On the Analysis of Experimental Results in Evolutionary Computation

Stjepan Picek
Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
Email:stjepan@computer.org

Marin Golub
Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
Email: marin.golub@fer.hr

Domagoj Jakobovic
Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
Email:domagoj.jakobovic@fer.hr

***Abstract*—Evolutionary computation methods are successfully applied in solving of combinatorial optimization problems. Since the "No Free Lunch" theorem states that there is no single best algorithm to solve all possible problems, throughout the years many algorithms and their modifications have emerged. When a new algorithm is developed, one question that naturally arises is how it compares to other algorithms, whether for some specific problem or in general performance. Because of the stochastic nature of systems involved, usually the only possible way of deriving the answer is to perform extensive experimental analysis. In this paper we provide an overview of possible approaches in the experimental analysis, and describe statistical methods that could be used. Furthermore, we outline similarities and differences between these methods, which lead to a discussion of important issues that need to be resolved when using these methods.**

I. Introduction

Evolutionary computation forms a subfield of computational intelligence that involves combinatorial optimization problems. Even broadly speaking, this field has a multitude of methods and their variations. Among those algorithms and theirs modifications it is often hard to single out the more successful ones. To single out the best overall algorithm would be impossible as is stated in the "No Free Lunch" theorem. Wolpert stated that, when averaged over all test problems, all search algorithms perform equally. This means that if algorithm A is better than algorithm B on some problems, then the algorithm B will be better than the algorithm A in exactly as many different problems [1]. However, it is important to state that"No Free Lunch" theorem assumes no knowledge of the problems. Since almost always we have at least partial knowledge of the problems at hand, it is possible to sort out the better algorithms for particular classes of problems.

If it is possible to find better, more suitable algorithms for a particular problem, the next question is how to recognize which of those algorithms are better. Due to the fact that the theory of evolutionary computation is often not developed enough, one way to do it is to perform an experimental analysis.

It is necessary to realize that the experimental analysis is usually not a trivial task. When we have algorithms that we want to compare, then it is first necessary to decide on the test problems that should be used. Deciding about the test problems is often more difficult then it seems. If the problems are too easy, then the algorithms will always find the global optimum. From the other side, if the problems are too hard, algorithms will typically get stuck in local optima. Optimal parameters must also be found for those problems (this alone can be a very difficult task where approaches range from guessing to complex parameter tuning methods). After these preparation steps have been done, the experiments can be made.

The next step is the analysis of experimental results which requires the usage of statistical methods. However, there are different possibilities regarding statistical procedures that can be used. Although often different statistical methods can reach the same or very similar results, the wrong choice of a method can lead to a misleading result [2]. Even if the analysis is properly conducted, there is still the question whether the results are more depending on the algorithm design or the problem design.

Our aim is to discuss statistical methods, with an accent on the nonparametric statistical procedures since those methods are rarely used, but are often a needed tool for conducting a proper statistical analysis.
In Section 2 we present relevant theory needed for the understanding of the experimental models and the statistics employed in the analysis of the results. In Section 3 a review of nonparametric statistical methods is done, in Section 4 we give a hypothetical test case where nonparametric statistic procedures should be used, and finally, in Section 5 a conclusion is given.

II. Preliminary

A. Experimental Analysis Models

When performing experiments, several options exist regarding the algorithms and problems choice. Regarding the chosen model, the appropriate statistical analysis should be chosen as well.

1) Single Algorithm Single Problem analysis: This approach is best when some real-world setting is investigated. It also brings the benefit of thorough investigation of dedicated algorithm on some particular problem which is often not possible when there are more algorithms or problems.

2) Single Algorithm Multiple Problem analysis: Main benefit of this approach is that it can help in better understanding of the algorithm and its parameters.

3) Multiple Algorithm Single Problem analysis: This approach can help in understanding of the algorithms and their similarities.

4) Multiple Algorithm Multiple Problem analysis: This approach usually gives the biggest mass of data but smallest actual insight in the inner mechanisms of algorithms or understanding of problems. This is because of the complexity of this approach.

Among the approaches mentioned above, the last one is the most common one, since that approach usually suggests (although that is not necessarily true) a serious research paper.

When the model of the analysis is a single problem scenario, it should be safe to use parametric statistical tests, however, when the model has a multiple problem scenario, an analysis should be done regarding the choice of appropriate statistical methods.

B. Statistical Analysis

1) Descriptive Statistics: Descriptive statistics represents statistics that allows us to present data in a way that is easier when interpreting the results. This kind of statistics does not allow making any conclusions beyond the data that have been analyzed.

2) Inferential Statistics: Inferential statistics represents techniques that allow us to make generalizations about the populations on the basis of the samples that are evaluated. Inferential statistics employs two methodologies: hypothesis testing, and estimation of population parameters.

a) Hypothesis Testing: Hypothesis represents a prediction about a population or relationship between two or more populations. Hypothesis testing is a procedure in which sample data is employed to evaluate a hypothesis. There is also a research hypothesis which represents a statement of what the research predicts, and statistical hypotheses which represent restatement of the research hypothesis.

Statistical hypotheses are null hypothesis which is a statement of no difference, e.g. a hypothesis that researcher expects to be rejected. Second statistical hypothesis is alternative hypothesis and it represents a statement of difference, e.g. a hypothesis that researcher expects to be supported [3].

The more conventional hypothesis testing model used in inferential statistics assumes *a priori* stating at what level of significance , the null hypothesis will be evaluated. Other approach is that instead of *a priori* setting of a level of significance, one calculates the smallest level of significance that results in rejection of null hypothesis where that is represented by p-value [3]. P-value is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true.

b) Estimation of Population Parameters: This framework is employed for estimating the value of one or more population parameters. There can be point estimation where a value is estimated from the computed value of statistics, and interval estimation where a range of values are computed and in which a true value of a parameter falls with a high degree of certainty [3].

C. Parametric and Nonparametric Statistics

Statistical procedure employed in inferential statistics can be divided in parametric statistical methods and nonparametric statistical methods. The distinction between parametric and nonparametric statistical methods can be on the basis that parametric tests make specific assumptions regarding population parameters that characterize the underlying distribution whereas nonparametric tests make no such assumptions. Because of that, the nonparametric statistical tests are also sometimes called distribution free statistical methods.

The second distinction that can be used is that parametric statistical methods work with categorical/nominal data and nonparametric methods work with ordinal/rank-order data [3].

In an analysis, it is necessary to decide whether to use parametric or nonparametric tests. Parametric tests present a more powerful option when testing alternative hypothesis than nonparametric tests, but if some assumptions for using parametric tests are violated then that advantage can be lost. If no assumptions of a parametric tests have been violated, and when the level of measurement for a set of data is interval or ratio then the data should be evaluated with parametric tests. If one or more assumptions are violated then it could be prudent to transform the data into format that is compatible with nonparametric tests. However, not all researchers agree on this approach because in the transformation of the data from interval/ratio data to the ordinal/rank-order format some of the information is lost. This is due to the fact that interval/ratio data contain more information. Because of that, some researchers advocate the usage of parametric test with some adjustments when assumptions for parametric test are violated [3]. Finally, usual conclusion is that in the end, parametric or nonparametric tests give the same or very similar results and so there is no real consequence in choosing the wrong tests. When there are conflicting results between these two approaches, often it could be enough to conduct multiple experiments [3].

However, as displayed in [2], it can be seen that wrong choice of statistical tests can lead to a misleading conclusion. Additionally, conducting multiple experiments may not be

always possible due to the complexity of the problems and often limited time frame on disposal.

To decide on the choice of nonparametric or parametric statistical tests, it is necessary to check all the conditions to see if it is allowed to use parametric tests. These conditions are independence, normality, and homoscedasticity. For details about the necessary conditions or tests check [2] [3].

1) Independence
   Two events are independent if the occurrence of one of the events gives no information about whether or not the other event will occur.
2) Normality
   Normal or Gaussian distribution is a continuous probability distribution that has a bell-shaped probability density function. An observation is normal if it follows normal distribution. Tests that can be used to indicate the presence of normality in a sample are:
   - Kolmogorov-Smirnov test
   - Shapiro-Wilk test
   - D'Agostino-Pearson test
3) Homoscedasticity
   Homoscedasticity is also called homogeneity of variance. It is a property which shows that all of the samples have the same variances. The tests that can be used to check homoscedasticity are:
   - Bartlett test
   - Levene test

For every parametric statistical test, there exists a nonparametric statistical equivalent as shown in Table IV.

TABLE I
PARAMETRIC VERSUS NONPARAMETRIC TESTS

	Parametric test	Nonparametric test
Pairwise	t-test	Sign test Wilcoxon signed rank test
Multiple	ANOVA Tukey, Tamhane	Friedman Holm, Hochberg, Rom, Li,...

III. NONPARAMETRIC STATISTICAL PROCEDURES

Once we decide to use nonparametric statistical methods, the first step is to decide whether to use pairwise or multiple comparison tests. Pairwise tests are a good option when two methods are compared. If more than two methods are compared, multiple comparison tests should be used because pairwise tests do not control error propagation of making more than one comparison.

As we stated before, since most of the research work today compare more than two methods (multiple algorithms multiple problem scenario), we will put an emphasis on the use of multiple comparison tests.

A. Discovering Global Differences

The simplest tests for multiple comparisons are Friedman test and its extension Iman-Davenport test. The purpose of those tests is to answer whether there are global differences between related samples obtained.

These tests do not give the answer which algorithm is better: they only show if the null hypothesis is rejected, i.e. if there are differences between algorithms. More details about Friedman and Iman-Davenport tests can be found in [3] [4] . The Friedman two-way analysis of variances by ranks test is a nonparametric analogue of the parametric two-way analysis of variance. Ranking of the data is the transformation that enables the usage of nonparametric statistical methods on the real data. There are several possibilities how to conduct that transformation, and depending on that, it is possible to improve the standard Friedman test.

B. Post-hoc Procedures

When we determine that the null hypothesis is rejected through the use of tests mentioned above, then we can use post-hoc procedures to find where those differences exactly are. Those tests can be done in the way that control method (usually one that is suggested by researchers) is tested against other methods (1 x N comparison) or to conduct multiple comparisons (N x N comparisons) among all methods.

With the post-hoc procedures we can obtain the p-values which determine the degree of rejection of each hypothesis. P-value gives information whether some statistical hypothesis is significant, and if it is, then how significant it is. The smaller the p-value is, the more strongly the null hypothesis is rejected. However, p-values obtained directly from the formulas are not suitable for multiple comparisons case. When a p-value is considered in a multiple comparison test, it reflects the probability error of a certain comparison, but it does not take into account the remaining comparisons.

This problem is possible to solve with the usage of adjusted p-values. Adjusted p-values can solve that problem since they take into account the accumulated error [3]. When a p-value is found, then it can be compared with desired level of significance α to see if the null hypothesis is rejected. In regards how the level of significance α can be adjusted to compensate for multiple comparisons there are following procedures and examples of tests:

- One-step: Bonferroni-Dunn test
- Step-down: Holm, Holland, Finner tests
- Step-up: Hochberg, Hommel, Rom tests
- Two-step: Li test

C. Contrast Estimation Procedure

A contrast estimation procedure based on medians can be used to estimate the differences between each two crossover operators. In this test the performance of the algorithms is reflected by the magnitudes of the differences in error rates [4].

TABLE II
BENCHMARK FUNCTIONS

Test function	Abbreviation
$f(x) = \sum_{i=1}^{D} x_i^2$	P_1
$f(x) = \sum_{i=1}^{D} i \cdot x_i^2$	P_2
$f(x) = \sum_{i=1}^{D} 5 \cdot i \cdot x_i^2$	P_3
$f(x) = \sum_{i=1}^{D} \left(\sum_{j=1}^{i} x_j^2 \right)$	P_4
$f(x) = \sum_{i=1}^{D-1} 100 \cdot \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2$	P_5
$f(x) = 10 \cdot D + \sum_{i=1}^{D} \left(x_i^2 - 10 \cdot \cos(2 \cdot \Pi \cdot x_i) \right)$	P_6
$f(x) = \sum_{i=1}^{D} -x_i \cdot \sin\left(\sqrt{\lvert x_i \rvert} \right)$	P_7
$f(x) = \sum_{i=1}^{D} x_i^2/4000 - \prod_{i=1}^{D} \cos\left(x_i/\sqrt{i} \right) + 1$	P_8
$f(x) = -20 \cdot e^{-0.2\sqrt{\sum_{i=1}^{D} x_i^2/D}} - e^{\sum_{i=1}^{D} \cos(2\Pi x_i)/D} + 20 + e$	P_9
$f(x) = -\sum_{i=1}^{D} \sin(x_i) \cdot \left(\sin\left(i \cdot x_i^2/\Pi \right) \right)^{20}$	P_{10}

IV. AN EXAMPLE APPLICATION OF NONPARAMETRIC STATISTICAL ANALYSIS

The purpose of this section is to go through one simple example of statistical analysis for a genetic algorithm (GA) case (of course, these steps are valid in general case). We compare four different crossover operators (single-point crossover - Alg. 1, uniform crossover - Alg. 2, shuffle crossover - Alg. 3, non-geometric crossover - Alg. 4) in a binary-coded genetic algorithm with a roulette-wheel selection. GAs with different crossover operators can be regarded as different algorithms. The comparison is made on a set of 10 standard benchmark functions. The formulas for these benchmark functions are given in Table II.

More details on the benchmark functions and the algorithms can be found in [5]. As a performance measure the error rate obtained for every operator (algorithm) is used. For all the test functions, the objective is to find global minimum. For each algorithm 30 independent runs are made.
From the each run of the algorithm, the individual with the smallest error rate is chosen as the best one from that run. Then, the mean value is taken from those 30 values for each algorithm and problem. Those mean values for all the algorithms and problems are displayed in Table III.

TABLE III
MEAN ERROR RATES FOR EVERY ALGORITHM AND PROBLEM

Problem	Alg. 1	Alg. 2	Alg. 3	Alg. 4
P_1	0.056	0.065	0.112	0.139
P_2	233.41	434.133	1197.863	1799.66
P_3	36.53	36.404	28.432	36.74
P_4	0.201	0.201	0.675	0.953
P_5	1.003	1.003	1.008	1.007
P_6	0.012	0.009	0.035	0.024
P_7	7.547	7.445	8.21	8.603
P_8	0.055	0.038	0.269	0.283
P_9	93.47	91.876	100.342	102.24
P_{10}	2.243	2.230	2.349	2.365

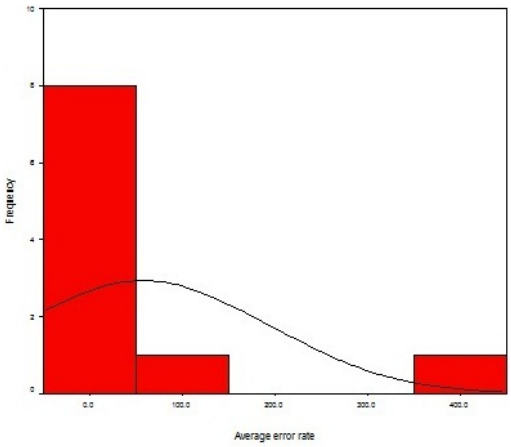

Fig. 1. Histogram for the uniform crossover case

First it is necessary to decide whether to use parametric or nonparametric tests so we test indenpendence, normality and homoscedasticity assumptions. Independence condition is easily checked since there are independent runs of the algorithms.

Normality and homoscedasticity conditions are estimated at the level of significance α of 0.01.
For normality condition we use Shapiro-Wilk test that can be easily conducted in SPSS tool [6]. Shapiro-Wilk is used instead of Kolmogorov-Smirnov because the number of samples is smaller than 50. The obtained values are smaller than 0.01, so we can conclude that normality condition is not satisfied. In Fig. 1 a histogram is displayed for the uniform crossover case where it can be seen that the solutions do not follow normal distribution.

For the homoscedasticity test we can use Levene test that can also be done in SPSS tool. However, since the normality assumption is not satisfied we do not need to conduct the homoscedasticity test.

Next, we can start with the nonparametric statistical tests. We use KEEL tool for conducting nonparametric statistical analysis [7]. There are many possible choices of the statistical packages that can be used, but we decided on KEEL since it is very easy to use and free. First we conduct Friedman and Iman-Davenport test. Table V gives the results for a Friedman two-way analysis of variances by ranks.

The transformation to rank data is done in the following way: for each row in Table III a rank of 1 is assigned to the lowest score in that row, a rank of 2 is assigned to the second smallest value in that row and so on until the rank 4. If two algorithms have the same value for a problem (row), then the average value of the ranks that are involved is assigned to all algorithms tied for a given rank. In Table IV are displayed all the ranked values and the scores for the each column.

To compute the chi-square approximation of the Friedman

TABLE IV
RANKED VALUES FOR ALL THE ALGORITHMS AND PROBLEMS

Problem	Alg. 1	Alg. 2	Alg. 3	Alg. 4
P1	1	2	3	4
P2	1	2	3	4
P3	3	2	1	4
P4	1.5	1.5	3	4
P5	1.5	1.5	4	3
P6	2	1	4	3
P7	2	1	3	4
P8	2	1	3	4
P9	2	1	3	4
P10	2	1	3	4
	$R_1 = 18$	$R_2 = 14$	$R_3 = 30$	$R_4 = 38$

TABLE V
AVERAGE RANKINGS OF THE ALGORITHMS (FRIEDMAN)

Algorithm	Ranking
Single-point	1.8
Uniform	1.4
Shuffle	3
Non-geometric	3.8

test statistic we use 1:

$$\chi_r^2 = \frac{12}{n * k(k+1)} \left[\sum_{j=1}^{k} \left(\frac{R_j}{n} \right)^2 \right] - 3 * n\,(k+1) \qquad (1)$$

Here *n* represents the number of problems, *k* the number of algorithms, and R_j are column scores from the Table IV.

With the level of significance α of 0.01 both the Friedman and Iman-Davenport statistic show significant differences on operators with test values of 21.84 and 24.08, respectively, and $p < 0.001$.

The next test that can be done is the Bonferroni-Dunn test. This procedure does not have a great resolution in distinguishing differences but is appropriate for graphical display. For the Bonferroni-Dunn test, first the critical difference (CD) must be found [2]. The equation for calculating critical difference is as follows:

$$CD = q_\alpha \sqrt{\frac{k * (k+1)}{6 * n}} \qquad (2)$$

where q_α is the critical value for two-tailed Bonferroni-Dunn test.
Critical difference computed according to 2 has a value of 1.38. The interpretation of this measure is that the performance of two algorithms is significantly different only if the corresponding mean ranks differ by at least a critical difference, which is depicted in Fig. 2. A cut line is drawn at height equal to the sum of critical difference and ranking of the control algorithm. The bars that exceed this line are associated with the algorithms that have worse performance than the control algorithm.

Now, post-hoc procedures can be applied. We use one procedure from every class mentioned previously: Bonferroni-Dunn, Holm, Hochberg, and Li test. For the control algorithm,

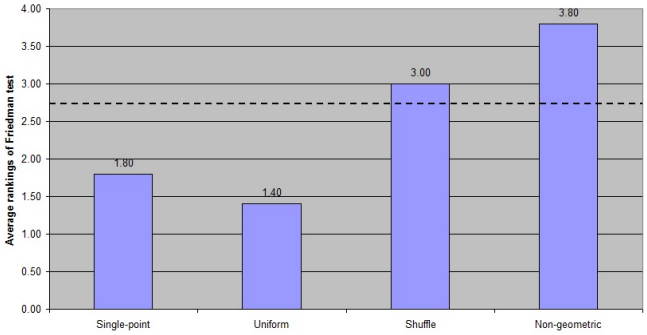

Fig. 2. Bonferroni-Dunn's test, critical difference = 1.38, control operator: uniform crossover

uniform crossover is selected as the best one from Friedman test. Results obtained from the post-hoc analysis are presented in Table VI.

TABLE VI
POST-HOC COMPARISON (CONTROL OPERATOR: UNIFORM CROSSOVER)

Algorithm	p_{Li}	p_{Bonf}	p_{Holm}	$p_{Hochberg}$
Non-geometric	0.000063	0.000097	0.000097	0.000097
Shuffle	0.010797	0.016751	0.011167	0.011167
Single-point	0.488422	1.465267	0.488422	0.488422

From the Table VI it is obvious that the Bonferroni-Dunn results are confirmed; and that the uniform crossover is better than the shuffle and the non-geometric crossover. However, we still can not answer the question whether the uniform or single-point crossover are better. Because of that, now we use the contrast estimation method.

TABLE VII
CONTRAST ESTIMATION

	Single-point	Uniform	Shuffle	Non-geometric
Single-point	0	0.026	-0.138	-0.258
Uniform	-0.026	0	-0.164	-0.284
Shuffle	0.138	0.164	0	-0.12
Non-geometric	0.258	0.284	0.12	0

A negative value for the operator in a given row indicates that the operator performs better than the operator in a given column. From the Table VII we can see that uniform crossover is the best operator for the selected benchmark functions.

This simple example should be regarded just as a roadmap through the nonparametric statistical tests, and not as some representative case. Often, we can not get the definitive answer which algorithm is the best one, but a set of better algorithms.

V. CONCLUSION

Experimental analysis often represents the only method of evaluating the performance in evolutionary computation. Because of that, it is important to use as appropriate statistical methods as possible. It may also be advisable to perform parametric statistical tests together with nonparametric statistical

tests.

Only by applying a rigorous statistical analysis on a set of well defined benchmark functions a proper performance evaluation can be made.

Only then can we properly evaluate not only existing algorithms but also the future ones. Some of those algorithms proved to be very successful and now are widely used. Some other algorithms were presented as successful (even better than the more common ones) but they soon disappeared from usage.

REFERENCES

[1] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[2] S. Garcia, D. Molina, M. Lozano, and F. Herrera, "A study on the use of non-parametric tests for analyzing the evolutionary algorithms? behaviour: a case study on the CEC 2005 special session on real parameter optimization," *Journal of Heuristics*, no. 15, pp. 617–644, 2009.

[3] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th ed. Chapman and Hall/CRC, 2007.

[4] J. Derrac, S. Garcia, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, pp. 3–18, 2011.

[5] S. Picek, M. Golub, and D. Jakobovic, "Evaluation of crossover operator performance in genetic algorithms with binary representation," *Lecture Notes in Computer Science*, vol. 6840, pp. 223–230, 2011.

[6] "SPSS predictive analytics software and solutions," Available from http://www-01.ibm.com/software/analytics/spss/, Feb. 2012.

[7] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garc?a, L. S?nchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, pp. 255–287, 2011.