there is additional information that is used to build tighter intervals. Of course, the price of such improvement is an increase of the number of runs and computational resources.

Figure 3 (a) depicts the coverage of standard CIs for several values of p and n. It can be seen that Figure 3 depicts a rather chaotic behavior, with many peaks and valleys without a clear pattern. We will show later that this behavior is not actually chaotic but rather the low resolution of Figure 4 which hides some phenomena very characteristic of binomial CIs. It will be analyzed in detail in the next section.

Some patterns can be found in Figure 3 (a). In particular, it is pretty clear that low coverage is associated to a low number of trials or a SR close to 0 and 1. This fact is also observed in the other methods and is intrinsic to the nature of the binomial distribution. However, there are quantitative differences among the methods. In the case of the standard interval the effects of low n values are dramatic because the normal approximation is no longer valid. This behavior is consistent with the one found in CIW: the wider is the CI, the less restrictive it is, and thus it is more likely that the real p was contained in the interval.

The rest of the methods present the same high level behavior described above, so all of them share some common properties which seem to be intrinsic to the problem, however, their performance differs significantly when analyzed in detail. It worths exploring this issue.

5.2 *Coverage of CIs*

An analysis of CP shows some remarkable facts. Figure 4 represents the coverage surface of the methods under study for n between 20 and 200 in steps of 1 and p takes $2,000$ values between 0 and 1. CP was calculated using R's function `binom::binom.coverage()`.

Figure 4 shows that the chaotic behavior of CP seen in Figure 3 (a) actually follows a pattern with symmetry in the axis $p = 0.5$. The low resolution and sampling noise of Figure 3 hid this pattern. As was previously seen, regardless of the used method, there are some areas with poor performance in terms of CP placed on the boundaries of p and low values of n. Coverage is particularly low in

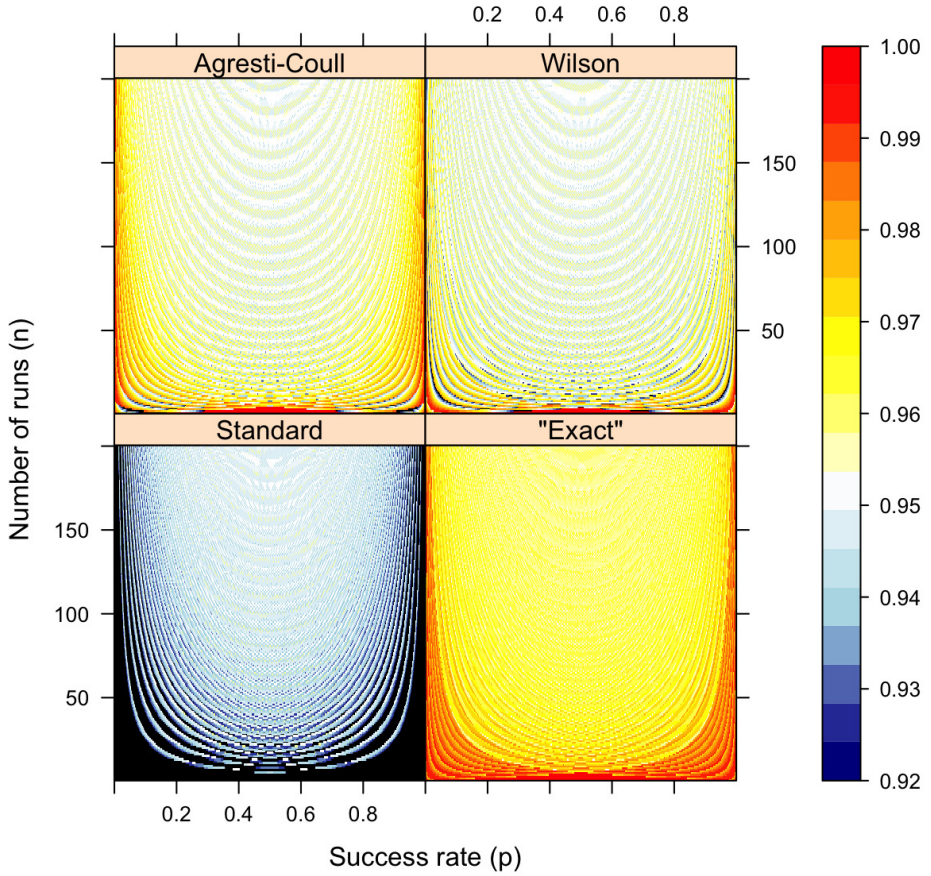


Figure 4. Dependence between coverage, n and p for CI methods under study: standard, 'exact', Agresti-Coull and Wilson, all calculated with $\alpha = 0.05$. X-axis represents success probability, p, while y-axis represents the number of runs, n. Coverage values lower than 0.92 have been represented in black, while coverage that equals the nominal value 0.95 is plotted in white.

the bottom corners of the graph, where both, n and p have negative influence on CP. This is a low coverage area, where simply there is not enough information to define reliable intervals. However, the coverage of the standard interval is dramatically poor in these corners in relation to the rest of the methods. Coverage of standard intervals achieves extremely low values when $n < 15$ and $p < 0.1$.

A new and striking phenomenon that was not observed previously in Figure 3 (a) is the presence of oscillations in the coverage regardless the CI method used. These oscillations are a well known phenomena (Agresti and Coull 1998, Brown et al. 2001) and they are generated by the discreteness of the binomial distribution. The magnitude of the oscillations has a great impact in the overall performance of the CI method. Oscillations appear in Figure 4 as waves whose magnitude is inversely proportional to the value of n. The magnitude of the oscillations also depends on p, and it is, like CP and CIW, symmetrical with respect $p = 0.5$.

Despite the fact that CP presents oscillations regardless the method being used, their magnitude changes. It is clear that, for instance, standard intervals coverage oscillates strongly when n is low in comparison to the rest of the methods. Wilson and 'exact' methods contains coverage oscillations less pronounced than those in the standard method.

It is important to verify how close is CP to the nominal coverage. This point can be appreciated more clearly in Figure 5, which shows several 'cuts' of the coverage surfaces at some values of n. In this way we obtain a detailed view of the

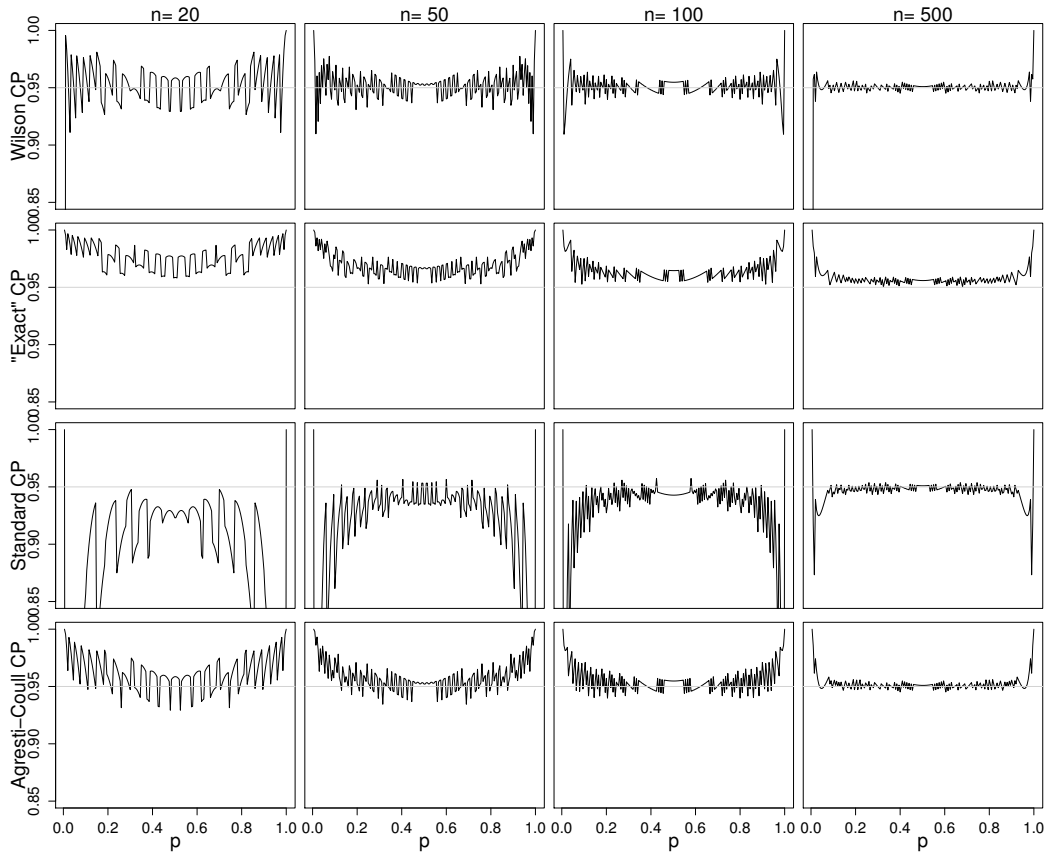


Figure 5. Comparison of CP for different CI methods (Wilson, 'exact', standardand Agresti-Coull) and number of samples (20, 50, 100 and 500). The nominal coverage, $\alpha = 0.95$, is represented with an horizontal grey line.

oscillations. An ideal CI method would generate intervals with a CP equal to the nominal one, nevertheless it is clear looking at Figure 5 that the real coverage is far from being equal to the nominal one, 0.95 in this case.

There are some common characteristics of the oscillations for all the methods under study. The magnitude of the oscillations and its shape is directly related to n. When n is large, for instance 500, CP gets a rather flat shape with small oscillations and CP gets pretty close to the nominal coverage. Looking at Figure 5 we conclude that, in the same line than the ones reported by (Brown et al. 2001), when n is large enough the differences of coverage properties of the CI methods under study are not significant.

Each method exhibits some particularities in the behavior of their coverage. 'Exact' intervals coverage is always higher than the nominal one. This conservatism produces wider intervals, as will be shown later. On the contrary, standard intervals coverage is lower than the nominal coverage, indeed when n is low, the coverage is much lower. Even when $n = 100$, which is a relatively high number of trials, its performance in the boundaries of p is poor compared to the other methods. Agresti-Coull interval does not exhibit such a clear behavior. It has areas where CP is higher than the nominal one, and other areas where CP is lower, however it tends to be more conservative in the boundaries of p. Finally, Wilson intervals show good coverage properties close to $1 - \alpha$ and it is neither conservative nor liberal.

It is worth comparing the exact CP with those obtained in the simulated executions of GP. Figure 6 depicts coverage diagrams obtained by the experiment described in the beginning of the section for $n = 20$. It can be seen that Figure 6

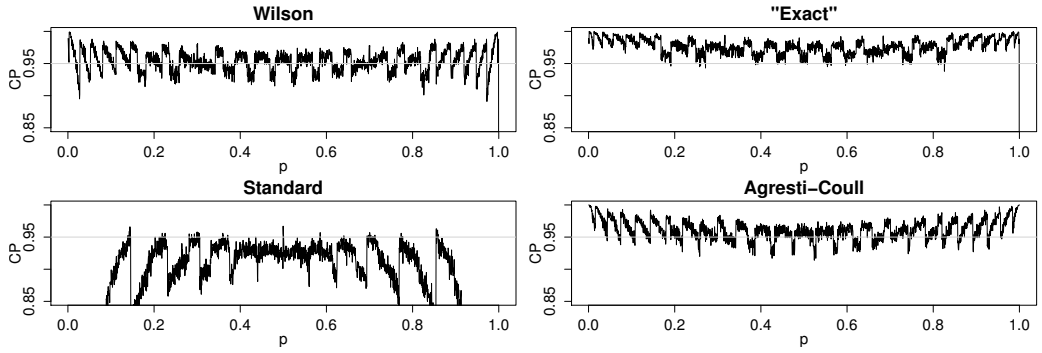


Figure 6. Empirical CP measured for simulated GP experiments when $n = 20$ and $\alpha = 0.05$. Coverage shows the same shape than the exact coverage shown in Figure 5 with some noise.

fits nicely with the analytical coverage represented in Figure 5, with the only exception of a high frequency noise produced by the sampling. Given that the CP diagram shown in Figure 6 is itself the estimation of a probability, the existence of this noise is now surprising. In any case, it seems clear that the shapes of both diagrams follow the same pattern, and thus the underlying probability distribution is likely to be the same, i.e., a binomial distribution, providing additional support to our hypothesis.

5.3 *Average CP*

Some characteristics of the coverage properties can be better viewed using average values of CP, as they are shown in Figure 7. It shows the CP averaged for $1,000$ values of p between 0 and 1 (top) and values of n between 5 and 49 (bottom). We used the `binom::binom.coverage()` function from the R `binom` package.

Figure 7 (top) shows how for low number of runs average coverage is degradated in all the methods, nevertheless, it does not affect equally to all. The standard method has very poor average performance when n is low. On the contrary, 'exact' method presents a rather high average CP for small number of runs, which is consistent with the conservative behavior of this method previously observed. Agresti-Coull method is also quite conservative, however less than the 'exact' method, its average CP is close to the nominal one. Finally, Wilson CIs have an outstanding performance with a low number of runs. When n is very low, around 5, its average CP is close to the one in the 'exact' method, nonetheless it dramatically decreases for higher number of runs, achieving an average CP very close to the nominal one.

It is interesting to observe the average CP when the number of runs is high. Figure 7 (top) shows that increasing the number of runs tends to reduce the difference among the methods, but not to the point of diluting all the differences. Even for a relative high number of runs ($n \approx 100$), the standard method has a disappointing performance with an average CP much lower than the nominal one. A glance to Figure 5 shows that the low average CP is due to its poor performance in the boundaries of p. In opposition to the standard method, the 'exact' method tends to generate conservative intervals even with high number of runs. Agresti-Coull and Wilson are the methods that are closest to the nominal coverage when n is high, with a small advantage to Wilson.

Figure 7 (bottom) adds a complementary perspective where CP has been averaged for values of n between 5 and 49. The standard method has very poor coverage properties, dramatically poor when p is close to 0 or 1. It should be pointed out that n values used to average is rather low, just where its performance is worse. The conservatism of the 'exact' method is evident observing the figure, this method

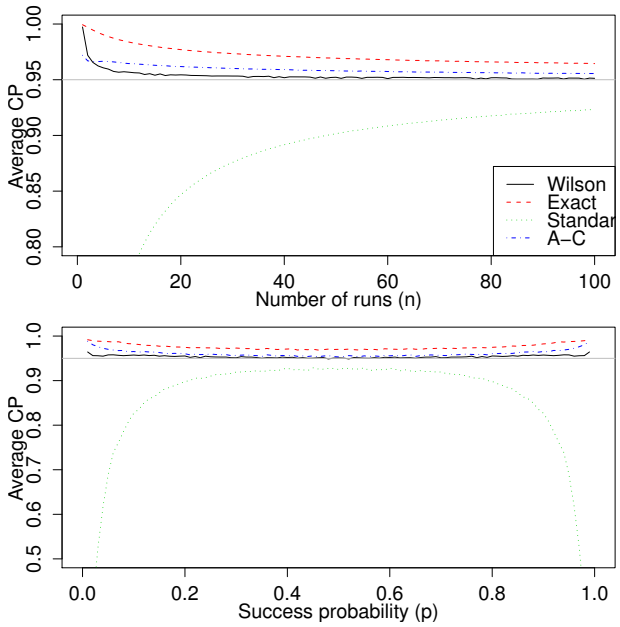


Figure 7. Comparison of average CP for different CI methods with fixed success probability p (top) and fixed number of runs n (bottom) (source: (Brown et al. 2001)).

generates the highest CP. This property makes it more difficult to find differences among algorithms, however it minimizes finding false differences, which might be of interest depending on the experimentation goals. Close to $p = 0.5$ all the methods have similar average CP, with the only exception of the 'exact' method, again, with a high CP in comparison with the rest of the methods. The method whose average CP is closest to the nominal coverage is clearly Wilson's method, achieving a quite flat average coverage plot, even in extreme values of p, where the average CP is slightly increased.

5.4 *Average CIW*

The overall picture of how CI methods perform should be completed looking at CIW. Unlike CP, CIW has not a random nature, given a certain n and p, the exact value of CIW can be determined. Average CIW values were calculated using n (Figure 8 top) and p (Figure 8 bottom) as independent variable. Many properties of the CI methods are equivalent or complementary to those observed for the average CP because of the close relationship between CP and CIW. The most notable differences among the CI methods are found when n is small and p is close to its boundaries, just like CP. Similarly, when n is large, average CIW tends to be rather similar in all the CI methods. The same happens when p is close to 0.5 between Wilson and Agresti-Coull.

Figure 8 (top) shows that there is a clear relationship between the average CIW and the number of trials: the smaller n is, the wider the interval is. Indeed it does not follow a linear relationship: when n is small adding few runs dramatically reduces CIW, but the effect of increasing n is less notable when n is greater, until a point where increasing n does not pay off due to the limited improvements in CIW.

CIW graphs explain some facts about coverage properties. The high CP found for the 'exact' method has its counterpart in CIW; high average CP is achieved at a cost of wider average intervals. This fact can be observed for almost all the values of p and n shown in Figure 8 (top and bottom). The normal approximation yields

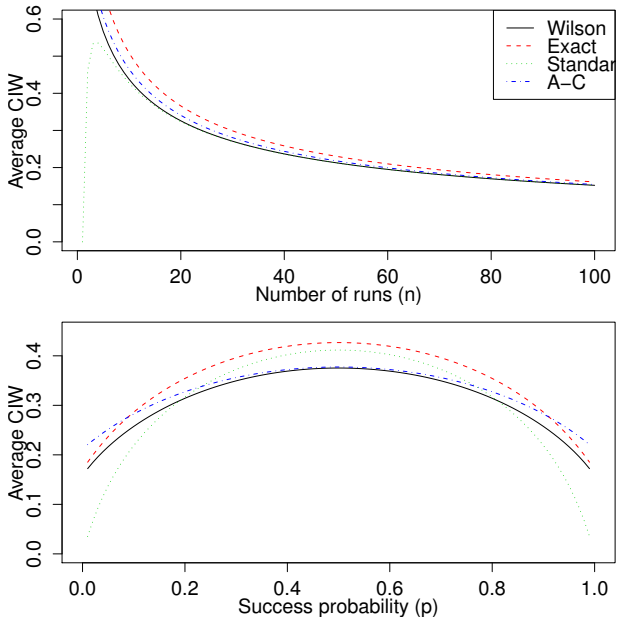


Figure 8. Comparison of average CIW for different CI methods with fixed success probability p (top) and fixed number of runs n (bottom) (source: (Brown et al. 2001)).

slightly wider CIs, except in case of low p values, just where CP is much worse. Wilson shows an excellent performance from the average CIW point of view with tight intervals.

5.5 *Discussion of the results*

Looking at the results shown in this section, we suggest not using the standard method in any case, its performance in terms of CP and CIW ranges from mediocre (when $np \gg 5$), to very poor ($np \leq 5$). The simplicity, availability and presence in the literature is a point to take into account in favor of Wilson's method. In any case, there is not a method with better CP and CIW in absolute terms. In average terms, Wilson seems to be a good election, but in order to be strict selecting the method with the best performance for an EA, we suggest to analyze first the area of the binomial parameter space in which SR would likely to be placed, and then look at CP and CIW of the methods in that area to select the method with better performance.

Another important subject to take into account when selecting a CI method, is the particularities of the experimentation. It might be important, for instance, being able to detect differences of SR between two EAs with a high level of confidence, avoiding type-II errors as much as possible. In this case using the 'exact' method might be interesting, at the price of making it harder to find these differences.

In any case, when the SR is very low, and it is not possible to run a large number of runs, the methods described in this paper are no longer recommended. When this situation is found, it is better to approximate the binomial $Bin(n, p)$ with a Poisson distribution with expectation $\lambda = np$ (Hodges and Le Cam 1960, Ehm 1991).

We have provided so far some theoretical and empirical evidences that support the binomiality of SR, as well as a glance to the performance of CIs. A natural question arises at this point: does the behavior of CI performance studied above also describe CI performance in real EA experiments?. It is clear that in case SR was binomial its CIs would have the same performance, suggesting an affirmative

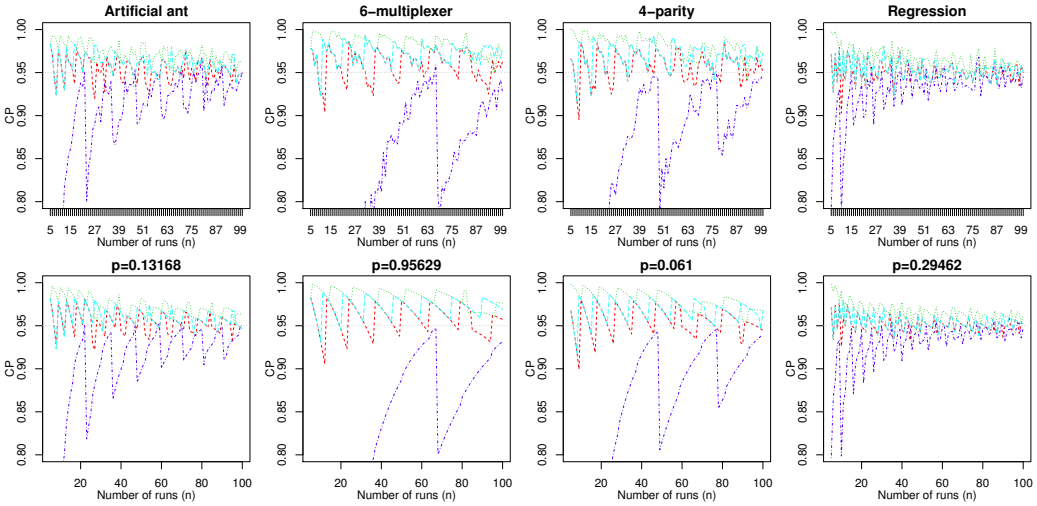


Figure 9. Empirical CP for the four domains (top) compared to CP of the binomial distribution (bottom) with the same SR ($\alpha = 0.05$).

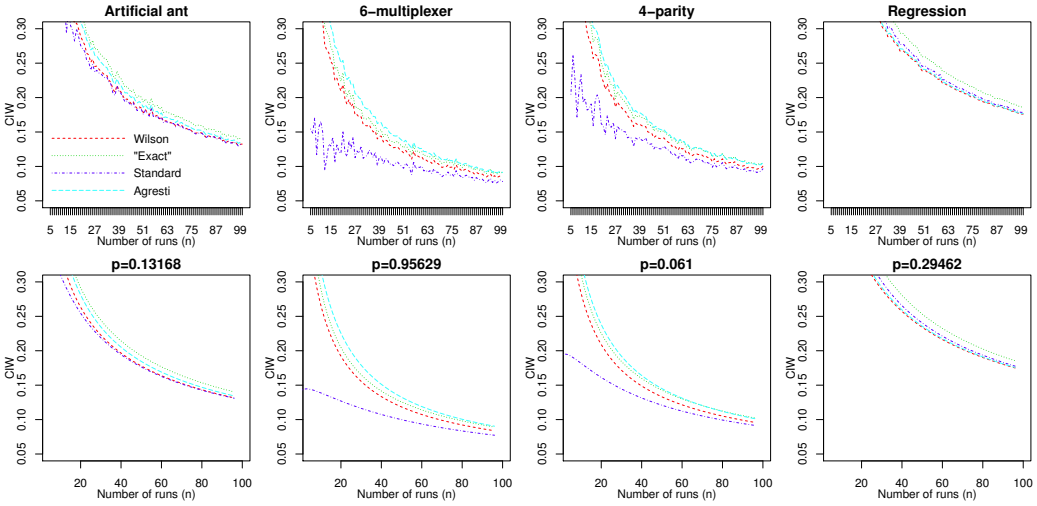


Figure 10. Experimental CIW for the four domains (top) compared to CIW of the binomial distribution (bottom) with the same SR ($\alpha = 0.05$).

answer, however a more direct evidence is actually desirable.

6. Binomial CIs in GP problems

We aim to test if the theoretical CP and CIW curves shown above are related to those ones obtained from real GP problems. So, CP and CIW curves have been generated using the experimental setup described in section 3.1 with its four GP problems: artificial ant with the Santa Fe trail, 6-multiplexer, even 6-parity and symbolic regression. CP and CIW were calculated using the same experimental procedure described in section 5.2, however, instead of using a dataset composed by pseudoexperiments, real GP experiments were used to generate $2,000$ intervals and calculate CP.

Figures 9 and 10 compare, respectively, CP and CIW of the four problems under study (first row) with the theoretical CP and CIW of a binomial $Bin(n, \hat{p}_{best})$ (second row). Theoretical and empirical plots match pretty well, with similar shapes and values. The only significant difference is the presence of statistical noise in

empirical data, which is logical. It is interesting to note that both, theoretical and empirical CP, contain almost the same oscilations, actually, empirical CP seems to be the theoretical CP contaminated with noise. This result supports our hypothesis that p fits actually a binomial distribution. An additional evidence supporting the binomiality of SR in GP is provided by Barrero et al. (Barrero et al. 2014), where the authors modeled the error in Koza's computational effort assuming the binomiality of SR. The theoretical model fitted quite well the error measured in the experiments.

7. Sample size of confidence intervals

We are interested in getting precise estimates of SR. Such an objective is rather simple to achieve: just by increasing the number of runs. However, the computational costs of running an EA might be high and so, increasing n without a well founded criteria might not be practical. It would be desirable using a well grounded mechanism to set *a priori* the number of runs needed to get intervals of the desired quality. Such a mechanism would, on the one hand, avoid wasting unnecessary computational resources running, and, on the other hand, provide a solid methodology to set the number of runs.

When someone computes the CI, he usually knows the number of algorithm executions and successes that have been achieved. However, we can state the problem from another point of view. We could make an initial rough estimation of p with a small number of runs, let us call it anticipated SR, or $\hat{p}_0$. Once $\hat{p}_0$ has been estimated, it is possible to set a certain CIW for the CI, and then obtain n from the equations of the CIs. In other words, we can try to set n using an initial estimate of p and a desired CIW. This approach was proposed by Piegorsch (Piegorsch 2004) to estimate the number of samples needed to obtain intervals with a certain CIW. In this section we summarize some of his results. There are other approaches, especially from the bayesian perspective, and a number of studies have been published (Sharma 1975, Rahme et al. 2000, M'Lan et al. 2008) addressing this topic from a statistical point of view.

Piegorsch (Piegorsch 2004) describes a method to compute b for the standard, Agresti-Coull, Wilson and Jeffreys methods. It is convenient to slightly change the notation. Instead of using the CIW to measure the interval width, in this section we follow Piegorsch adopting the half of the interval to represent the width. Let us denote the half of the interval as ε; it is clear that, by definition, $\varepsilon = CIW/2$.

We can apply the method previously described to the different CIs. For the Standard intervals we can state that the half of the interval is, using eq. (1), $\varepsilon = z_{\alpha/2}\sqrt{\hat{p}_0(1-\hat{p}_0)/n_s}$, solving that equation for the number of trials for the standard method (n_s) is straightforward, yielding

$$n_s = \frac{z^2_{\alpha/2}\hat{p}_0(1-\hat{p}_0)}{\epsilon^2} \tag{7}$$

We have used n_s instead of n to explicit the CI method that the equation applies to in an attempt to avoid notation ambiguity. The rest of the discussion follows this criteria.

The same procedure can be used for Agresti-Coull, the half of the interval in this case is given by eq. (5) as $z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$, where $\tilde{p}$ and $\tilde{n}$ are given by (3) and

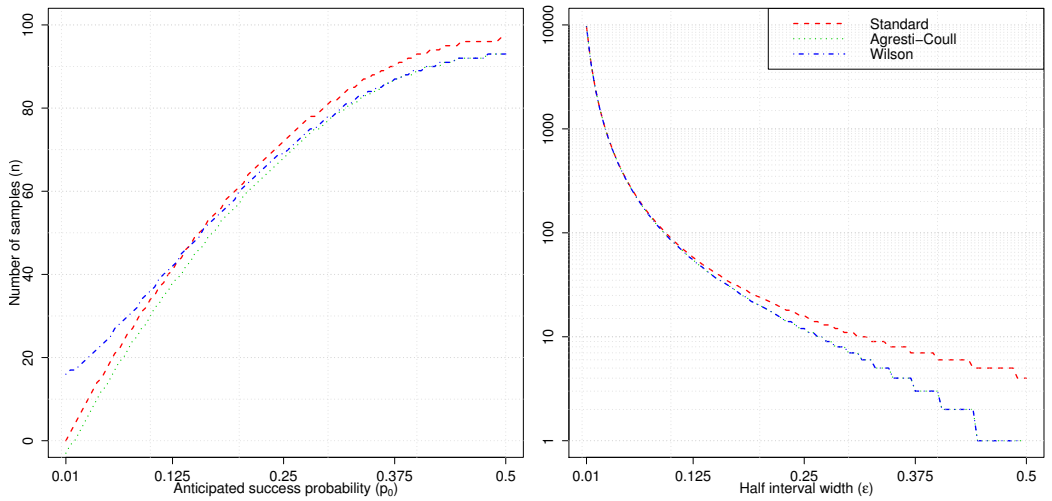


Figure 11. Sample size for standard, Agresti-Coull and Wilson CI methods for several anticipated success probabilities $\hat{p}_0$ with $\varepsilon = CIW/2 = 0.1$ (left) and different half interval widths when $\hat{p}_0 = 1/2$. Notice the logarithmic scale in the latter one. Unless for low values of $\hat{p}_0$, the sample size needed by Wilson is lower than for the other methods. Confidence level is set to $\alpha = 0.05$.

(4). Solving for the number of trials, n_{AC}, we obtain eq. (8),

$$n_{AC} = \frac{z^2_{\alpha/2}\hat{p}_0(1-\hat{p}_0)}{\epsilon^2} - z^2_{\alpha/2} = n_s - z^2_{\alpha/2} \tag{8}$$

It should be mentioned that Piegorsch does not recommend using (8) with less than 40 samples.

Similarly to the standard and Agresti-Coull intervals, Wilson sample size (n_W) is determined by (6) when it equals ε, yielding the following expression:

$$n_W = z^2_{\alpha/2}\frac{\hat{p}_0(1-\hat{p}_0) - 2\varepsilon^2 + \sqrt{\hat{p}^2(1-\hat{p}_0)^2 + 4\varepsilon^2(\hat{p}_0 - \frac{1}{2})^2}}{2\varepsilon^2} \tag{9}$$

In-depth discussion of the equations described above is out of the scope of this article, but it can be found in (Piegorsch 2004). However, it is worth a brief discussion to bring some interesting points to the EC community. Comparing (7) and (8) it is clear that since $z^2_{\alpha/2}$ is positive, Agresti-Coull always requires less samples than the standard method to achieve an interval of the same length. It is interesting to point out that when $\hat{p}_0 = 1/2$, n_W equals n_{AC}, a fact consistent with the relationship between Agresti-Coull and Wilson interval previously shown.

A major concern to calculate the sample size method described in this section is the anticipated success probability, $\hat{p}_0$, which is, itself, the problem we face when calculating binomial CIs. A conservative solution to deal with this problem without estimating $\hat{p}_0$ is to use the fact that CIs are widest when $p = 1/2$. If we consider the worst case scenario with $\hat{p}_0 = 1/2$, it is guaranteed that the resulting sample size will generate intervals, at least, of the desired ε. It could be better understood looking at Figure 11 (left), this figure represents the sample size as a function of $\hat{p}_0$ when $\varepsilon = 0.1$, i.e., an interval of the form $[\hat{p} - 0.1, \hat{p} + 0.1]$. The same behavior is observed for different values of ε. Figure 11 (left) clearly shows that, no matter which method is used, the value of $\hat{p}_0$ that originates the highest sample size is $1/2$, so it is a good conservative election when there is no information about its value.

Observing Figure 11 (left) with more detail is interesting. Agresti-Coull requires always less samples than the standard CI, and this method takes the same sample

size than Wilson next to $\hat{p}_0 = 1/2$. Figure 11 (left) could lead to mistakenly conclude that Agresti-Coull is the best choice to reduce the sample size if $\hat{p}_0 < 0.3$ and $\varepsilon = 0.1$. To get a complete picture, CP should also be considered. From the CIW point of view it is clear that Agresti-Coull would be the best choice, but looking at CP we can see that the small sample size is at a cost of a bad CP performance. So in this case Agresti-Coull needs fewer runs, but it generates less reliable intervals. In the low coverage region (small $\hat{p}_0$) Wilson has a slightly better CP performance, at a cost of a higher sample size, as can be seen in Figure 11 (left). In summary, there is a trade-off between CP and CIW.

When $\hat{p}_0$ is unknown, a conservative value $\hat{p}_0 = 1/2$ might be a good choice. Figure 11 (right) represents the sample size as function of ε in this situation. The sample size dramatically increases with the inverse of ε. This behavior is explained by the presence of ε in the denominator of (7), (8) and (9), and is logical, if we desire a tighter CI, we would need more information to build it, which is translated into more runs. As it was previously noticed, Agresti-Coull and Wilson intervals generate exactly the same sample size because when $p = 1/2$ both intervals are actually the same. Finally, it is interesting to notice that for wide intervals the sample size is as low as 1. When $\varepsilon \lesssim 0.5$ the interval takes the form $[L, U] \approx [0, 1]$, and therefore it almost covers all the possible values of p. In other words, the interval is so wide that it does not need much information to be constructed, yielding extremely low sample sizes.

In conclusion, the selection of the sample size has to take into account several criteria, some of them mutually exclusive, so a compromise is needed. Usually the goal is to obtain an interval with a certain CIW (or ε) with the lowest number of runs to save computational resources. Eqs. (7), (8) and (9) provide a mean to calculate the number of runs required to achieve a CI with a given CIW regardless of the associated CP. Figure 4 provides a mean to estimate the expected CP for the calculated sample size for $\alpha = 0.05$. In case that CP is not the expected one, it is necessary to increase the number of samples until CP achieves an acceptable value. So, n is determined by the maximum sample size between the desirable CP and CIW.

8. Conclusions

In this article we have provided theoretical and empirical evidences suggesting that SR in an EA can be modeled with a binomial distribution. Hence, the extensive literature about binomials can be applied, including CIs, determination of the sample size, hypothesis test for difference between proportions and so on. An important problem related to EC experimentation is the measurement of SR. Due to the binomial nature of SR, its estimation is the same problem that the estimation of the parameters of a binomial distribution. Unfortunately, a common practice in EC literature is to report SR without a measure of its uncertainty; the binomial nature of SR gives us sound statistical methods, such as CIs, to report it.

CIs are a statistical tool with a potential role when studying the performance of an EA. We have described some binomial CI methods with some of their main properties, drawing a picture useful to generate more robust experimental designs in EC. It was found that Wilson is the method that provides better average performance, even for low number of samples and SR next to the boundaries, nonetheless there is no method with the best CP for all the parameter space. Depending on the nature of the experimentation, other methods might be interesting due to their properties, such as Agresti-Coull or the 'exact' method when a conservative method is needed. In any case, experiments shown in this paper and related lit-

erature strongly discourage the use of the standard interval. Despite the method chosen, we encourage EA researchers to report n and k, as well as the interval, to ease further statistical manipulation and comparability of the results.

Finally, we reported some guidelines to select the number of runs to generate SR intervals with a certain expected performance in terms of CIW and CP. A straightforward method is summarized in Figure 11, which relates the number of runs with an upper bound of the uncertainty for several CI methods. Of course, SR is not usually the only measure that is taken from an EA, it only gives a partial view of the algorithm performance, that, with other measures, helps to understand the behavior of an EA. From this perspective, it would be interesting to study how the algorithm parametrization -for instance, the population size- affects the SR.

Acknowledgment

The work has been funded by UAH project 2015/00297/001, MINECO project EphemeCH TIN2014-56494-C4-4-P and JCLM project PEII-2014-015-A. The authors would like to thank Bonifacio Castaño Martín and Concepción Alonso Rodríguez for their valuable suggestions, and Mike de Pumpo for proof reading this article.

References

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

Angeline, P.J. (1996), "An investigation into the sensitivity of genetic programming to the frequency of leaf selection during subtree crossover," in *GECCO '96: Proceedings of the First Annual Conference on Genetic Programming*, Stanford, California, Cambridge, MA, USA: MIT Press, pp. 21–29.

Barr, R., Golden, B., Kelly, J., Resende, M., and Stewart, W. (1995), "Designing and reporting on computational experiments with heuristic methods," *Journal of Heuristics*, 1, 9–32, 10.1007/BF02430363.

Barrero, D.F., Camacho, D., and R-Moreno, M.D. (2010), "Confidence intervals of success rates in evolutionary computation," in *GECCO '10: Proceedings of the 12th annual conference on Genetic and evolutionary computation*, Portland, Oregon, USA, New York, NY, USA: ACM, pp. 975–976.

Barrero, D.F., Castaño, B., R-Moreno, M.D., and Camacho, D. (2014), "A study on Koza's performance measures," *Genetic Programming and Evolvable Machines*, 16(3), 327–349.

Barrero, D.F., Castaño, B., R-Moreno, M.D., and Camacho, D. (2011), "Statistical Distribution of Generation-to-Success in GP: Application to Model Accumulated Success Probability," in *Proceedings of the 14th European Conference on Genetic Programming, EuroGP 2011*, eds. S. Silva, J.A. Foster, M. Nicolau, M. Giacobini and P. Machado, 27-29 Apr., Vol. 6621 of *LNCS*, Turin, Italy: Springer-Verlag, pp. 155–166.

Barreto, A.M., Bernardino, H.S., and Barbosa, H.J. (2010), "Probabilistic performance profiles for the experimental evaluation of stochastic algorithms," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, Portland, Oregon, USA, GECCO '10, New York, NY, USA: ACM, pp. 751–758.

Bartz-Beielstein, T. (2003), "Tuning evolutionary algorithms: overview and comprenhensive introduction," Technical report 148/03, Universität Dortmund.

Bartz-Beielstein, T. (2006), *Experimental Research in Evolutionary Computation: The New Experimentalism*, Natural Computing Series, Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Bartz-Beielstein, T., Chiarandini, M., Paquete, L., and Preuss, M. (2010), *Experimental*

Methods for the Analysis of Optimization Algorithms, 1st ed., New York, NY, USA: Springer-Verlag New York, Inc.

Birattari, M. (2009), *Tuning metaheuristics: a machine learning perspective*, Vol. 197, Springer Verlag.

Brown, L.D., Cai, T.T., and Dasgupta, A. (2001), "Interval Estimation for a Binomial," *Statistical Science*, 16, 101–133.

Brown, L.D., Cai, T.T., and Dasgupta, A. (2002), "Confidence Intervals for a Binomial Proportion and Asymptotic Expansions," *Annals of Statistics*, 30(1), 160–201.

Chen, T., Tang, K., Chen, G., and Yao, X. (2010), "Analysis of Computational Time of Simple Estimation of Distribution Algorithms," *Evolutionary Computation, IEEE Transactions on*, 14(1), 1 –22.

Chen, T., Tang, K., Chen, G., and Yao, X. (2012), "A large population size can be unhelpful in evolutionary algorithms," *Theoretical Computer Science*, 436(0), 54 – 70.

Christensen, S., and Oppacher, F. (2002), "An Analysis of Koza's Computational Effort Statistic for Genetic Programming," in *EuroGP '02: Proceedings of the 5th European Conference on Genetic Programming*, London, UK: Springer-Verlag, pp. 182–191.

Clopper, C., and Pearson, S. (1934), "The use of confidence or fiducial limits illustrated in the case of the Binomial," *Biometrika*, 26, 404–413.

Cohen, P.R. (1995), *Empirical methods for artificial intelligence*, Cambridge, MA, USA: MIT Press.

Cruz, C., González, J., and Pelta, D. (2011), "Optimization in dynamic environments: a survey on problems, methods and measures," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 15, 1427–1448, 10.1007/s00500-010-0681-0.

Daida, J., Ross, S., Mcclain, J., Ampy, D., and Holczer, M. (1997), "Challenges with Verification, Repeatability, and Meaningful Comparisons in Genetic Programming," in *Genetic Programming 1997: Proceedings of the Second Annual Conference*, Morgan Kaufmann, pp. 64–69.

Derrac, J., García, S., Molina, D., and Herrera, F. (2011), "A Practical Tutorial on the Use of Nonparametric Statistical Tests as a Methodology for Comparing Evolutionary and Swarm Intelligence Algorithms.," *Swarm and Evolutionary Computation*, 1, 3–18.

Ehm, W. (1991), "Binomial approximation to the Poisson binomial distribution," *Statistics & Probability Letters*, 11(1), 7–16.

Eiben, A.E., and Jelasity, M. (2002), "A critical note on experimental research methodology in EC," in *In: Proceedings of the 2002 Congress on Evolutionary Computation (CEC 2002)*, IEEE, pp. 582–587.

Eiben, A.E., and Smith, J.E. (2009), "Working with Evolutionary Algorithms," *Introduction to Evolutionary Computing*, Springer-Verlag, pp. 241–258.

Hodges, J.L., and Le Cam, L. (1960), "The Poisson Approximation to the Poisson Binomial Distribution," *The Annals of Mathematical Statistics*, 31(3), 737–740.

Holland, J.H. (1992), *Adaptation in natural and artificial systems*, Cambridge, MA, USA: MIT Press.

Keijzer, M., Babovic, V., Ryan, C., O'Neill, M., and Cattolico, M. (2001), "Adaptive Logic Programming," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, eds. L. Spector, E.D. Goodman, A. Wu, W.B. Langdon, H.M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M.H. Garzon and E. Burke, 7-11 Jul., San Francisco, California, USA: Morgan Kaufmann, pp. 42–49.

Koza, J. (1992), *Genetic Programming: On the programming of Computers by Means of Natural Selection*, Cambrige, MA: MIT Press.

Laplace, P.S. (1812), *Théorie Analytique des probabilités*, Paris, France: Mme. Ve Courcier.

Luke, S., and Panait, L. (2002), "Is the Perfect the Enemy of the Good?," in *In Genetic and Evolutionary Computation Conference*, Morgan Kaufmann, pp. 820–828.

M'Lan, C.R., Lawrence, J., and Wolfson, D.B. (2008), "Bayesian Sample Size Determination for Binomial Proportions," *Bayesian Analysis*, 3(2), 269–296.

Newcombe, R.G. (1998), "Two-sided confidence intervals for the single proportion: comparison of seven methods," *Statistics in Medicine*, 17(8), 857–872.

Niehaus, J., and Banzhaf, W. (2003), "More on Computational Effort Statistics for Genetic Programming," in *Genetic Programming, Proceedings of EuroGP'2003*, eds. C. Ryan,

T. Soule, M. Keijzer, E. Tsang, R. Poli and E. Costa, 14-16 Apr., Vol. 2610 of *LNCS*, Essex: Springer-Verlag, pp. 164–172.

Piegorsch, W.W. (2004), "Sample sizes for improved binomial confidence intervals," *Computational Statistics & Data Analysis*, 46(2), 309–316.

Pires, A.M., and Amado, C.a. (2008), "Interval Estimators for a binomial proportion: comparison of twenty methods," *Statistical Journal*, 6(2), 165–197.

Poli, R., Vanneschi, L., Langdon, W., and McPhee, N. (2010), "Theoretical results in genetic programming: the next ten years?," *Genetic Programming and Evolvable Machines*, 11(3), 285–320–320.

Rahme, E., Joseph, L., and Gyorkos, T.W. (2000), "Bayesian sample size determination for estimating binomial parameters from data subject to misclassification," *Journal Of The Royal Statistical Society Series C*, 49(1), 119–128.

Rardin, R.L., and Uzsoy, R. (2001), "Experimental Evaluation of Heuristic Optimization Algorithms: A Tutorial," *Journal of Heuristics*, 7, 261–304.

Ross, T.D. (2003), "Accurate confidence intervals for binomial proportion and Poisson rate estimation," *Computers in Biology and Medicine*, 33(6), 509–531.

Sharma, R.S. (1975), "Bayes approach to interval estimation of a binomial parameter," *Annals of the Institute of Statistical Mathematics*, 27(1), 259–267.

Vollset, S.E. (1993), "Confidence intervals for a binomial proportion," *Statistics in Medicine*, 12(9), 809–827.

Wald, A. (1943), "Test of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54(3), 426–482.

Walker, M., Edwards, H., and Messom, C. (2007b), "The reliability of confidence intervals for computational effort comparisons," in *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, London, England, New York, NY, USA: ACM, pp. 1716–1723.

Walker, M., Edwards, H., and Messom, C.H. (2007a), "Confidence Intervals for Computational Effort Comparisons," in *EuroGP*, pp. 23–32.

Wilson, E.B. (1927), "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, (22), 309–316.

Wineberg, M., and Christensen, S. (2010a), "Statistical analysis for evolutionary computation: advanced techniques," in *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, Portland, Oregon, USA, GECCO '10, New York, NY, USA: ACM, pp. 2661–2682.

Wineberg, M., and Christensen, S. (2010b), "Statistical analysis for evolutionary computation: introduction," in *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, Portland, Oregon, USA, GECCO '10, New York, NY, USA: ACM, pp. 2413–2440.

Zitzler, E., Deb, K., and Thiele, L. (2000), "Comparison of Multiobjective Evolutionary Algorithms: Empirical Results," *Evol. Comput.*, 8, 173–195.