Confidence Intervals of Success Rates in Evolutionary Computation

David F. Barrero
University of Alcalá
Escuela Politécnica
Alcalá de Henares
Madrid, Spain
david@aut.uah.es

David Camacho
Autonomous University of Madrid
Computer Science Dept.
Madrid, Spain
david.camacho@uam.es

María D. R-Moreno
University of Alcalá
Escuela Politécnica
Alcalá de Henares
Madrid, Spain
mdolores@aut.uah.es

ABSTRACT

Success Rate (SR) is a statistic straightforward to use and interpret, however a number of non-trivial statistical issues arises when it is examinated in detail. We address some of those issues, providing evidence that suggests that SR follows a binomial density function, therefore its statistical properties are independent of the flavour of the Evolutionary Algorithm (EA) and its domain. It is fully described by the SR and the number of runs. Moreover, the binomial distribution is a well known statistical distribution with a large corpus of tools available that can be used in the context of EC research. One of those tools, confidence intervals (CIs), is studied.

Categories and Subject Descriptors

I.2 [**computing Methodologies**]: Artificial Intelligence

General Terms

Experimentation, measurement

1. INTRODUCTION

Regardless of the particular nature of the EA under study, the procedure to estimate its SR is similar. In a generational EA we run the algorithm n times and use heuristics to identify whether a particular run has been successful. Then we count the number of sucessful runs in generation i, $k(i)$. Finally, SR is estimated as $\hat{p}(i) = k(i)/n$. If we can assume that the experiments are independent, which indeed is not a very restrictive assumption, estimating p is equivalent to estimate the number of successes k in n independent experiments. It is well known in Statistics that k, under the described assumptions, is a random variable described by a binomial distribution which gives the probability of getting k successes in n trials. A binomial distribution is fully described by two parameters: the number of trials (n), and the number of successes (k). Alternately, the success probability $p = k/n$ can also be used, which can be directly calculated from n and k. It is interesting from an EC point of view because it decouples its study from the particular EA used, and thus a general domain-independent study can be performed.

GECCO'10, July 7–11, 2010, Portland, Oregon, USA.
ACM 978-1-4503-0072-8/10/07.

2. BINOMIALITY OF SR

In order to get empirical evidence to support our hypothesis, we have selected a classical GP problem: the Artificial Ant with the Santa Fe Trail. This problem has been widely used in the GP literature. We used the implementation made in ECJ v18 with the default configuration. Since the real SR of this problem is not known, we tried to obtain a reliable SR with a large number of $100,000$ runs, yielding $13,168$ successes. Therefore, our best estimation of SR in the Santa Fe problem is 0.13168. $2,000$ estimations of $\hat{p}$ were bootstrapped for each $n \in \{5, 10, 25, 50, 100, 200, 500, 1000\}$ and they were plotted in a Q-Q plot against a binomial distribution. Since the real SR of the problem is not known, the binomial used in the plot was calculated using the best estimation available, i.e., the estimation that used the entire dataset. It was seen that the binomial fits nicely to the data, supporting the binomiality assumption.

It seems reasonable to approximate the estimation of p to a binomial function. This fact means that the problem of estimating the SR can be generalized to the problem of estimating the parameters of a binomial distribution, which has been a subject of intense research in Statistics. There is a wide corpus in the literature [1] that can be applied to this problem, one of the most interesting ones are CIs.

3. CONFIDENCE INTERVALS

It is clear that providing just a puntual estimation of p is generally inconclusive. Then, it is necessary to provide additional information of how far the real p is expected to be from the estimated $\hat{p}$. This information can be provided using CIs. There are many methods to calculate CIs for a binomial distribution, and it is not possible to consider all of them in this study. Thus, we have selected the most relevant ones: Standard (also known as asymptotic, normal approximation or Wald), Clopper-Pearson or "exact", Agresti-Coull, Wilson and Bayes.

Some authors have studied the performance of CI methods using rigorous statistical approaches [1]. Brown recommends, for small n (40 or less), Wilson or Bayes. For larger n values (more than 40) he also recomends Agresti-Coull. Some GP related studies have been focused in the specific problem of estimating the *computational effort* in GP, [3, 2] and they did not considered the bayesian approach. All of them have noticed the poor performance of normal approximation, and recommended the use of Wilson to calculate CIs of Koza's computational efford.

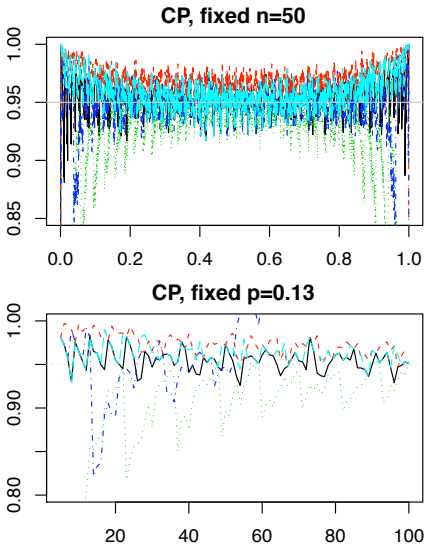

Figure 1: CP for different CI methods. Upper figure has a fixed $n = 50$ and the x-axis represents SR. Bottom figure has a fixed $p = 0.13$ and the x-axis represents the number of runs.

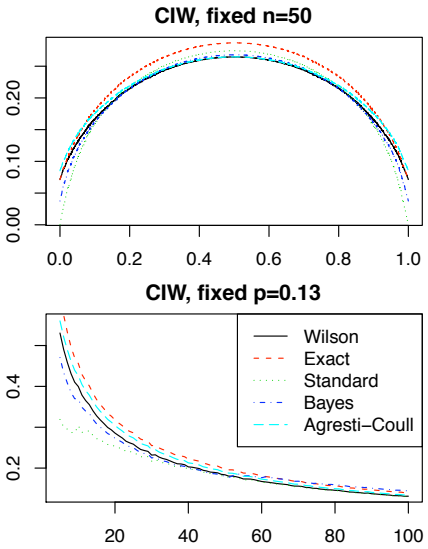

Figure 2: CIW for different CI methods. Upper figure has a fixed $n = 50$ and the x-axis represents SR. Bottom figure has a fixed $p = 0.13$ and the x-axis represents the number of runs.

4. CI PERFORMANCE

We use two metrics to measure the performance of the CI methods: the coverage and the interval width. On the one hand *Coverage Probability* (CP) is defined as the pobability of a CI to contain the real parameter p. On the other hand, CI width (or CIW) is defined as the difference between the upper and lower bounds.

Two experiments were run to characterize CI's performance. First, a dataset of $100,000$ simulated EA executions was build for each $p \in \{0.005 \times i, i = 0, ..., 2001\}$ to simulate a large EA number of runs. Then, for each point, $\hat{p}$ was estimated bootstrapping its value $1,000$ times, and CP as well as CIW were calculated. Results are represented in Figures 1 and 2 (top). Additionally, the same procedure was used with n, CP and CIW were calculated bootstrapping $\hat{p}$ $2,000$ times for each $n \in \{5, 6, ..., 100\}$. CP and CIW are represented in Figures 1 and 2 (bottom). Confidence level has been set to $\alpha = 0.95$.

Real CP might be quite different to the theoretical one. This fact is quite evident when p is close to 0 or 1 (Figure 1 top) or there are a low number of runs (Figure 1 bottom). The "exact" method is quite conservative, with real CP higher than α in any case, which lead to wider intervals. Coverage gets worse next to the boundaries of p, 0 and 1, regardless of the chosen method. An interesting phenomena can be found in Figure 1 (bottom). One may expect that increasing the number of runs would improve the performance of CIs, however the reality is other one. A small increase of n can lead to worse coverage properties. This fact is explained by the discreteness of the binomial distribution.

Figure 2 (top) shows that CIs are wider when p is close to 0.5, aditionally, CIs are tighter as the number of runs increases (Figure 2 bottom). It is interesting to observe the effects of increasing the number of runs: when n is small, adding few runs dramatically reduces CIW, but the effect of increasing n are less apparent when n is greater, until a point where increasing n does not pay off. Performance of Wilson and Bayes methods are quite similar, with a small advantage for Wilson for small n, while Bayes performs slightly better for p next to 0.5. However, the simplicity and availability of Wilson is a strong point in its favor.

It is interesting to verify if the behaviour described is similar to the one found in a real EA application. In order to provide some light to this issue, CP and CIW for the Artificial Ant were obtained using bootstrapping, and it was confirmed that the CP and CIW properties are very close to those shown in Figures 1 and 2.

5. CONCLUSIONS

In this article we have argumented the necessity to strengthen SR measurement in EC with more robust statistical tools. There are theoretical and empirical evidences that suggest that SR in an EA can be modelled with a binomial distribution. Statistical properties of the estimation are function of n and p, regardless of the algorithm's internals. The experiments carried out discourage the use of normal approximation and "exact" CI methods, while they showed that the best performance is achieved by Wilson.

6. REFERENCES

[1] L. D. Brown, T. T. Cai, and A. Dasgupta. Interval estimation for a binomial. *Statistical Science*, 16:101–133, 2001.

[2] S. Christensen and F. Oppacher. An analysis of koza's computational effort statistic for genetic programming. In *EuroGP '02*, pages 182–191, London, UK, 2002. Springer-Verlag.

[3] M. Walker, H. Edwards, and C. H. Messom. Confidence intervals for computational effort comparisons. In *EuroGP*, pages 23–32, 2007.