A SURVEY OF NONPARAMETRIC TESTS FOR THE STATISTICAL ANALYSIS OF EVOLUTIONARY COMPUTATIONAL EXPERIMENTS

Rafael Lahoz-Beltra, Carlos Perales-Gravan

***Abstract**: One of the main problems in the statistical analysis of Evolutionary Computation (EC) experiments is the 'statistical personality' of data. A main feature of EC algorithms is the sampling of solutions from one generation to the next. Sampling is based on Holland's schema theory, having a greater probability to be chosen those solutions with best-fitness (or evaluation) values. In consequence, simulation experiments result in biased samples with non-normal, highly skewed, and asymmetric distributions. Furthermore, the main problem arises with the noncompliance of one of the main premises of the central limit theorem, invalidating the statistical analysis based on the average fitness $\overline{f}$ of the solutions. In this paper, we address a tutorial or 'How-to' explaining the basics of the statistical analysis of data in EC. The use of nonparametric tests for comparing two or more medians combined with Exploratory Data Analysis is a good option, bearing in mind that we are only considering two experimental situations that are common in EC practitioners: (i) the performance evaluation of an algorithm and (ii) the multiple experiments comparison. The different approaches are illustrated with different examples (see http://bioinformatica.net/tests/survey.html) selected from Evolutionary Computation and the related field of Artificial Life.*

***Keywords**: Evolutionary Computation, Statistical Analysis and Simulation.*

***ACM Classification Keywords**: G.3 PROBABILITY AND STATISTICS*

***Conference topic**: Evolutionary Computation.*

Introduction

Evolutionary Computation (EC) refers to a class of stochastic optimization algorithms inspired in the evolution of organisms in Nature by means of Darwinian natural selection [Lahoz-Beltra, 2004][Lahoz-Beltra, 2008]. Nowadays, this class of algorithms is applied in many diverse areas, such as scheduling, machine learning, optimization, electronic circuit design [Lahoz-Beltra, 2001][Perales-Gravan and Lahoz-Beltra, 2008], pattern evolution in biology (i.e. zebra skin pattern) [Perales-Gravan and Lahoz-Beltra, 2004], etc. All methods in EC are *bioinspired* in the fundamental principles of neo-Darwinism, evolving a set of potential solutions by a selection procedure to sort candidate solutions for breeding. At each generation, a new set of solutions is selected for reproduction, contributing with one or more copies of the selected individuals to the offspring representing the next generation. The selection is carried out according to the goodness or utility of the solutions x_i, thus calculating the values $f(x_i)$ which are known as fitness values. Once the selection has concluded, the next generation of solutions is transformed by the simulation of different genetic mechanisms (Fig. 1). The genetic mechanisms are mainly crossover or recombination (combination of two solutions) and/or mutation (random change of a solution). These kinds 'genetic procedures' evolve a set of solutions, generation after generation, until a set of solutions is obtained with one of them representing an optimum solution. Genetic algorithms, evolutive algorithms, genetic programming, etc. are different types of EC algorithms. However all of them share with some variations the following general steps:

1. Generate at random an initial set of solutions $x_i \in S(0)$.
2. Evaluate the fitness of each solution $f(x_i)$.
3. Select the best-fitness solutions to reproduce.
4. Breed a new generation (t=t+1) of solutions $S(t)$ through crossover and/or mutation and give birth to offspring.
5. Replace a part of solutions with offspring.
6. Repeat 2-5 steps until {terminating condition}.

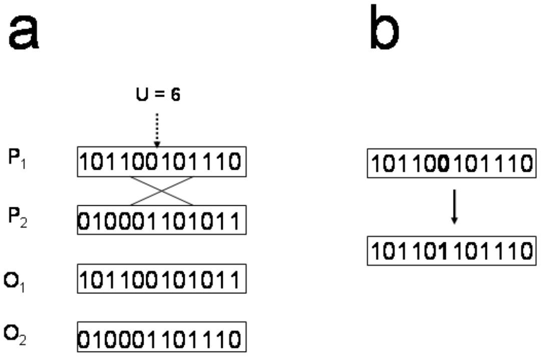

Figure 1.- Evolutionary Computation methods are based on genetic mechanisms simulation such as crossover and/or mutation. In (a) crossover two parental solutions represented in 1D-arrays called chromosomes (P_1 and P_2) exchange their segments (in this example, U is the one-point crossover randomly selected) obtaining two recombinant solutions, O_1 and O_2, thus the offspring solutions. However, in (b) mutation a random ‘genetic’ change occurs in the solution, in this example replacing or inverting in position 6 a bit value 0 by 1.

However, at present EC methods lack of a general statistical framework to compare their performance [Czarn et al., 2004] and evaluate the convergence to the optimum solution. In fact, most EC practitioners are satisfied with obtaining a simple performance graph [Goldberg, 1989][Davis, 1991] displaying the *x*-axis the number of generation, simulation time or epoch and the *y*-axis the average fitness per generation (other such possibilities exist such as the maximum fitness per generation at that point in the run).

One of the main problems in the statistical analysis of EC experiments is the ‘statistical personality’ of data. The reason is that selection of the best-fitness solutions generation after generation leads to non-normal, highly skewed, and asymmetric distributions of data (Fig. 2). At present, there are many available techniques that are common in EC algorithms to select the solutions to be copied over into the next generation: fitness-proportionate selection, rank selection, roulette-wheel selection, tournament selection, etc. A main feature of selection methods is that all of them generate new random samples of solutions x_1, x_2 ,.., x_N but *biased* random samples. Thus, solutions are chosen at random but according to their fitness values $f(x_i)$, having a greater probability to be chosen those x_i solutions with the best-fitness values. The consequence is that EC algorithms select the solutions x_1, x_2 ,.., x_N (or sample) from one generation to the next based on Holland’s schema theory [Holland, 1992]. This theorem – the most important theorem in EC- asserts the following: the number *m(H)* of ‘short’ (distance between the first and last positions) and ‘low-order’ (number of fixed positions) solutions *H* (called *schema*) with above-average fitness *f(H)* increase exponentially in successive generations:

$$m(H,t+1) \geq \frac{m(H,t).f(H)}{\bar{f}(t)}[1-p] \qquad (1)$$

where $\overline{f}(t)$ is the average fitness of the set of solutions at time *t*, and p is the probability that crossover or mutation will destroy a solution *H*.

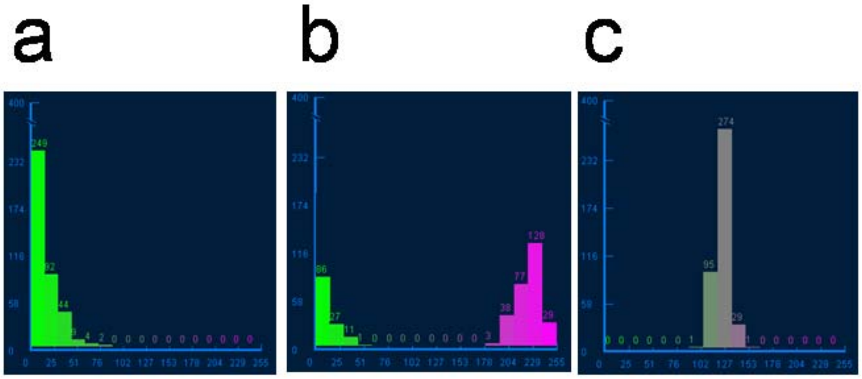

Figure 2.- Darwinian natural selection modes (see Manly, 1985). (a) Directional selection, (b) disruptive, and (c) stabilizing selection. Histograms were obtained with EvoTutor selection applet (see http://www.evotutor.org/TutorA.html).

The main statement and motivation of the present paper is as follows. EC practitioners frequently use parametric methods such as *t*-student, ANOVA, etc. assuming that x_1, x_2 ,.., x_N sample is a sequence of independent and identically distributed values (i.i.d.). In such cases, the non-compliance of one of the main premises of the central limit theorem (CLT), invalidate the statistical analysis based on the average fitness $\overline{f}$ of the solutions x_1, x_2 ,.., x_N :

$$\overline{f} = \frac{f(x_1) + f(x_2) + \ldots + f(x_N)}{N} \tag{2}$$

In consequence, there is no convergence of $\sqrt{N}\left(\overline{f} - \mu\right)$ towards the standard normal distribution $N(0, \sigma^2)$. We suggest that nonparametric tests for comparing two or more medians could provide a simple statistical tool for the statistical analysis of data in EC. Furthermore, since important assumptions about the underlying population are questionable and the fitness values of the solutions can be put in order, thus $f(x_1)$, $f(x_2)$,.., $f(x_N)$ are ranked data, then the statistical inference based on ranks [Hettmansperger, 1991] provide a useful approach to compare two or more populations.

In the present paper and according with the above considerations, we illustrate how the statistical analysis of data in EC experiments could be addressed using assorted study cases and general and simple statistical protocols. The protocols combine the Exploratory Data Analysis approach, in particular Box-and-Whisker Plots [Tukey, 1977], with simple nonparametric tests [Siegel and Castellan, 1988][Hollander and Wolfe, 1999][Gibbons and Chakraborti, 2003]. The different approaches are illustrated with different examples chosen from Evolutionary Computation and Artificial Life [Prata, 1993][Lahoz-Beltra, 2008].

Performance analysis and comparison of EC experiments

Most of the general research with EC algorithms usually addresses two type of statistical analysis (Table I).

Table I.- Statistical analysis protocols in Evolutionary Computation

Performance evaluation	Robust Performance Graph Statistical Summary Table		
Simulation experiments comparison	*K_e= 2 experiments*	Multiple Notched Box-and-Whisker Plot Statistical Summary Table	$\sigma_i = \sigma_j$ Mann-Whitney (Wilcoxon) test
			$\sigma_i \neq \sigma_j$ Studentized Wilcoxon test
	K_e> 2 experiments	Multiple Notched Box-and-Whisker Plot Statistical Summary Table Kruskal-Wallis test Dunn's post-test	

The evaluation of the algorithm performance is one of the most common tasks in EC experiments. In such a case, the evaluation could be carried out combining a robust performance graph with a statistical summary table. A robust performance graph is as a plot with the *x*-axis displaying the number of generation, simulation time or epoch, depicting for each generation a Notched Box-and-Whisker Plot [McGill et al., 1978]. The Notched Box-and-Whisker Plot shows the distribution of the fitness values of the solutions, displaying the *y*-axis the scale of the batch of data, thus the fitness values of the solutions. The statistical summary table should include the following descriptive statistics: the average fitness or other evaluation measure (i.e. mean distance, some measure of error) computed per generation, the median and the variance of the fitness, as well as the minimum, maximum, Q_1, Q_3, the interquartile range (IQR), and the standardized skewness and standard kurtosis.

Comparative studies are common in simulation experiments with EC algorithms. In such a case researchers use to compare different experimental protocols, genetic operators (i.e. one-point crossover, two-points crossover, uniform crossover, arithmetic crossover, heuristic crossover, flip-a-bit mutation, boundary mutation, Gaussian mutation, roulette selection, tournament selection, etc.) and parameter values (i.e. population size, crossover probability, mutation probability). According to tradition, a common straightforward approach is the performance graph comparison. In this approach, practitioners have a quick look at the plot lines of different simulation experiments, without any statistical test to evaluate the significance or not of performance differences. In the case of two experiments (k_e=2) with non-normal data, the statistical comparison could be addressed resorting to a Multiple Notched Box-and-Whisker Plot, the statistical summary table and a Mann-Whitney (Wilcoxon) test [Mann and Whitney, 1947]. The statistical summary table (i.e. IQR or the box length in the Box-and-Whisker Plot) is important since differences in population medians are often accompanied by other differences in spread and shape, being not sufficient merely to report a *p* value [Hart, 2001]. Note that Mann-Whitney (Wilcoxon) test assumes two populations with continuous distributions and similar shape, although it does not specify the shape of the distributions. The Mann-Whitney test is less powerful than *t*-test because it converts data values into ranks, but more powerful than the median test [Mood, 1954] presented in many statistical textbooks and popular statistical software packages (SAS, SPSS, etc.). Freidlin and Gastwirth [Freidlin and Gastwirth, 2000] suggested that the median test should be "retired" from routine use, showing the loss of power of this test in the case of highly unbalanced samples. Nevertheless, EC simulation experiments are often designed with balanced samples. It is important to note that in this tutorial, we only consider the case of two EC experiments where distributions may differ only in medians. However, when testing two experiments (k_e=2) researchers should take care of the statistical analysis under heteroscedasticity. For instance, Table XIII (see http://bioinformatica.net/tests/survey.html) shows three simulation experiments (*study case* 5) with significant differences among variances. In such a case, if our goal were for testing the significance or not of performance of two simulation experiments a studentized Wilcoxon test [Fung, 1980] should be used instead the Mann-Whitney test.

In the case of more than two experiments (k_e>2) with non-normal data the Multiple Box-and-Whisker Plot and the statistical summary table can be completed making inferences with a Kruskal-Wallis test [Kruskal and Walis, 1952]. This approach can be applied even when variances are not equal in the *k* simulation experiments. In such a case, thus under heteroscedasticity, medians comparisons also could be carried out using the studentized Brown and Mood test [Fung, 1980] as well as the S-PLUS and R functions introduced by Wilcox [Wilcox, 2005][Wilcox, 2006]. However, the loss of information involved in substituting ranks for the original values makes this a less powerful test than an ANOVA. Once again, the statistical summary table is important, since the Kruskal-Wallis test assumes a similar shape in the distributions, except for a possible difference in the population medians. Furthermore, it is well suited to analyzing data when outliers are suspected. For instance, solutions with fitness values laying more 3.0 times the IQR. If the Kruskal-Wallis test is significant then we should perform multiple comparisons [Hochberg and Tamhane, 1987] making detailed inferences on $\binom{k_e}{2}$ pairwise simulation experiments. One possible approach to making such multiple comparisons is the Dunn's post-test [Dunn, 1964].

The Box-and-Whisker plot

The Box-and-Whisker Plot, or boxplot, was introduced by Tukey [Tukey, 1977] as a simple but powerful tool for displaying the distribution of univariate batch of data. A boxplot is based on five number summary: minimum (Min), first quartile (Q_1), median (Me), third quartile (Q_3), and maximum (Max). One of the main features of a boxplot is that is based on robust statistics, being more resistant to the presence of outliers than the classical statistics based on the normal distribution. In a boxplot [Frigge et al., 1989; Benjamini, 1988] a central rectangle or box spans from the first quartile to the third, representing the interquartile range (IQR = Q_3-Q_1, where IQR=1.35xσ for data normally distributed), which covers the central half of a sample. In the simplest definition, a central line or segment inside the box shows the median, and a plus sign the location of the sample mean. The whiskers that extend above (upper whisker) and below (lower whisker) the box illustrate the locations of the maximum (Q_3+k(Q_3-Q_1) and the minimum (Q_1-k(Q_3-Q_1)) values respectively, being usually k=1.5. In consequence, small squares or circles (its depend on the statistical package) outside whiskers represent values that lie more than 1.5 times the IQR above or below the box, whereas those values that lie more 3.0 times the IQR are shown as small squares or circles sometimes including a plus sign. Usually, the values that are above or below 3xIQR are considered outliers whereas those above or below 1.5xIQR are suspected outliers. However, outlying data points can be displayed using a different criterion, i.e. unfilled for suspected outlier and filled circles for outliers, etc. Boxplots can be used to analyze data from one simulation experiment or to compare two or more samples from different simulation experiments, using medians and IQR during analysis without any statistical assumptions. It is important to note that a boxplot is a type of graph that shows information about: (a) location (displayed by the line showing the median), (b) shape (skewness by the deviation of the median line from box central position as well as by the length of the upper whisker in relation with length of the lower one), and (c) variability of a distribution (the length of the box, thus the IQR value, as well as the distance between the end of the whiskers).

A slightly different form boxplot is the Notched Box-and-Whisker Plot or notched boxplot [McGill et al., 1978]. A notched boxplot is a regular boxplot including a notch representing an approximate confidence interval for the median of the batch of data. The endpoints of the notches are located at the median $\pm 1.5\frac{IQR}{\sqrt{n}}$ such that the medians of two boxplots are significantly different at approximately the 0.05 level if the corresponding notches do not overlap. It is important to note that sometimes a 'folding effect' is displayed at top or bottom of the notch. This folding can be observed when the endpoint of a notch is beyond its corresponding quartile, occurring when the sample size is small.

In the following site **http://bioinformatica.net/tests/survey.html** we describe how the statistical analysis of six selected study cases was accomplished using the statistical package STATGRAPHICS 5.1 (Statistical Graphics Corporation), excepting the Dunn test which was performed using Prisma® (Graph Pad Software, Inc.) software. In this web site we included the study cases explanation as well as the statistical summary Tables of this paper.

Hands-on statistical analysis

Statistical performance

The Fig. 3 shows a robust performance graph obtained with a simple genetic algorithm (*study case* 1). Note how the performance has been evaluated based on a sequential set of Notched Box-and-Whisker Plots, one boxplot per generation. The dispersion measured with the IQR of the fitness values decreases during optimization, being the average fitness per generation, thus the classical performance measure in genetic algorithms, greater or equal to the median values during the first six generations. After the sixth generation some chromosomes have an outlying fitness value. Maybe some of them, i.e. those represented in generations 8 and 10, would be representing optimal solutions. The Table IV summarizes the minimum and maximum fitness, the mean fitness, median fitness and the IQR values per generation.

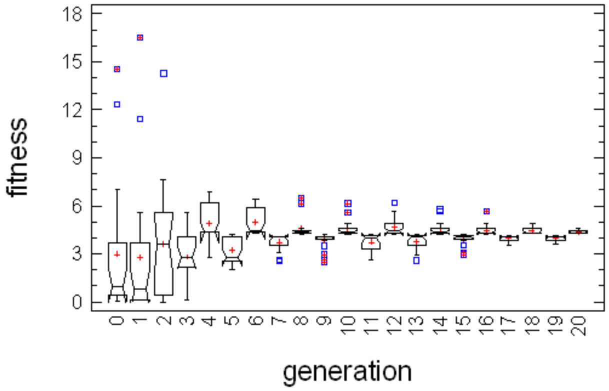

Figure 3.- Robust performance graph in a simple genetic algorithm showing the medians (notches) and means (crosses) of the fitness. Squares and squares including a plus sign indicate suspected outliers and outliers respectively.

The performance of the second and third study cases was evaluated using the Hamming and Euclidean distances. Such distances are useful when the individuals in the populations are defined by binary and integer chromosomes, respectively. In particular, in the *study case* 2, thus the experiment carried out with the ant population, since ants are defined by a 10 bits length chromosome, a robust performance graph (Fig. 4) is obtained based on the Hamming distance. The Kruskal-Wallis test (Table V) shows with a *p*-value equal to 0.0384 that there is a statistically significant difference among medians at the 95.0% confidence level. Note that the Multiple Box-and-Whisker Plot (Fig. 4) shows an overlapping among notches. At first glance, we perceive an overlapping among the 0, 1, 2 and 3 sample times, and between 4 and 5 sample times. Table VI summarizes the genetic drift, thus the average Hamming distance per generation, the median and the variance of the Hamming distance, as well as the minimum, maximum, Q_1, Q_3, IQR and the standardized skewness and standard kurtosis. The special importance is how the genetic drift decreases with the sample time, illustrating the population evolution towards the target ant. Genetic drift values are related with the Hamming ball of radius *r*, such that with *r*=2 the number of ants with distance $d(a,t) \leq 2$ increases with sample time (Table VII). An important fact is the similar shapes of distributions, according to one of the main assumptions of the Kruskal-Wallis test. In particular, for any sample time the standard skewness (Table VI) is lesser than zero. Thus, the distributions are all asymmetrical showing the same class of skewed tail.

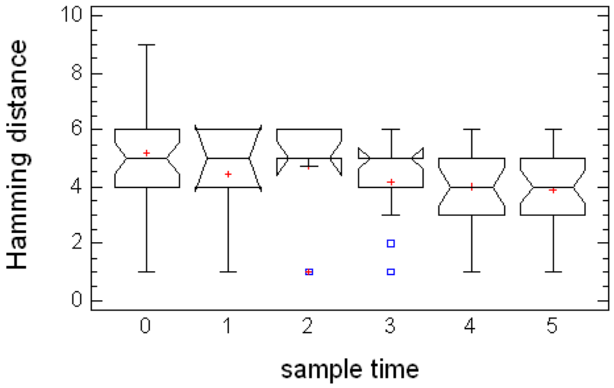

Figure 4.- Robust performance graph in the ant population evolution experiment showing the medians (notches) and means (crosses) of the Hamming distance. Squares and squares including a plus sign indicate suspected outliers and outliers respectively.

The Fig. 5 shows the robust performance graph obtained in the third *study case*, thus the simulation of evolution in Dawkin’s biomorphs (Fig. 6).

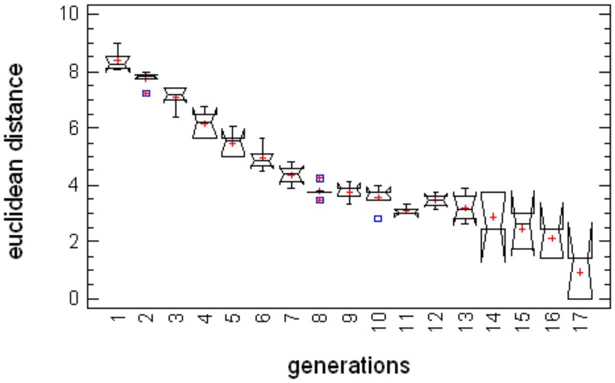

Figure 5.- Robust performance graph in the Dawkin’s biomorphs evolution experiment showing the medians (notches) and means (crosses) of the Euclidean distance. Squares and squares including a plus sign indicate suspected outliers and outliers.

Since each biomorph is defined by a chromosome composed of 8 integer values, the performance graph is based on the Euclidean distance. The performance graph shows how the medians as well as the average of the Euclidean distances per generation (or genetic drift) become smaller as a consequence of the convergence of the population towards a target biomorph. Table VIII summarizes the genetic drift, thus the average of the Euclidean distance computed per generation, the median and the variance of the Euclidean distance, as well as the minimum, maximum, Q_1, Q_3, IQR, and the standardized skewness and standard kurtosis. Note how in this case the different shape of the distributions, thus the different sign of the standardized skewness values, breaks one of the main assumptions in the Kruskal-Wallis test.

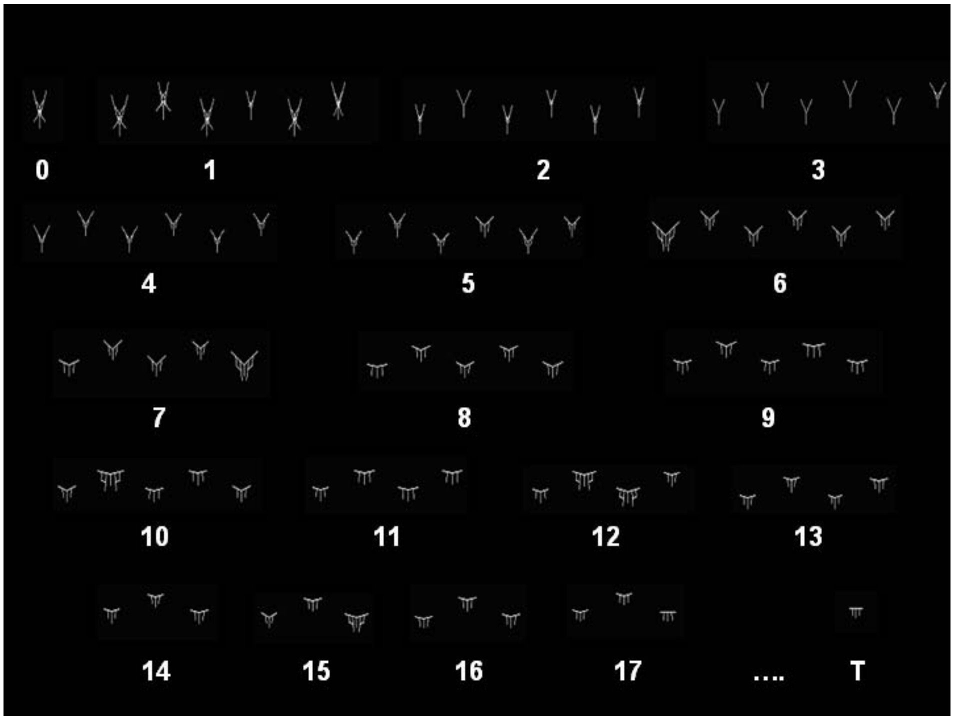

Figure 6.- Dawkin's biomorphs evolution during 17 generations showing T the target biomorph.

Statistical comparisons

Using the SGA program three simulation experiments were carried out (*study case* 4) representing in a Multiple Notched Box-and-Whisker Plot (Fig. 7) the obtained results.

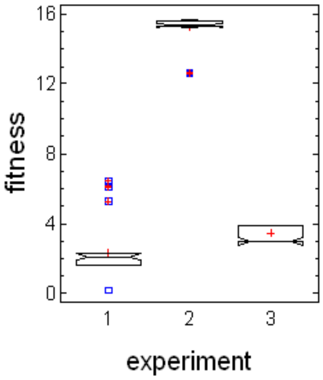

Figure 7.- Multiple Notched Box-and-Whisker Plot showing the medians (notches) and means (crosses) of the fitness in three different SGA experiments. Squares and squares including a plus sign indicate suspected outliers and outliers.

Table IX summarizes the mean, median and variance of the fitness, as well as the minimum, maximum, Q_1, Q_3, IQR, and the standardized skewness and standard kurtosis. A first statistical analysis compares the genetic algorithm experiment with crossover and mutation probabilities of 75% and 5% (first experiment) and the genetic algorithm without mutation and crossover with a probability of 75% (second experiment). Since the standardized skewness and standard kurtosis (Table IX) do no belong to the interval [-2, 2] then it suggests that data do not come from a Normal distribution. In consequence, we carried out a Mann-Whitney (Wilcoxon) test to compare the medians of the two experiments. The Mann-Whitney test (Table X) shows with a *p*-value equal to zero that there is a statistically significant difference between the two medians at the 95.0% confidence level. Finally, we compared the three experiments together, thus the first and second simulation experiments together with a third one consisting in a genetic algorithm with only mutation with a probability equal to 5%. A Kruskal-Wallis test was carried out comparing the medians of the three experiments, with Dunn's post-test for comparison of all pairs. The Kruskal-Wallis test (Table XI) shows with a *p*-value equal to zero that there is a statistically significant difference among medians at the 95.0% confidence level. In agreement with the Dunn test (Table XII) all pairs of experiments were significant, considering the differences with the value $p < 0.05$ statistically significant. The statistical analysis results illustrate the importance and role of the crossover and mutation in an optimization problem.

The statistical analysis of the three TSP simulation experiments performed with ACO algorithm (*study case* 5), was a similar to the protocol for the SGA experiments (*study case* 4). Fig. 8 shows the obtained results represented in a Multiple Notched Box-and-Whisker Plot, and Table XIII summarizes the mean, median and variance of the tour length, as well as the minimum, maximum, Q_1, Q_3, IQR, and the standardized skewness and standard kurtosis. In the case of the second experiment the standardized skewness and standard kurtosis do no belong to the interval [-2, 2] suggesting that data do not come from a Normal distribution.

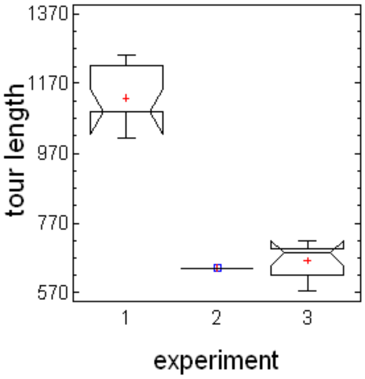

Figure 8.- Multiple Notched Box-and-Whisker Plot showing the medians (notches) and means (crosses) of the tour length in three different TSP experiments carried out with ACO algorithm.

Furthermore, in agreement with Table XIII and the statistical tests we carried out for determining homoscedasticity or variance homogeneity, thus whether significant differences exist among the variances σ_1^2, σ_2^2 and σ_3^2 of the three ACO experiments (Fig. 9), we concluded that variances were very different. Since the *p*-values for Cochran's C test and Bartlett test were both lesser than 0.05, in particular 0.0016 and 0.0000 respectively, we concluded that variance differences were statistically significant. Once again, a Kruskal-Wallis test was carried out comparing the medians of the three experiments, with Dunn's post-test for comparison of all pairs. The Kruskal-Wallis test (Table XIV) shows with a *p*-value equal to zero that there is a statistically significant difference among medians at the 95.0% confidence level. In agreement with the Dunn test (Table XV) the first

TSP tour experiment significantly differs from the second and third TSP tours, whereas the differences observed between the second TSP tour and third one are not significant, considering the differences with the value $p < 0.05$ as statistically significant. No matter such differences, the best tour in the first, second and third simulation experiments were as follows: 5-4-3-2-1-0-9-8-7-6 with a tour length equal to 984.04, 7-5-4-3-0-1-9-2-8-6 with a tour length equal to 641.58, and 9-0-4-2-1-3-7-5-6-8 with a tour length equal to 576.79, respectively.

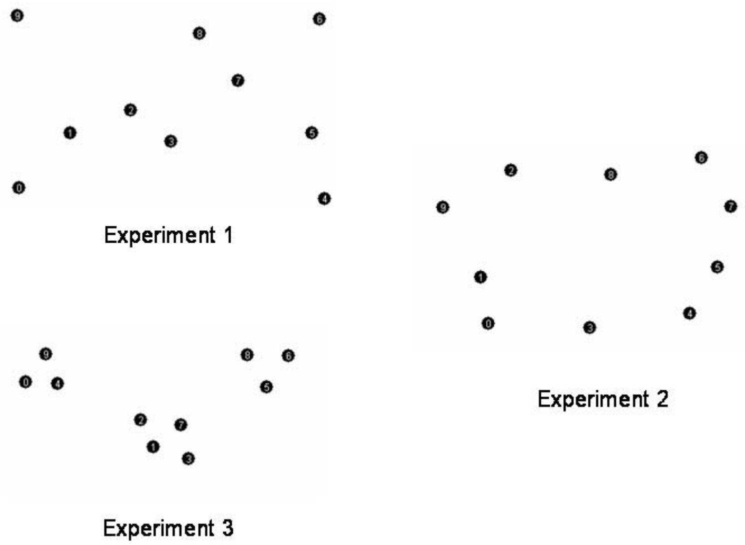

Figure 9.- Three different simulation experiments performed with ACO algorithm of the popular TSP experiment with ten cities labeled from 0 to 9.

The symbolic regression experiment (*study case* 6) illustrates once again the general protocol continued with the SGA and ACO experiments. The Fig. 10 shows the obtained results represented in a Multiple Notched Box-and-Whisker Plot, and Table XVI summarizes the mean, median and variance of the fitness –calculated using the error between the approximated and the target functions-, as well as the minimum, maximum, Q_1, Q_3, IQR, and the standardized skewness and standard kurtosis. Note that the best ‘genetic protocol’ is used in the first simulation experiment (Fig. 10).

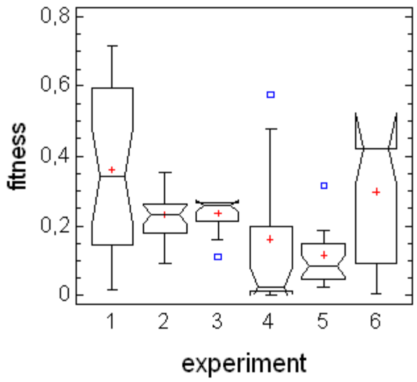

Figure 10.- Multiple Notched Box-and-Whisker Plot showing the medians (notches) and means (crosses) of the fitness in six different simulation experiments carried out with a symbolic regression problem. The problem consisted in the search of an approximation of $3x^4 – 3x + 1$ function. Squares indicates suspected outliers.

Thus, the experiment where the method of selection is the fitness proportionate, and the crossver and mutation probabilities are equal to 80% and 20%, respectively. In this experiment, the best individual was found in generation 406 with a fitness value of 0.7143, being the evolved approximated function:

```
mul(
 sub(
  add(
   add(
    x,
    add(
     add(3.3133816502639917,-3.740923112937817),
     add(3.3133816502639917,-3.740923112937817)
    )
   ),
   x
  ),
  x
 ),
 add(
  add(
   add(
    x,
    add(
     add(3.3133816502639917,-3.740923112937817),
     x
    )
   ),
   add(
    mul(
     x,
     add(
      add(
       add(3.3133816502639917,-3.740923112937817),
       add(
        mul(
         x,
         add(
          mul(
           x,
           mul(
            x,
            mul(x,x)
           )
          ),
          add(
           x,
           add(3.3133816502639917,-
3.740923112937817)
          )
         )
        ),
        x
       )
      ),
      add(
       x,
       add(3.3133816502639917,-3.740923112937817)
      )
     )
    ),
    x
   )
  ),
  add(
   add(
    x,
    add(3.3133816502639917,-3.740923112937817)
   ),
   add(
    mul(
     x,
     add(
      mul(
       add(
        x,
        add(3.3133816502639917,-3.740923112937817)
       ),
       add(
        mul(x,x),
        add(
         x,
         add(3.3133816502639917,-
3.740923112937817)
        )
       )
      ),
      x
     )
    ),
    add(
     mul(
      x,
      add(3.3133816502639917,-3.740923112937817)
     ),
     add(3.3133816502639917,-3.740923112937817)
    )
   )
  )
 )
)
```

In agreement with the Kruskal-Wallis test (Table XVII) the obtained *p*-value (0.0000) means that the medians for the six experiments were significantly different. Since in the Dunn's post-test (Table XVIII) for comparison of all pairs of simulation experiments we only obtained significant differences between the first and fourth experiments, and first and fifth experiments, as well as between the fourth and sixth experiments, and fifth and sixth experiments, then we reach the following conclusion. Even when the first, fourth and fifth experiments were carried out with a crossover probability of 80%, the first experiment significantly differs of the other two simulations because in the experiments fourth and fifth the method of selection is the tournament approach instead the fitness proportionate. Surprisingly, when the method of selection is the tournament approach, the best performance is obtained when crossover has a low probability, only a 5% in the sixth simulation experiment, with a high mutation rate of 20% differing significantly from the fourth and fifth experiments where crossover had a high probability of 80%.

Conclusion

This tutorial shows how assuming that EC simulation experiments result in biased samples with non-normal, highly skewed, and asymmetric distributions, the statistical analysis of data could be carried out combining Exploratory Data Analysis with nonparametric tests for comparing two or more medians. In fact and except for the initial random population, the normality assumption was only fulfilled in two examples (see http://bioinformatica.net/tests/survey.html): in the symbolic regression problem (s*tudy case* 6) and in the Darwkin’s biomorphs experiment (s*tudy case* 3). Note that for *case* 3 normality was accomplished once data were transformed with the logarithmic transformation. In both examples the Kolmogorov-Smirnov *p*-value was greater than 0.01. The use of nonparametric tests combined with a Multiple Notched Box-and-Whisker Plot is a good option, bearing in mind that we only considered two experimental situations that are common in EC practitioners: the performance evaluation of an algorithm and the multiple experiments comparison. The different approaches have been illustrated with different examples chosen from Evolutionary Computation and the related field of Artificial Life.

Bibliography

[Benjamini, 1988] Y. Benjamini. 1988. Opening the box of a boxplot. The American Statistician 42: 257-262.

[Czarn et al., 2004] A. Czarn, C. MacNish, K. Vijayan, B. Turlach, R. Gupta. 2004. Statistical exploratory analysis of genetic algorithms. IEEE Transactions on Evolutionary Computation 8: 405-421.

[Davis, 1991] L. Davis (Ed.). 1991. Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold.

[Dawkins, 1986] R. Dawkins. 1986. The Blind Watchmaker. New York: W. W. Norton & Company.

[Dorigo and Gambardella, 1997] M. Dorigo, L.M. Gambardella. 1997. Ant colonies for the travelling salesman problem. BioSystems 43: 73-81.

[Draper ans Smith, 1998] N.R. Draper, H. Smith. 1998. Applied regression analysis (Third Edition). New York: Wiley.

[Dunn, 1964] O.J. Dunn. 1964. Multiple comparisons using rank sums. Technometrics 6: 241-252.

[Frederick et al., 1993] W.G. Frederick, R.L. Sedlmeyer, C.M. White. 1993. The Hamming metric in genetic algorithms and its application to two network problems. In: Proceedings of the 1993 ACM/SIGAPP Symposium on Applied Computing: States of the Art and Practice, Indianapolis, Indiana: 126-130.

[Frigge et al., 1989] M. Frigge, D.C. Hoaglin, B. Iglewicz. 1989. Some implementations of the boxplot. The American Statistician 43: 50-54.

[Gerber, 1998] H.Gerber. 1998. Simple Symbolic Regression Using Genetic Programming à la John Koza. http://alphard.ethz.ch/gerber/approx/default.html

[Gibbons and Chakraborti, 2003] J.D. Gibbons, S. Chakraborti. 2003. Nonparametric Statistical Inference. New York: Marcel Dekker.

[Goldberg, 1989] D.E. Goldberg. 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley.

[Hart, 2001] A. Hart. 2001. Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ 323: 391-393.

[Hochberg and Tamhane, 1987] Y. Hochberg, A.C. Tamhane. 1987. Multiple Comparison Procedures. New York: John Wiley & Sons.

[Holland, 1992] J.H. Holland. 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Cambridge: The MIT Press. (Reprint edition 1992, originally published in 1975).

[Hollander and Wolfe, 1999] M. Hollander, D.A. Wolfe. 1999. Nonparametric Statistical Methods. New York: Wiley.

[Koza, 1992] J.R. Koza. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA: MIT Press.

[Koza, 1994] J.R. Koza. 1994. Genetic Programming II: Automatic Discovery of Reusable Programs. Cambridge, MA: MIT Press.

[Kruskal and Wallis, 1952] W.H. Kruskal, W.A. Wallis. 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47: 583-621.

[Lahoz-Beltra, 2001] R. Lahoz-Beltra. 2001. Evolving hardware as model of enzyme evolution. BioSystems 61: 15-25.

[Lahoz-Beltra, 2004] R. Lahoz-Beltra. 2004. Bioinformática: Simulación, Vida Artificial e Inteligencia Artificial. Madrid: Ediciones Diaz de Santos. (Transl.: Spanish).

[Lahoz-Beltra, 2008] R. Lahoz-Beltra. 2008. ¿Juega Darwin a los Dados? Madrid: Editorial NIVOLA. (Transl.: Spanish).

[Manly, 1985] B.F.J. Manly. 1985. The Statistics of Natural Selection on Animal Populations. London: Chapman and Hall.

[Mann and Whitney, 1947] H.B. Mann, D.R. Whitney. 1947. On a test of whether one or two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18: 50-60.

[McGill eat al., 1978] R. McGill, W.A. Larsen, J.W. Tukey. 1978. Variations of boxplots. The American Statistician 32: 12-16.

[Mühlenbein and Schlierkamp-Voosen, 1993] H. Mühlenbein, D. Schlierkamp-Voosen. 1993. Predictive models for the breeder genetic algorithm I: Continuous parameter optimization. Evolutionary Computation 1: 25-49.

[Perales-Gravan and Lahoz-Beltra, 2004] C. Perales-Gravan, R. Lahoz-Beltra. 2004. Evolving morphogenetic fields in the zebra skin pattern based on Turing’s morphogen hypothesis. Int. J. Appl. Math. Comput. Sci. 14: 351-361.

[Perales-Gravan and Lahoz-Beltra, 2008] C. Perales-Gravan, R. Lahoz-Beltra. 2008. An AM radio receiver designed with a genetic algorithm based on a bacterial conjugation operator. IEEE Transactions on Evolutionary Computation 12(2): 129-142.

[Prata, 1993] Prata, S. (1993). Artificial Life Playhouse: Evolution at Your Fingertips (Disk included). Corte Madera, CA: Waite Group Press.

[Siegel and Castellan, 1988] S. Siegel, N.J.. Castellan. 1988. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill.

[Sinclair, 2006] M.C. Sinclair. 2006. Ant Colony Optimization. http://uk.geocities.com/ markcsinclair/aco.html

[Tukey, 1977] J.W. Tukey. 1977. Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Authors' Information

Rafael Lahoz-Beltra *–Chairman, Dept. of Applied Mathematics, Faculty of Biological Sciences, Complutense University of Madrid, 28040 Madrid, Spain ; e-mail: lahozraf@bio.ucm.es*

Major Fields of Scientific Research: evolutionary computation, embryo development modeling and the design of bioinspired algorithms

Carlos Perlaes Gravan *– Member at Bayes Forecast; e-mail: soyperales@gmail.com*

Major Fields of Scientific Research: design and development of artificial intelligence methods and its application to decision making processes, optimization and forecasting as well as the modeling and simulation of biological systems