Confidence Intervals for Success Probability in Evolutionary Computation

David F. Barrero[1], David Camacho[2] and María D. R-Moreno[1]

[1]Universidad de Alcalá, Madrid, Spain
david,mdolores@aut.uah.es
[2]Universidad Autónoma de Madrid, Madrid, Spain
david.camacho@uam.es

Introduction

- Experimental research in EC needs performance measures
- Several performance measures have been defined in EC
- One common measure is success probability, or success rate (SR).
  - SR is defined as the proportion p between successfull (k) and total runs (n)
  - Some measures such as computational efford use SR [CO02, WEM07, BCD01]
- SR has a random nature, which leads our research question: How can SR be rigorously estimated?

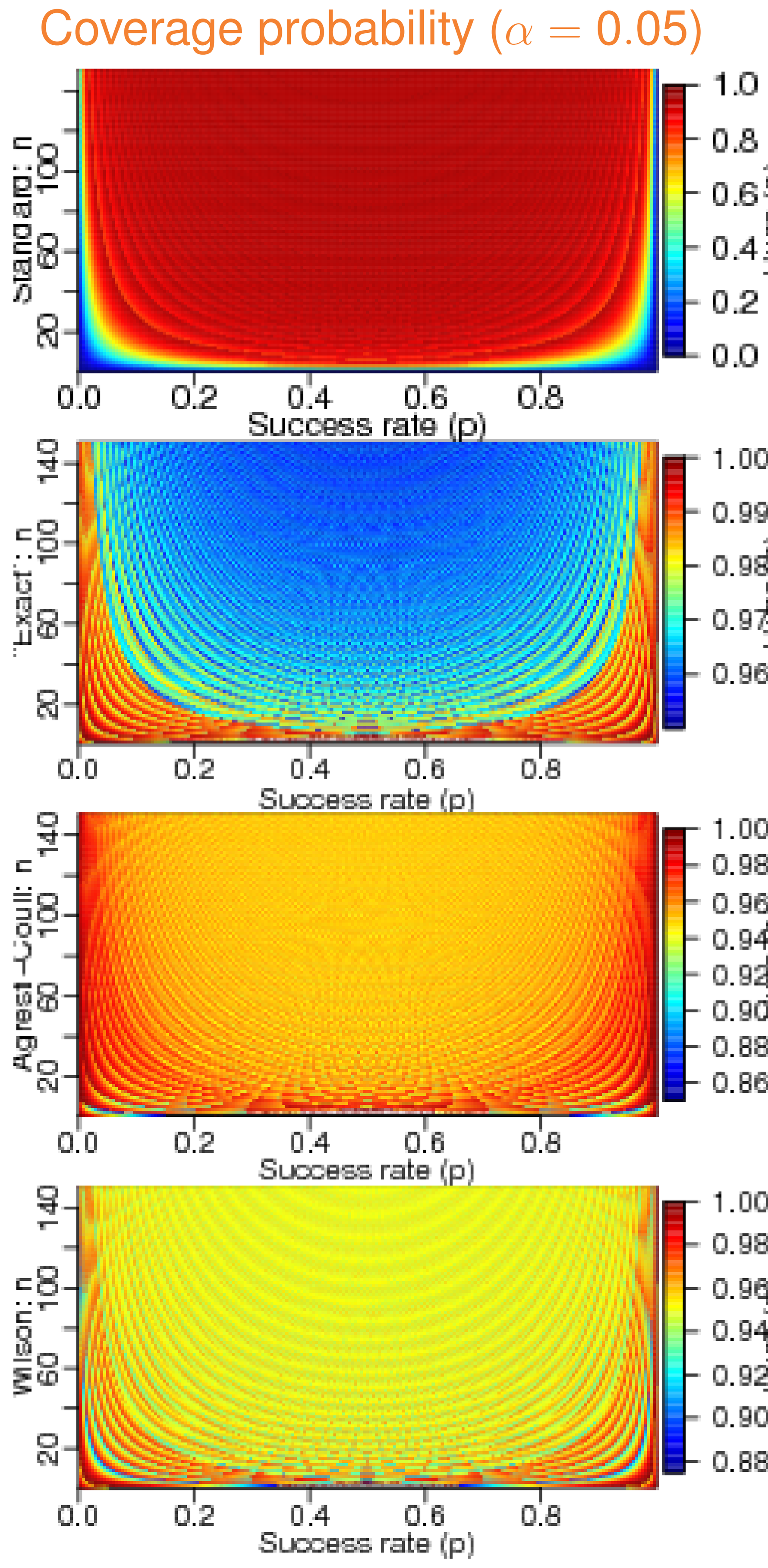
Confidence intervals

- A confidence interval of p is a range of values where p is likely to be contained
- An interval is defined by a lower and an upper bound $\hat{p} \in [L, U]$
- The confidence level α is defined as $P(L < p < U) = 1 - \alpha$
- Methods under study
  - Standard, Clopper-Pearson, Agresti-Coull and Wilson

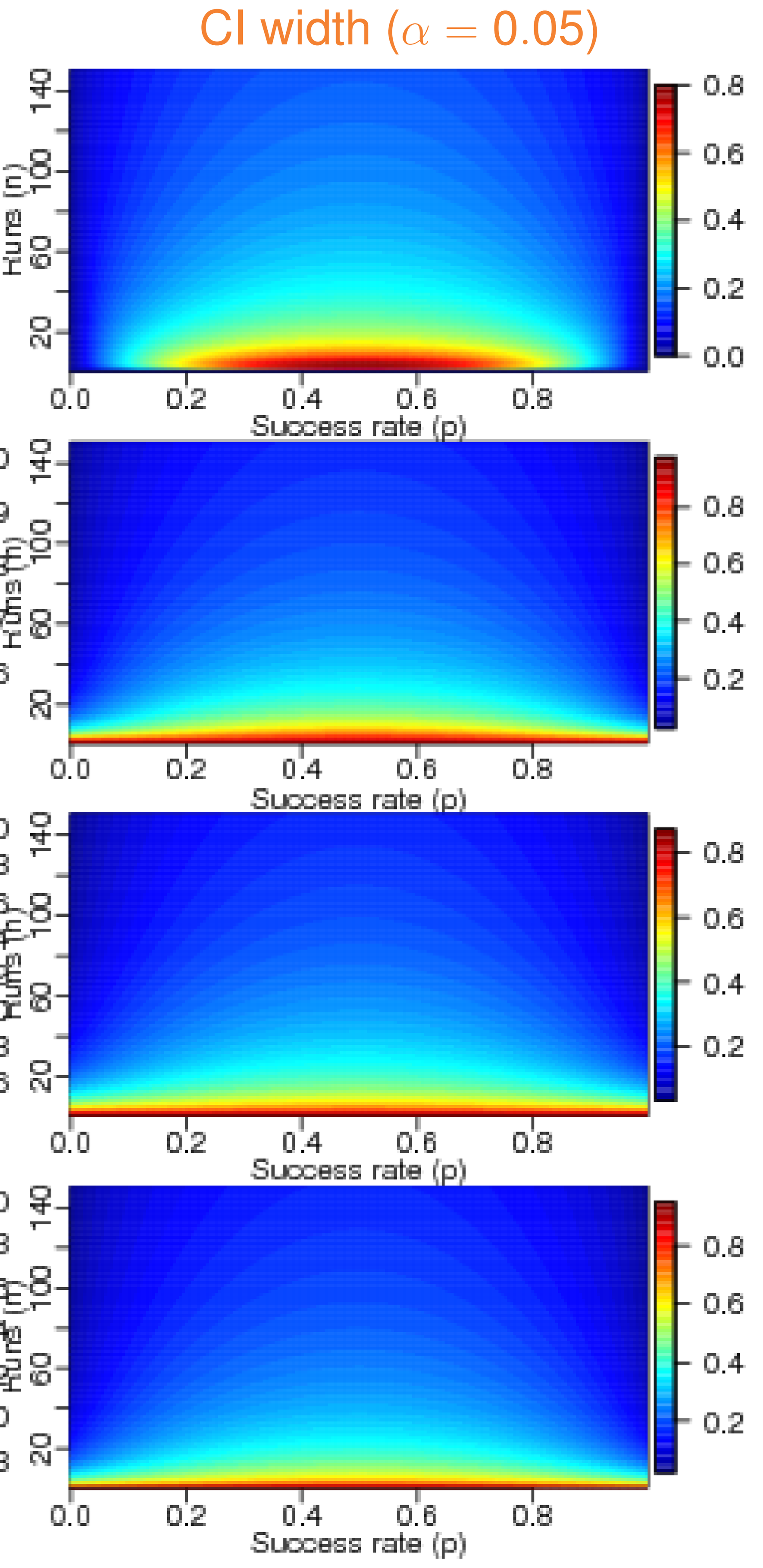
Confidence intervals performance

- We use two performance measures
  - Coverage probability (CP): Probability of the interval to contain the real parameter
  - Confidence interval width: Length of the interval
- Confidence intervals performance is a well known problem
- Best performance when:
  - $CP \approx 1 - \alpha$, $\forall\, p, n$
  - $U - L <<$

Coverage probability ($\alpha = 0.05$)



CI width ($\alpha = 0.05$)



- When $np <<$, coverage is poor
- Clopper-Pearson tends to be conservative
- Standard tends to be liberal
- Agresti-Coull is conservative next to 0 and 1
- Average Wilson coverage is close to α
- Wide intervals next to $p = 1/2$
- Clopper-Pearson creates wider intervals
- Standard creates tight intervals
- Agresti-Coull intervals are wide next to 0 and 1
- Wilson has an excelent ratio coverage/length

Which statistical model does SR follow?

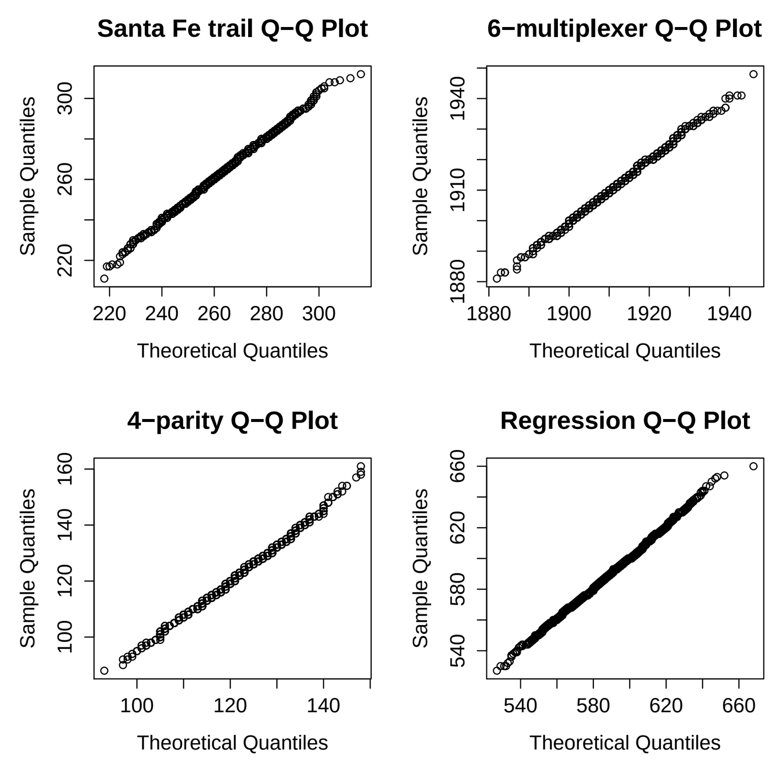
Theoretical approach

- SR is the probability of getting k successes and $n - k$ failures when an experiment is run n times
  - k successes: p^k
  - $(n - k)$ failures: $(1 - p)^{(n-k)}$
  - Successes might happen in any combination: $C(n, k)$
- Then $p(k, n) = C(n, k)p^k(1 - p)^{(n-k)} \Longrightarrow$ binomial distribution

Experimental approach

- Four classical GP problems were selected: Santa Fe trail, 6-multiplexer, even 4-parity, symbolic regression without ERC
- A large number of runs (100, 000) were executed
  - It gave a precise estimation of p
  - $\hat{p}$ was bootstrapped for several values of n
  - Two types of statistical tests: Q-Q plot and Pearson's χ^2 test for fit

Quantile plot



Pearson's χ^2 test for fit

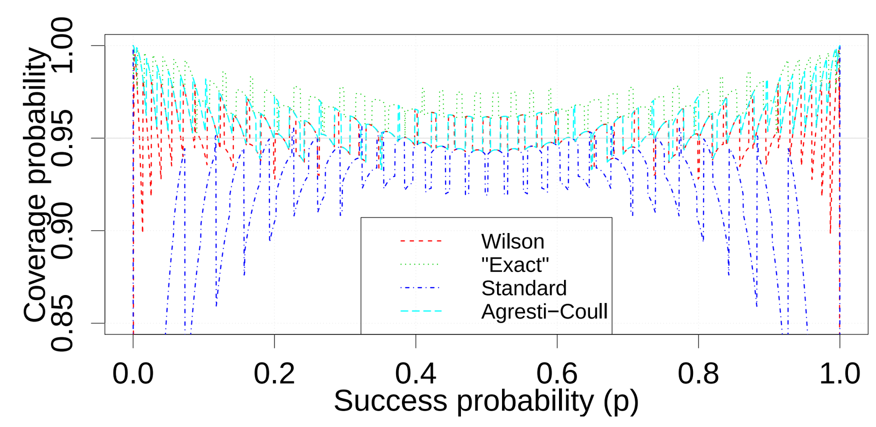
($\alpha = 0.05$)

GP Problem	p-value	sd
Santa Fe	0.2374	0.0197
6-Multiplexer	0.2293	0.0053
4-Parity	0.2327	0.0048
Regression	0.2453	0.0382

QQ plot and χ^2 test are shown with $n = 100$, different values of n yield similar results

Coverage oscilations

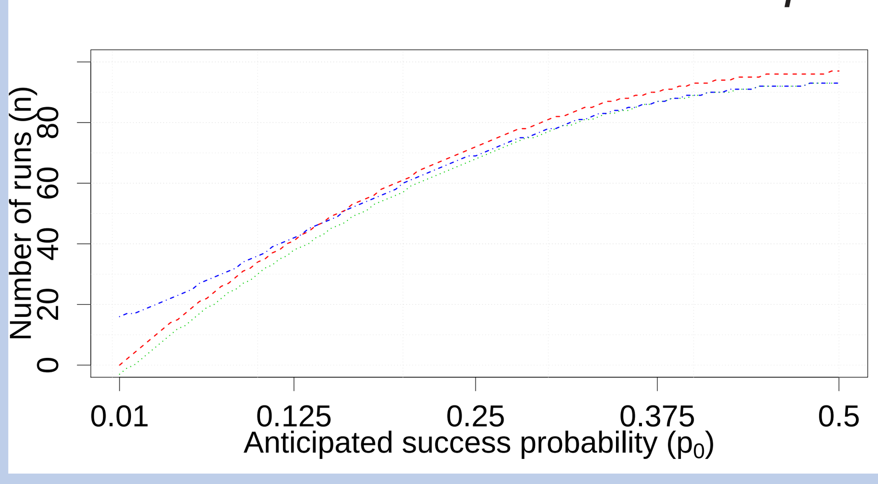
- Due to the discrete nature of binomials, coverage presents oscilations
- Increasing the number of runs decrease the magnitude of the oscilations



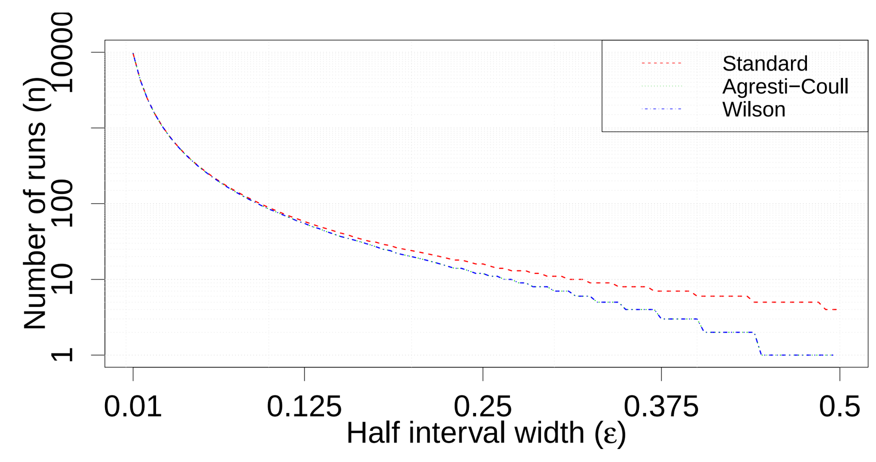
Determination of sample size

How many runs do we need?

- As function of a estimation of p



- As function of the interval width



Conclusions

- SR can be modelled as a binomial random variable
- Statistical methods used for binomials can also be used with SR
- Wilson is the most versatile confidence interval method
- In some conditions, Agresti-Coull and Clopper-Pearson methods might be a better choice

References

- L. D. Brown, T. T. Cai, and A. Dasgupta. Interval estimation for a binomial. *Statistical Science*, 16:101–133, 2001.
- S. Christensen and F. Oppacher. An analysis of koza's computational effort statistic for genetic programming. In *EuroGP '02*, pages 182–191, 2002.
- M. Walker, H. Edwards, and C. H. Messom. Confidence intervals for computational effort comparisons. In *EuroGP*, pages 23–32, 2007.

The Genetic and Evolutionaty Computation Conference (GECCO 2010) - Portland, Oregon