An Improved Small-Sample Statistical Test for Comparing the Success Rates of Evolutionary Algorithms

Bo Yuan
Division of Informatics
Graduate School at Shenzhen
Tsinghua University
Shenzhen 518055, P.R. China
+86-755-26036067
yuanb@sz.tsinghua.edu.cn

Marcus Gallagher
School of Information Technology and Electrical Engineering
The University of Queensland
QLD 4072, Australia
+61-7-33656197
marcusg@itee.uq.edu.au

ABSTRACT
Success rate is a commonly adopted performance criterion for evaluating Evolutionary Algorithms due to their inherent randomness. However, the classical large-sample binomial test based on normal distributions is only valid with a relatively large number of trials, which may not be feasible when experimental studies are very time consuming or expensive. In this paper, we give an alternative statistical test, which is suitable for situations where results from only a small number of trials are available.

Categories and Subject Descriptors
G.3 [**Probability and Statistics**]: Experimental Design.

General Terms
Algorithms, Measurement, Performance, Experimentation

Keywords: Performance Metric, Success Rate, Hypothesis Test

1. INTRODUCTION
Evolutionary Algorithms (EAs) and other metaheuristic algorithms are often relatively simple in terms of implementation. However, it is still generally impossible to investigate theoretically the behavior of these algorithms unless significant assumptions are made on either the algorithms or the problems to be solved. Consequently, EAs are primarily evaluated and compared empirically, using artificial benchmark problems or problems derived from real-world domains [1-3].

Success rate is one of the commonly used performance measures for assessing the performance of EAs, which is defined as the proportion of trials in which the termination criterion was met (e.g., the desired fitness level was obtained). From the hypothesis testing point of view, comparing the success rates of two algorithms is equal to testing the difference between two population proportions (the results of each algorithm are regarded as samples from an underlying binominal distribution with unknown proportion). According to the Central Limit Theorem, the sample proportion follows approximately a normal distribution when the number of samples is large.

However, in some cases, each trial may take a significant amount of time to finish and/or the objective function is very expensive to evaluate. In such cases, it is likely that only a few trials can be allowed and the hypothesis test on success rates has to be performed based on a small number of sample points. Unfortunately, the classical large-sample binominal test is no longer valid and in this paper we will focus on small-sample inferences concerning the difference between population proportions for the purpose of comparing the success rates of EAs.

GECCO'09, July 8–12, 2009, Montréal Québec, Canada.
ACM 978-1-60558-325-9/09/07.

2. LARGE-SAMPLE HYPOTHESIS TEST
The probability density function of a binominal distribution is:

$$b(x,n,p)=\binom{n}{x}p^x(1-p)^{n-x}=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \quad (1)$$

In Eq. 1, x is the number of successes among n trials where the probability of success is p (the proportion). The expected value and the variance of x are np and $np(1-p)$ respectively.

The unbiased estimator for the difference in population proportions (p_1-p_2) is the difference in the corresponding sample proportions (x_1/m-x_2/n) where x_1 and x_2 are the number of successes in the two samples respectively. In this paper, it is always assumed that the samples are of equal sizes (m=n).

The variance of the sample proportion is:

$$V(\hat{p})=V(\frac{x}{n})=\frac{1}{n^2}\cdot V(x)=\frac{p(1-p)}{n} \quad (2)$$

The variance of the difference of the two sample proportions is:

$$V(\hat{p}_1-\hat{p}_2)=V(\frac{x_1}{n})+V(\frac{x_2}{n})=\frac{p_1(1-p_1)+p_2(1-p_2)}{n} \quad (3)$$

When n is large, the sample proportions as well as their difference are approximately normally distributed. In this paper, the null hypothesis is H_0: p_1-p_2=0 and the corresponding test statistic is:

$$Z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{2\hat{p}(1-\hat{p})}{n}}},\qquad \hat{p}=\frac{x_1+x_2}{2n} \quad (4)$$

The alternative hypothesis is H_a: p_1-p_2>0 (one-tailed) and the rejection region is $Z \geq Z_\alpha$ where Z_α is determined by the predefined significance level α.

3. SMALL-SAMPLE HYPOTHESIS TEST

The major advantage of the test procedure described in Section 2 is that the test statistic can be computed very efficiently and the critical values (for a range of significance levels) are available in many statistics textbooks.

However, the Central Limit Theorem holds only for large sample sizes and alternative statistical techniques are required when the number of available trials is small. In this section, we will show how to calculate the *p*-value analytically, without using the approximation of normal distributions.

The *p*-value (observed significance level) can be interpreted as the probability of observing samples equally or more contradictory to the null hypothesis compared to the value that actually resulted. Since the resulted difference between the two samples is $d^*=x_1-x_2$, for the one-tailed test discussed in this paper, the *p*-value is equal to the probability of observing $d \geq d^*$ (i.e., the difference between the number of successes in the two samples is equal to or even greater than x_1-x_2). Without loss of generalization, in this paper it is assumed that $x_1-x_2 \geq 0$.

Given the common proportion *p* as defined in Eq. 4, the sum of the probabilities of observing all possible sample data satisfying this condition is given by:

$$C(p) = \sum_{i=d^*}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left\{ \sum_{j=0}^{i-d^*} \binom{n}{j} p^j (1-p)^{n-j} \right\} \quad (5)$$

It can be seen from Eq. 5 that the number of successes in the first sample (*i*) is always no less than d* greater than the number of successes in the second sample (*j*).

A similar statistical test has been proposed in [4, 5]:

$$C(p) = \sum_{i=x_1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \cdot \sum_{j=0}^{x_2} \binom{n}{j} p^j (1-p)^{n-j} \quad (6)$$

Eq. 6 represents the cumulative probability of observing at least x_1 successes in the first sample and at most x_2 successes in the second sample. However, what matters is the difference between x_1 and x_2, instead of their actual values. In fact, given the null hypothesis, we are only interested in how likely it is that the numbers of successes in the two samples *differ* from each other to a certain extent. Since Eq. 6 imposes a stricter condition on the samples, it will only consider a subset of the sample specified by Eq. 5. Consequently, it will underestimate the *p*-value, making it more likely to make the Type I error (the null hypothesis is rejected when there is actually not enough evidence).

It should be mentioned that, in [4, 5], the common proportion is defined differently from its counterpart in Eq. 4 and Eq. 5. Instead of using the pooled estimate, the specific value that maximizes Eq. 6 is used. Unfortunately, there is no analytical solution to this optimization problem and the required *p* has to be obtained through searching, which involves heavy computation.

The *p*-values for different combinations of x_1, x_2 and *n* values are given in Table 1 in which A, B and C represent the statistical tests based on Eq. 5, Eq. 4 and Eq. 6 respectively. These examples cover both small-sample cases where the classical test (Eq. 4) is not applicable as well as large-sample cases in order to provide a comprehensive head-to-head comparison.

Table 1. The *p*-values for some selected examples

x_1	x_2	n	Statistical Tests		
			A	B	C#
5	3	10	0.2470	0.1807	0.1403
8	3	10	0.0202	0.0123	0.0102
9	6	10	0.0963	0.0607	0.0551
25	15	50	0.0260	0.0260	0.0093
40	15	50	2.48e-7	2.51e-7	6.25e-8
45	30	50	0.0188	0.0145	0.0067
50	30	100	0.0024	0.0019	6.72e-4
80	60	100	0.0013	0.0010	3.46e-4
90	80	100	0.0298	0.0238	0.0106

#: The *p*-values of test C were obtained using the online calculator available at: http://qualopt.eivd.ch/stats/?page=stats

Firstly, by comparing test A and test B, it is clear that, with a small sample size, the classical test often underestimates the true *p*-value and as the sample size goes up, the difference between test A and test B becomes less and less obvious. Since the results from test B with large sample sizes are known to be trustable, it empirically verifies the correctness of the proposed test A.

Secondly, there is a distinct contrast between test C and others. In all of the above examples, it gives a much smaller *p*-value, which means that it is likely to reject the null hypothesis when there is actually not enough evidence against it. As a result, it is very dangerous to establish conclusions based on this test.

4. CONCLUSION

This paper addresses an important question in experimental studies of EAs: how to compare the success rates of algorithms with only a few trials. In review of the classical large-sample hypothesis test based on normal distributions, a new test procedure is developed, which is well suited for comparing the success rates without imposing any constrains on the sample size.

Comparative experiments show that the new test procedure can produce similar results as the classical test when the sample size is large and it is still applicable with small sample sizes. In the meantime, the existing small-sample test always produces biased results (smaller *p*-values) and is not recommended.

5. REFERENCES

[1] Eiben, A. and Jelasity, M. A critical note on experimental research methodology in EC. In *Proceedings of 2002 Congress on Evolutionary Computation*, 2002, 582-587.

[2] McGeoch, C. Toward an experimental method for algorithm simulation. *INFORMS J. Comput.*, 8(1), 1-15, 1996.

[3] Radin, R. and Uzsoy, R. Experimental evaluation of heuristic optimization algorithms: A tutorial. *Journal of Heuristics*, 7, 261-301, 2001.

[4] Taillard, É. A statistical test for comparing success rates. In *Proceedings of 2003 Metaheuristic International Conference*, 2003 (extended abstract).

[5] Taillard, É., Waelti, P. and Zuber, J. Few statistical tests for proportions comparisons. *European Journal of Operational Research*, 185(3), 1336-1350, 2008.