On Sampling Error in Evolutionary Algorithms

Dirk Schweim
schweim@uni-mainz.de
Johannes Gutenberg University
Mainz, Germany

David Wittenberg
wittenberg@uni-mainz.de
Johannes Gutenberg University
Mainz, Germany

Franz Rothlauf
rothlauf@uni-mainz.de
Johannes Gutenberg University
Mainz, Germany

ABSTRACT

The initial population in evolutionary algorithms (EAs) should form a representative sample of all possible solutions (the search space). While large populations accurately approximate the distribution of possible solutions, small populations tend to incorporate a sampling error. A low sampling error at initialization is necessary (but not sufficient) for a reliable search since a low sampling error reduces the overall random variations in a random sample. For this reason, we have recently presented a model to determine a minimum initial population size so that the sampling error is lower than a threshold, given a confidence level. Our model allows practitioners of, for example, genetic programming (GP) and other EA variants to estimate a reasonable initial population size.

CCS CONCEPTS

• **Computing methodologies** → **Model development and analysis**; • **Mathematics of computing** → *Stochastic processes*; *Mathematical analysis*; *Evolutionary algorithms*; • **Theory of computation** → *Design and analysis of algorithms.*

KEYWORDS

Sampling Error, Initial Supply, Evolutionary Algorithms, Building Blocks, Initial Population, *n*-Grams

ACM Reference Format:
Dirk Schweim, David Wittenberg, and Franz Rothlauf. 2021. On Sampling Error in Evolutionary Algorithms. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3462726

1 INTRODUCTION

In optimization, evaluating all solutions for a problem instance (complete enumeration) is often too difficult, expensive, or time-consuming. Therefore, population-based heuristic search methods like EAs start with a small sample taken from the set of all solutions and improve these solutions. When using a sample, there are usually differences between the *properties* of the statistical population and the information obtained from the sample. These differences are called *errors*. Non-systematic errors, describing random variations caused by observing only a subset of the statistical population are called *sampling errors*. The expected amount of sampling error can be reduced by using larger samples.

Sampling errors are a problem in evolutionary algorithms (EAs), leading to unreliable search results due to random variations (e.g., [1–3, 6]). Reducing the sampling error to a low amount is especially relevant for estimation of distribution algorithms (EDAs), where standard variation operators such as crossover and mutation are replaced by model building and sampling from the learned model. In EDAs, early sampling errors are learned by the model and, as a consequence, finding favorable solutions can be difficult. Therefore, we argue that an initial EA population should form a representative sample of the statistical population of possible solutions and that the sampling error in the initial EA population should be low.

To address the problem of sampling error in EA populations, we present a model that estimates the minimum size of an EA population that is required for a sampling error to be below a certain value that can be specified a priori by an EA user. Our suggested approach [8] consists of two steps:

(1) Identify relevant properties: the sampling error is measured with respect to a relevant property; differences of the frequencies in the sample in comparison to the statistical population of such a property define the size of the sampling error. Thus, a decision is needed on what is a relevant property.
(2) Determine a lower bound for the population size: the Cochran formula is used to estimate a lower bound for the size of the population. The model allows EA practitioners to estimate a minimum initial population size in such a way that the sampling error is lower than a threshold.

2 IDENTIFY RELEVANT PROPERTIES

Sampling error is a problem in evolutionary algorithms (EAs), leading to unreliable search results due to random variations. The problem has been discussed in the genetic algorithm (GA) literature. For example, Goldberg et al. [6] note that small initial populations in a genetic algorithm (GA) can be problematic when relevant *building blocks*[1] (BBs) are not represented by the sample. However, at the beginning of a search run, it is not known if a BB is relevant or not. Therefore, it is argued that the initial GA population should be large enough to ensure that at least one copy of each BB is present in the initial population.

Following the GA literature, papers about population sizing in GP focus on BB supply. In the context of GP, BBs describe relationships between nodes in GP parse trees. BBs in GP were usually defined as subtrees of a GP parse tree by many authors. For example, GP subtrees can be described by using *n-grams of ancestors*. An *n*-gram of ancestors in a GP parse tree is the sequence of the values

[1]BBs are defined by [5] as "short, low-order, and highly fit schemata". A schema is a similarity template describing a subset of solutions within a population with similarities at certain positions of the genotype [5].

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3462726

represented by a node i and its $n-1$ ancestor nodes on the same branch (parent, grandparent, greatgrandparent, etc.) [7]. Previous work finds that n-grams of ancestors represent relevant relationships between nodes of a GP parse tree. The difference between the expected and the observed frequencies of n-grams in a sample is the sampling error in initial GP populations.

3 DETERMINE A LOWER BOUND FOR THE POPULATION SIZE

The Cochran formula [4] is a standard method in statistics to estimate a minimum sample size N for a large statistical population. The Cochran formula needs an estimate of the relative frequency p of the property that is evaluated (e.g., the relative frequency of an n-gram of ancestors). In general, it is a problem to estimate p. In our case, the expected relative frequency can either be modeled (for an example, refer to [8]) or approximated. To approximate the frequencies, we suggest to initialize a large population without evaluating fitness values and measuring the frequency of the relevant property. We assume that p is normally distributed [4].

Furthermore, we need to choose an acceptable confidence level. For this, the Cochran formula uses z-scores of a normal distribution. For example, if a confidence level of 95% is chosen, the corresponding z-score is 1.96. Last, we define a desired margin r of the relative statistical error e, so that $e \leq r$, where e is the absolute difference between the expected frequency p and the measured frequency p' relative to p

$$e = \frac{|p' - p|}{p}.$$

The Cochran formula [4] is

$$N = \frac{z^2(1-p)}{r^2 p},$$

where p is the expected frequency, r is a margin of the relative error, and the confidence level is determined by a z-score.

Thus, if we take a sample of size N, the value of p will be in the interval

$$[p(1-r), p(1+r)]$$

with a probability equal to the confidence level. For example, we decide to use a confidence level of 95% ($z = 1.96$) and it is known that $p = 7\%$ of a statistical population have the respective property; the desired level of precision is $r = 10\%$. Then, using Eq. (3), we estimate $N = 5103.84$. As a result, if we take a random sample of size N, with a probability of 0.95 we measure p with $0.063 \leq p \leq 0.077$ ($P(0.063 \leq p \leq 0.077) = 0.95$).

The decision for a confidence level and a relative error is, to some extent, arbitrary [4]. Values widely used in the literature and also recommended by [4] are a confidence level of at least 95% ($z \geq 1.96$) and a relative error of not more than 5%. Estimated sample sizes calculated by using these values have a high precision and a high confidence. Given the expected frequency of a property (e.g., an n-gram), as the value for p, we can estimate the size of an EA population.

So far, we are only able to estimate the necessary EA population size for one statistical item, i.e., a specific value of the property (e.g., one specific n-gram). However, a relevant property typically can have several different values. For such a case, Cochran recommends to first identify the most important statistical items and afterwards estimate the sample size separately for each of these items. Then, Cochran's pragmatic recommendation is to simply select the largest estimate for a sample size of any of the items [4].

4 DISCUSSION AND CONCLUSION

We presented an application of Cochrans formula to determine a minimum size of an initial EA population, given a desired degree of sampling error and a confidence level. For the example of GP, we make our code publicly available in the form of a GP population size calculator, so that users of GP can calculate the lower bound for GP population sizes themselves.[2]

Of course, EA search is not only influenced by the initial population but also by other factors. In particular, we cannot guarantee a certain solution quality with our model since competing BBs or expressions for some specific problem domain are not considered. Thus, future studies need to extend our model, taking variation and selection into account (temporal models). Furthermore, combining our initialization model and an adaptive population size approach would be promising.

REFERENCES

[1] Bogdan Burlacu, Michael Affenzeller, Michael Kommenda, Gabriel Kronberger, and Stephan Winkler. 2018. Analysis of Schema Frequencies in Genetic Programming. In *Computer Aided Systems Theory – EUROCAST 2017*, Roberto Moreno-Díaz, Franz Pichler, and Alexis Quesada-Arencibia (Eds.). Springer International Publishing, Cham, 432–438.

[2] Bogdan Burlacu, Michael Affenzeller, Michael Kommenda, Gabriel Kronberger, and Stephan Winkler. 2018. Schema Analysis in Tree-Based Genetic Programming. In *Genetic Programming Theory and Practice XV*, Wolfgang Banzhaf, Randal S. Olson, William Tozier, and Rick Riolo (Eds.). Springer International Publishing, Cham, 17–37.

[3] Bogdan Burlacu, Michael Kommenda, and Michael Affenzeller. 2015. Building Blocks Identification Based on Subtree Sample Counts for Genetic Programming. In *Proceedings of the 2015 Asia-Pacific Conference on Computer Aided System Engineering (APCASE '15)*. IEEE Computer Society, 152–157.

[4] William Gemmell Cochran. 1977. *Sampling Techniques* (3 ed.). John Wiley, New York.

[5] David E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, Inc., Boston, MA.

[6] David E. Goldberg and Philip Segrest. 1987. Finite Markov Chain Analysis of Genetic Algorithms. In *Proceedings of the Second International Conference on Genetic Algorithms and Their Application*. L. Erlbaum Associates Inc., Hillsdale, NJ, 1–8. http://dl.acm.org/citation.cfm?id=42512.42513

[7] Erik Hemberg, Kalyan Veeramachaneni, James McDermott, Constantin Berzan, and Una-May O'Reilly. 2012. An Investigation of Local Patterns for Estimation of Distribution Genetic Programming. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation (GECCO '12)*. ACM, New York, NY, 767–774.

[8] Dirk Schweim, David Wittenberg, and Franz Rothlauf. 2021. On sampling error in genetic programming. *Natural Computing* (2021).

[2] The calculator can be found at https://gitlab.rlp.net/schweim/sampling-error-in-GP/