Statistical Analysis of Experiments in Data Mining and Computational Intelligence

How must I conduct statistical comparisons in my Experimental Study? On the use of Nonparametric Tests and Case Studies.

Salvador García, Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI^{2}S)
Dept. of Computer Science and A.I.
University of Granada, Spain

Emails: sglopez@ujaen.es, herrera@decsai.ugr.es
http://sci2s.ugr.es

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Motivation

The experimental analysis on the performance of a new method is a crucial and necessary task to carry out in a research.

Deciding when an algorithm is better than other one may not be a trivial task.

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Motivation Example for classification

Large Variations in Accuracies of Different Classifiers

Is really the alg.3 the best performing one because it obtains the best average value?

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	**81.2**	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	**85.7**
bal	45.0	76.2	87.2	**90.4**	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	**68.2**
bps	51.6	83.2	82.8	78.6	80.1	79.0	**83.3**
bre	65.5	96.0	**96.7**	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	**72.6**
h-c	54.5	77.4	83.2	**83.6**	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	**95.3**	**95.3**	94.7	**95.3**	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	**95.2**	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	**74.9**	75.1	74.8
lym	55.0	83.3	83.6	**85.6**	77.0	71.5	79.0
mmg	56.0	63.0	**65.3**	64.7	64.8	61.9	63.4
mus	51.8	**100.0**	**100.0**	96.4	**100.0**	**100.0**	99.8
mux	49.9	78.6	99.8	61.9	99.9	**100.0**	**100.0**
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	**50.8**	41.6	39.8	43.7
seg	14.3	**97.4**	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	**98.4**	97.0	96.7
soyb	13.5	89.5	90.3	**92.8**	91.4	90.3	76.2
tao	49.8	**96.1**	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	**92.1**	**92.1**	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	**96.5**	95.4
vow	9.1	99.1	**96.6**	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	**97.8**	94.6	92.9	96.3
zoo	41.7	94.6	92.5	**95.4**	91.6	92.5	92.6
Avg	44.8	80.0	82.4	78.0	82.1	81.8	81.7

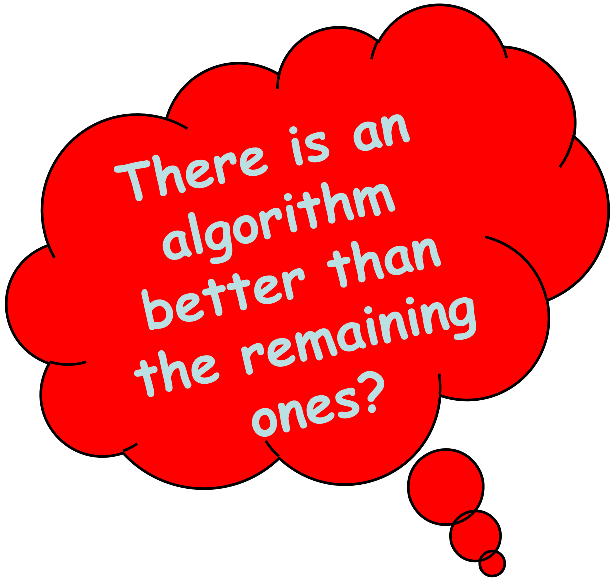
Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Motivation

Alg. 4 is the winner in 8 problems with average 78.0

Alg. 2 is the winner for 4 problems with average 80.0

What is the best between both?


There is an algorithm better than the remaining ones?

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	**81.2**	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	**85.7**
bal	45.0	76.2	87.2	**90.4**	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	**68.2**
bps	51.6	83.2	82.8	78.6	80.1	79.0	**83.3**
bre	65.5	96.0	**96.7**	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	**72.6**
h-c	54.5	77.4	83.2	**83.6**	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	**95.3**	**95.3**	94.7	**95.3**	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	**95.2**	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	**74.9**	75.1	74.8
lym	55.0	83.3	83.6	**85.6**	77.0	71.5	79.0
mmg	56.0	63.0	**65.3**	64.7	64.8	61.9	63.4
mus	51.8	**100.0**	**100.0**	96.4	**100.0**	**100.0**	99.8
mux	49.9	78.6	99.8	61.9	99.9	**100.0**	**100.0**
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	**50.8**	41.6	39.8	43.7
seg	14.3	**97.4**	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	**98.4**	97.0	96.7
soyb	13.5	89.5	90.3	**92.8**	91.4	90.3	76.2
tao	49.8	**96.1**	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	**92.1**	**92.1**	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	**96.5**	95.4
vow	9.1	99.1	**96.6**	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	**97.8**	94.6	92.9	96.3
zoo	41.7	94.6	92.5	**95.4**	91.6	92.5	92.6
Avg	**44.8**	**80.0**	**82.4**	**78.0**	**82.1**	**81.8**	**81.7**

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Motivation

We must use statistical tests for comparing the algorithms.

The problem:

How must I do the statistical experimental study?

What tests must I use?

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	**81.2**	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	**85.7**
bal	45.0	76.2	87.2	**90.4**	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	**68.2**
bps	51.6	83.2	82.8	78.6	80.1	79.0	**83.3**
bre	65.5	96.0	**96.7**	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	**72.6**
h-c	54.5	77.4	83.2	**83.6**	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	**95.3**	**95.3**	94.7	**95.3**	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	**95.2**	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	**74.9**	75.1	74.8
lym	55.0	83.3	83.6	**85.6**	77.0	71.5	79.0
mmg	56.0	63.0	**65.3**	64.7	64.8	61.9	63.4
mus	51.8	**100.0**	**100.0**	96.4	**100.0**	**100.0**	99.8
mux	49.9	78.6	99.8	61.9	99.9	**100.0**	**100.0**
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	**50.8**	41.6	39.8	43.7
seg	14.3	**97.4**	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	**98.4**	97.0	96.7
soyb	13.5	89.5	90.3	**92.8**	91.4	90.3	76.2
tao	49.8	**96.1**	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	**92.1**	**92.1**	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	**96.5**	95.4
vow	9.1	99.1	**96.6**	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	**97.8**	94.6	92.9	96.3
zoo	41.7	94.6	92.5	**95.4**	91.6	92.5	92.6
Avg	44.8	80.0	82.4	78.0	82.1	81.8	81.7

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Objective

To show some results on the use of statistical tests

(nonparametric tests) for comparing algorithms.

We will not discuss the performance measures that can be used neither the choice on the set of benchmarks.

Some guidelines on the use of appropriate nonparametric tests depending on the situation will be given

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
  - Theoretical background
  - Checking the conditions in Parameter Optimization Experiments
- Basic non-parametric tests and case studies:
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Evolutionary Algorithms: CEC’05 Special Session on Parameter Optimization
- Lessons Learned
  - Recommendations on the use of nonparametric tests
  - Frequent Questions
- Books of Interest and References
- Software

OUTLINE

- **Introduction to Inferential Statistics**
- **Conditions for the safe use of parametric tests**
  - **Theoretical background**
  - **Checking the conditions in Parameter Optimization Experiments**
- **Basic non-parametric tests and case studies:**
  - **For Pairwise Comparisons**
  - **For Multiple Comparisons involving control method**
  - **Evolutionary Algorithms: CEC’05 Special Session on Parameter Optimization**
- **Lessons Learned**
  - **Recommendations on the use of nonparametric tests**
  - **Frequent Questions**
- **Books of Interest,References and Software**
- **Software**

Introduction to Inferential Statistics

Inferential Statistics

provide measures of how well your data (results of experiments) support your hypothesis and if your data are generalizable beyond what was tested (*significance tests*)

For example: Comparing two or various sets of experiments in a computational problem.

Parametric versus Nonparametric Statistics – When to use them and which is more powerful?

Introduction to Inferential Statistics

What is an hypothesis?

a prediction about a single population or about the relationship between two or more populations.

Hypothesis testing is a procedure in which sample data are employed to evaluate a hypothesis.

The null hypothesis is a statement of no effect or no difference and it is expected to be rejected by the experimenter.

Introduction to Inferential Statistics

Examples of Null-Hypothesis

H_o: The 2 samples come from populations with the same distributions.

Or,

median of population 1 = median of population 2

(generalization with n samples)

Significance level α

- **It is a confidence threshold that informs us whether or not to reject the null hypothesis.**
- **It must be pre-defined by the experimenter and a significance level of 90% (0.1) or 95% (0.05) is usually used, also 99% (0.01).**

Introduction to Inferential Statistics

Significance level α

- If you decide for a significance level of 0.05 (95% certainty that there indeed is a significant difference), then a **p-value** (datum provided by the test) smaller than 0.05 indicates that you can reject the **null-hypothesis.**

- **Remember:** the null-hypothesis generally is associated to an hypothesis of equality or equivalence (equal means or distributions).

- So, if a test obtains p = 0.07, it means that you **cannot reject** the null hypothesis of equality ⇨ **there is no significant differences in the analysis conducted**

Introduction to Inferential Statistics

p-value

- Instead of stipulating a priori level of significance ɑ, one could calculate the smallest level of significance that results in the rejection of the null hypothesis.

- **This is the p-value, it provides information about "how significant" the result is.**

- **It does it without commiting to a particular level of significance.**

Introduction to Inferential Statistics

There is at least one nonparametric test equivalent to a basic parametric test

- **Compare two variables**
- **If more than two variables**

Parametric	Nonparametric
t-test	Sign test
	Wilcoxon signed rank test
ANOVA and derivatives	Friedman test and more…
Tukey, Tamhane, ...	Bonferroni-Dunn, Holm, etc…

Introduction to Inferential Statistics

Parametric Assumptions

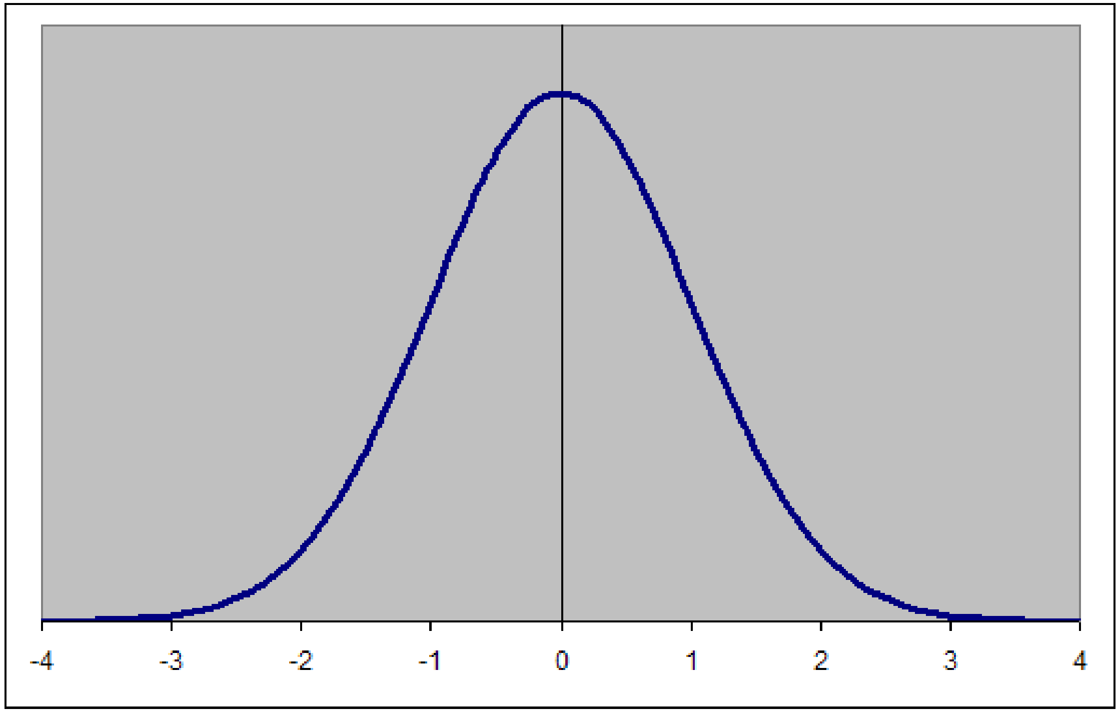
(t-test, ANOVA, …)

- The observations must be independent
- Normality: The observations must be drawn from normally distributed populations
- Homoscedasticity: These populations must have the same variances

Introduction to Inferential Statistics

Normality Tip

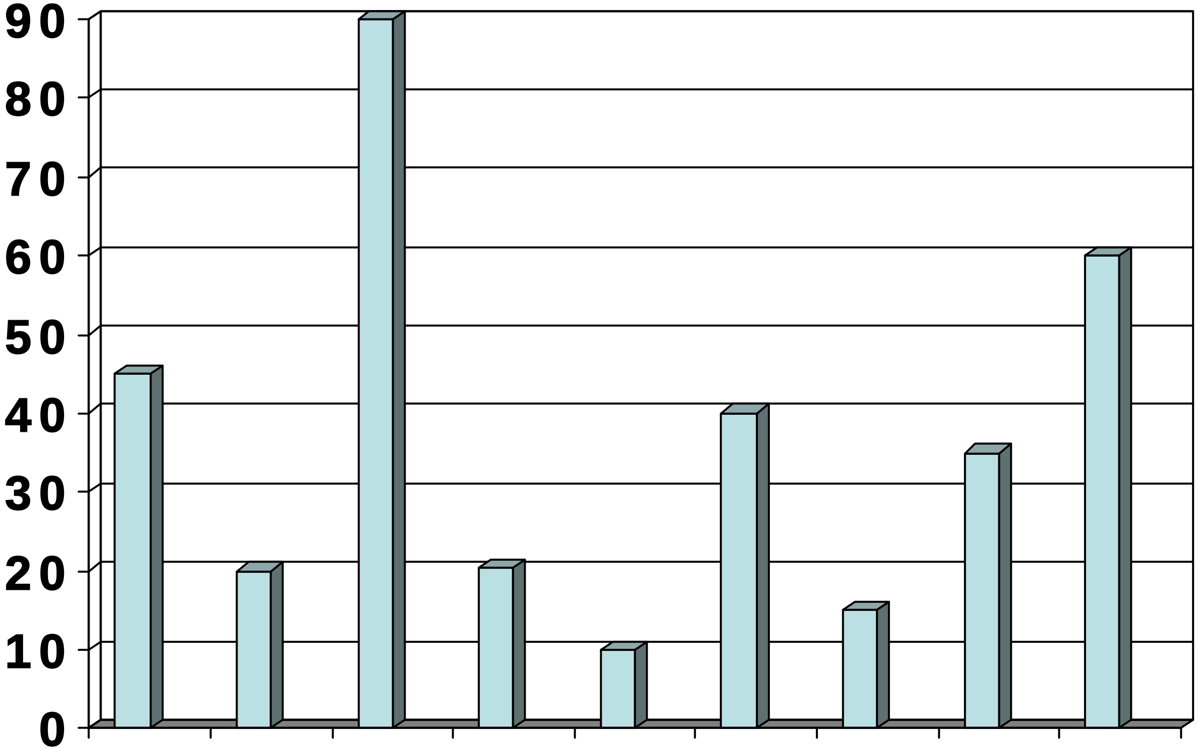
If a histogram representing your data looks like this, you can conduct a parametric test!



Introduction to Inferential Statistics

Otherwise, don't conduct a parametric test!

The conclusions could be erroneous



Histogram

Nonparametric Assumptions

(Wilcoxon, Friedman, …)

- The observations must be independent

- The data must be represented by ordinal numbering.

How do nonparametric tests work?

❑ Most nonparametric tests use ***ranks*** instead of raw data for their hypothesis testing.

❑ They apply a transformation procedure in order to obtain ranking data.

OUTLINE

- Introduction to Inferential Statistics
- **Conditions for the safe use of parametric tests**
  - **Theoretical background**
  - **Checking the conditions in Parameter Optimization Experiments**
- Basic non-parametric tests and case studies:
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Evolutionary Algorithms: CEC’05 Special Session on Parameter Optimization
- Lessons Learned
  - Recommendations on the use of nonparametric tests
  - Frequent Questions
- Books of Interest and References
- Software

Conditions for the safe use of parametric tests

- Theoretical background
- Checking the conditions in Parameter Optimization Experiments

Conditions for the Safe Use of Parametric Tests

Theoretical Background

The distinction between parametric and nonparametric test is based on the level of measure represented by the data which will be analyzed.

A parametric test is able to use data composed by real values: But when we dispose of this type of data, we should not always use a parametric test.

There are some assumptions for a safe usage of parametric tests ad the non fulfillment of these conditions might cause a statistical analysis to lose credibility.

Conditions for the Safe Use of Parametric Tests

Theoretical Background

In order to use the parametric tests, is necessary to check the following conditions:

Independence: In statistics, two events are independent when the fact that one occurs does not modify the probability of the other one occurring.

- **When we compare two optimization algorithms they are usually independent.**
- **When we compare two machine learning methods, it depends on the partition:**
  - **The independency is not truly verified in 10-fcv (a portion of samples is used either for training and testing in different partitions.**
  - **Hold out partitions can be safely take as independent, since training and test partitions do not overlap.**

Conditions for the Safe Use of Parametric Tests

Theoretical Background

Parametric tests assume that the data are taken from normal distributions

Normality: An observation is normal when its behaviour follows a normal or Gauss distribution with a certain value of average μ and variance σ. A normality test applied over a sample can indicate the presence or absence of this condition in observed data.

- **Kolmogorov-Smirnov**
- **Shapiro-Wilk**
- **D'Agostino-Pearson**

Conditions for the Safe Use of Parametric Tests

Theoretical Background

Kolmogorov-Smirnov: It compares the accumulated distribution of observed data with the accumulated Gaussian distribution expected.

Shapiro-Wilk: It analyzes the observed data to compute the level of symmetry and kurtosis (shape of the curve) in order to compute the difference with respect to a Gaussian distribution afterwards.

D'Agostino-Pearson: It computes the skewness and kurtosis to quantify how far from the Gaussian distribution is in terms of asymmetry and shape.

Conditions for the Safe Use of Parametric Tests

Theoretical Background

Heteroscedasticity: This property indicates the existence of a violation of the hypothesis of equality of variances.

Levene’s test is used for checking if k samples present or not this homogeneity of variances (homoscedasticity).

Conditions for the safe use of parametric tests

- Theoretical background
- Checking the conditions in Parameter Optimization Experiments

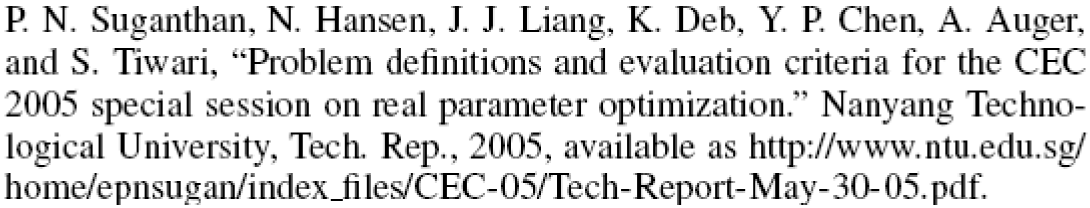
Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

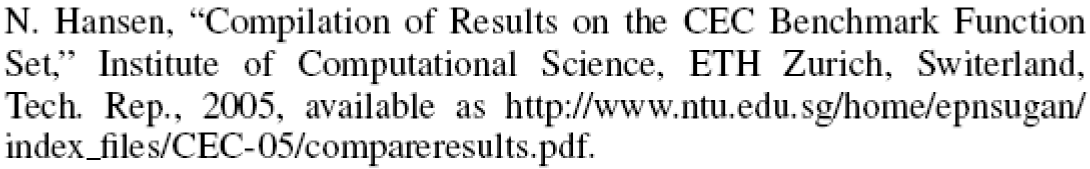
Special Session on Real-Parameter Optimization at CEC-05, Edinburgh, UK, 2-5 Sept. 2005

25 functions with real parameters, 10 variables:

f1-f5 unimodal functions   f6-f25 multimodal functions



P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y. P. Chen, A. Auger, and S. Tiwari, "Problem definitions and evaluation criteria for the CEC 2005 special session on real parameter optimization." Nanyang Technological University, Tech. Rep., 2005, available as http://www.ntu.edu.sg/home/epnsugan/index_files/CEC-05/Tech-Report-May-30-05.pdf.



N. Hansen, "Compilation of Results on the CEC Benchmark Function Set," Institute of Computational Science, ETH Zurich, Switerland, Tech. Rep., 2005, available as http://www.ntu.edu.sg/home/epnsugan/index_files/CEC-05/compareresults.pdf.

Source: S. García, D. Molina, M. Lozano, 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 15 (2009) 617-644, doi: 10.1007/s10732-008-9080-4*.

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

- Algorithms involved in the comparison:
  - **BLX-GL50 (Garcia-Martinez & Lozano, 2005):** Hybrid Real-Coded Genetic Algorithms with Female and Male Differentiation
  - **BLX-MA (Molina *et al.*, 2005):** Adaptive Local Search Parameters for Real-Coded Memetic Algorithms
  - **CoEVO (Posik, 2005):** Mutation Step Co-evolution
  - **DE (Ronkkonen *et al.*,2005):**Differential Evolution
  - **DMS-L-PSO:** Dynamic Multi-Swarm Particle Swarm Optimizer with Local Search
  - **EDA (Yuan & Gallagher, 2005):** Estimation of Distribution Algorithm
  - **G-CMA-ES (Auger & Hansen, 2005):** A restart Covariance Matrix Adaptation Evolution Strategy with increasing population size
  - **K-PCX (Sinha *et al.*, 2005):** A Population-based, Steady-State real-parameter optimization algorithm with parent-centric recombination operator, a polynomial mutation operator and a niched -selection operation.
  - **L-CMA-ES (Auger & Hansen, 2005):** A restart local search Covariance Matrix Adaptation Evolution Strategy
  - **L-SaDE (Qin & Suganthan, 2005):** Self-adaptive Differential Evolution algorithm with Local Search
  - **SPC-PNX (Ballester *et al.*,2005):** A steady-state real-parameter GA with PNX crossover operator

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

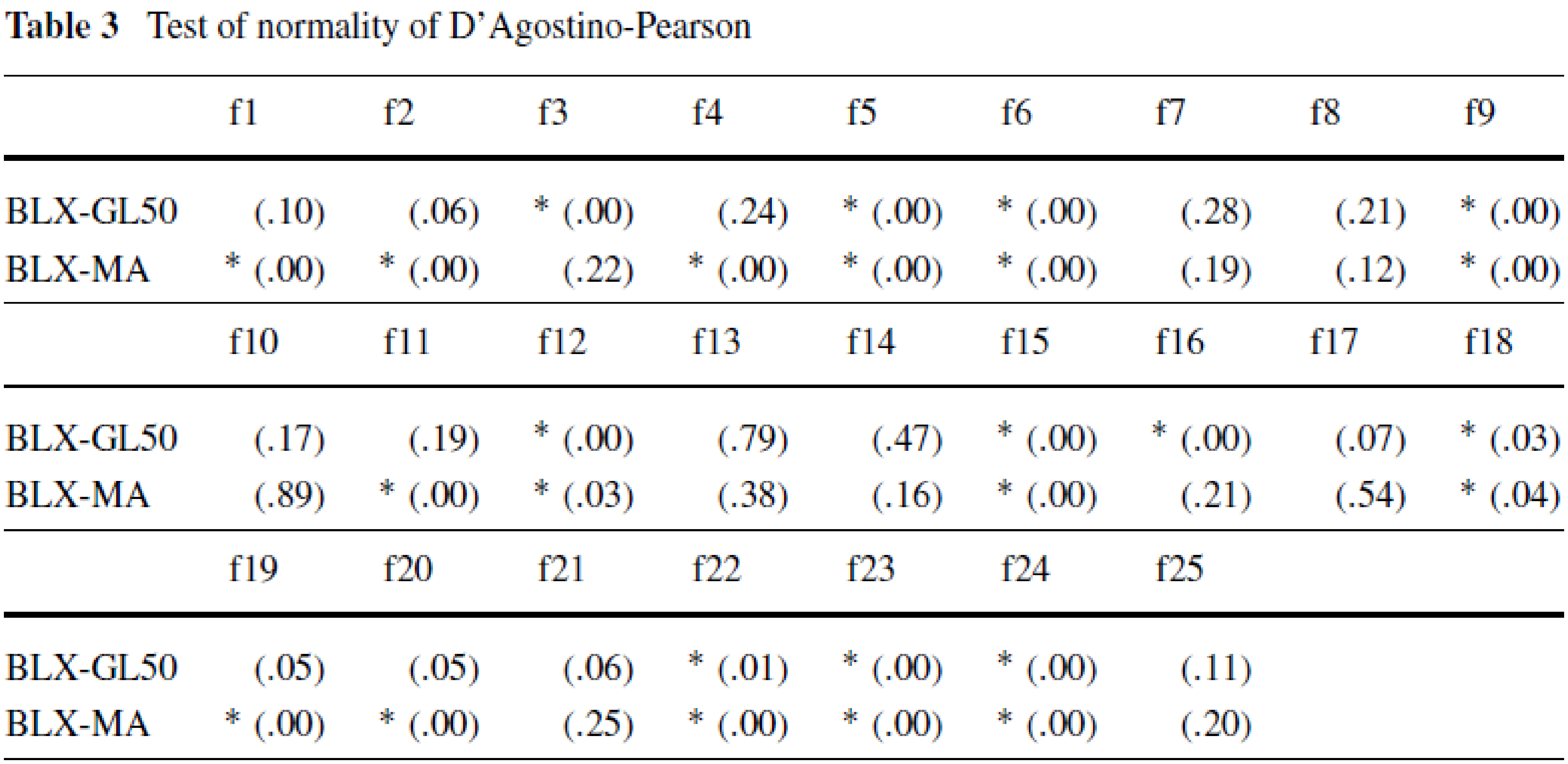


Table 3 Test of normality of D'Agostino-Pearson

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25
BLX-GL50	(.10)	(.06)	* (.00)	(.24)	* (.00)	* (.00)	(.28)	(.21)	* (.00)	(.17)	(.19)	* (.00)	(.79)	(.47)	* (.00)	* (.00)	(.07)	* (.03)	(.05)	(.05)	(.06)	* (.01)	* (.00)	* (.00)	(.11)
BLX-MA	* (.00)	* (.00)	(.22)	* (.00)	* (.00)	* (.00)	(.19)	(.12)	* (.00)	(.89)	* (.00)	* (.03)	(.38)	(.16)	* (.00)	(.21)	(.54)	* (.04)	* (.00)	* (.00)	(.25)	* (.00)	* (.00)	* (.00)	(.20)

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

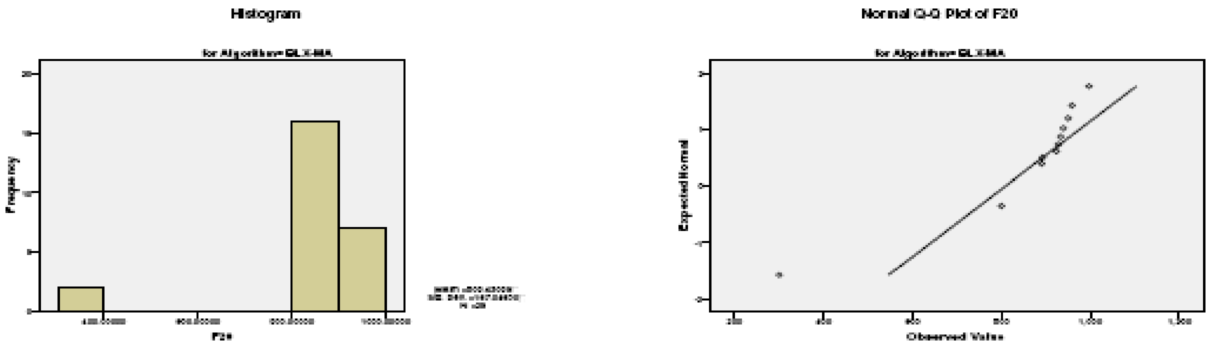


Figure 1: Example of non-normal distribution: Function f20 and BLX-GL50 algorithm: Histogram and Q-Q Graphic.

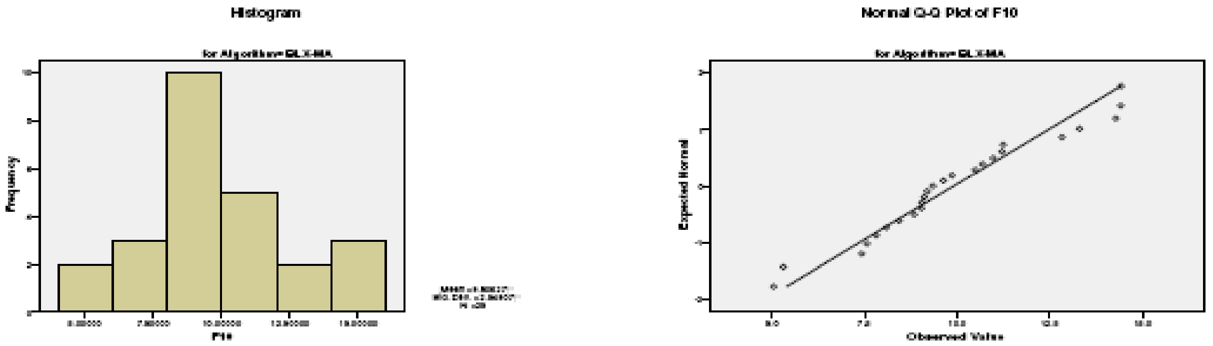


Figure 2: Example of normal distribution: Function f10 and BLX-MA algorithm: Histogram and Q-Q Graphic.

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

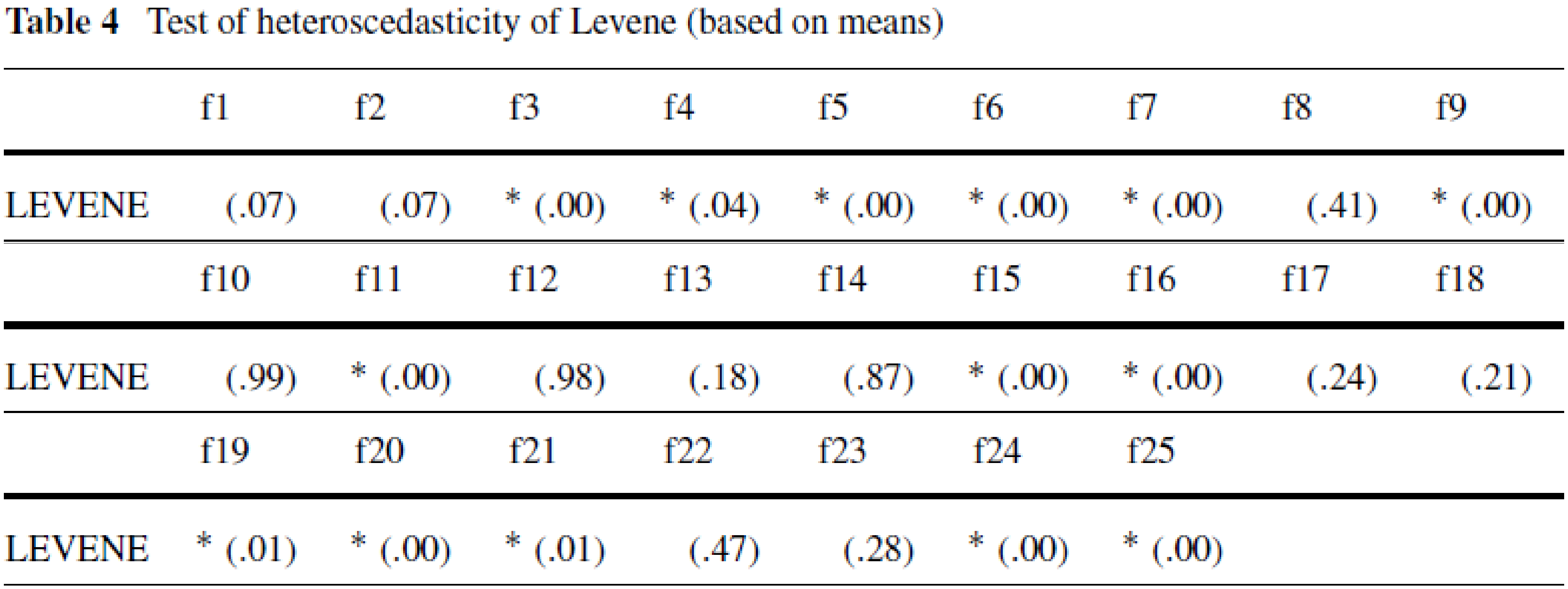


Table 4 Test of heteroscedasticity of Levene (based on means)

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25
LEVENE	(.07)	(.07)	* (.00)	* (.04)	* (.00)	* (.00)	* (.00)	(.41)	* (.00)	(.99)	* (.00)	(.98)	(.18)	(.87)	* (.00)	* (.00)	(.24)	(.21)	* (.01)	* (.00)	* (.01)	(.47)	(.28)	* (.00)	* (.00)

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
  - Theoretical background
  - Checking the conditions in Parameter Optimization Experiments
- **Basic non-parametric tests and case studies:**
  - **For Pairwise Comparisons**
  - **For Multiple Comparisons involving control method**
  - **Evolutionary Algorithms: CEC'05 Special Session on Parameter Optimization**
- Lessons Learned
  - Considerations on the use of nonparametric tests
  - Recommendations on the use of nonparametric tests
  - Frequent Questions
- Books of Interest, References and Software

Basic Non-Parametric Tests and Case Studies

- For Pairwise Comparisons
- For Multiple Comparisons involving a Control Method
- Evolutionary Algorithms: CEC’05 Special Session of Parameter Optimization

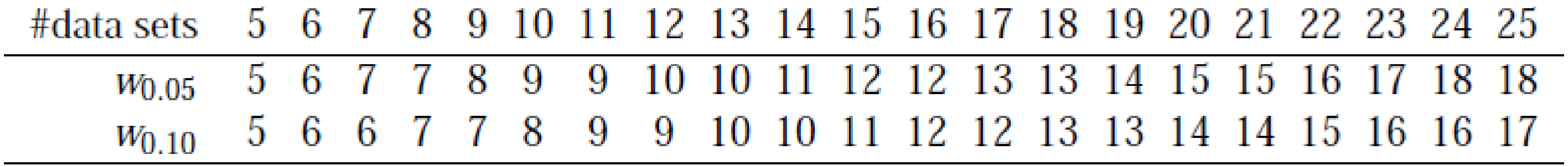
Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

Count of Wins, Losses and Ties: The Sign Test

It a classic form of inferential statistics that use the binomial distribution. If two algorithms compared are, assumed under the null-hypothesis, equivalent, each should win approximately N/2 out of N datasets/problems.

The number of wins are distributed following a binomial distribution. The critical number of wins are presented in the following Table for α=0.05 and α=0.1:



#data sets	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$w_{0.05}$	5	6	7	7	8	9	9	10	10	11	12	12	13	13	14	15	15	16	17	18	18
$w_{0.10}$	5	6	6	7	7	8	9	9	10	10	11	12	12	13	13	14	14	15	16	16	17

Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

Wilcoxon Signed-Ranks Test for Paired Samples

The Wilcoxon Signed-Ranks test is used in exactly the same situations as the paired t-test (i.e., where data from two samples are paired).

In general, the Test asks:

H_o: The 2 samples come from populations with the same distributions. Or, median of population 1 = median of population 2

The test statistic is based on ranks of the differences between pairs of data.

NOTE: If you have ≤ 5 pairs of data points, the Wilcoxon Signed-Ranks test can never report a 2-tailed p-value < 0.05

Example of the Wilcoxon Signed-Ranks Test

dataset	C4.5	C4.5m	Difference	Rank
Adult	0.763	0.768	+0.005	3.5
Breast	0.599	0.591	-0.008	7
Wisconsin	0.954	0.971	+0.017	9
Cmc	0.628	0.661	+0.033	12
Ionosphere	0.882	0.888	+0.006	5
Iris	0.936	0.931	-0.005	3.5
Bupa	0.661	0.668	+0.007	6
Lung	0.583	0.583	0.000	1.5
Lymphograph	0.775	0.838	+0.063	14
Mushroom	1.000	1.000	0.000	1.5
Tumor	0.940	0.962	+0.022	11
Rheum	0.619	0.666	+0.047	13
Voting	0.972	0.981	+0.009	8
Wine	0.957	0.978	+0.021	10

R^+ = 3.5 + 9 + 12 + 5 + 6+ 14+ 11 + 13 + 8 + 10 + 1.5 = 93

R^- = 7 + 3.5 + 1.5 = 12

Example of the Wilcoxon Signed-Ranks Test

$R^+ = 3.5 + 9 + 12 + 5 +$

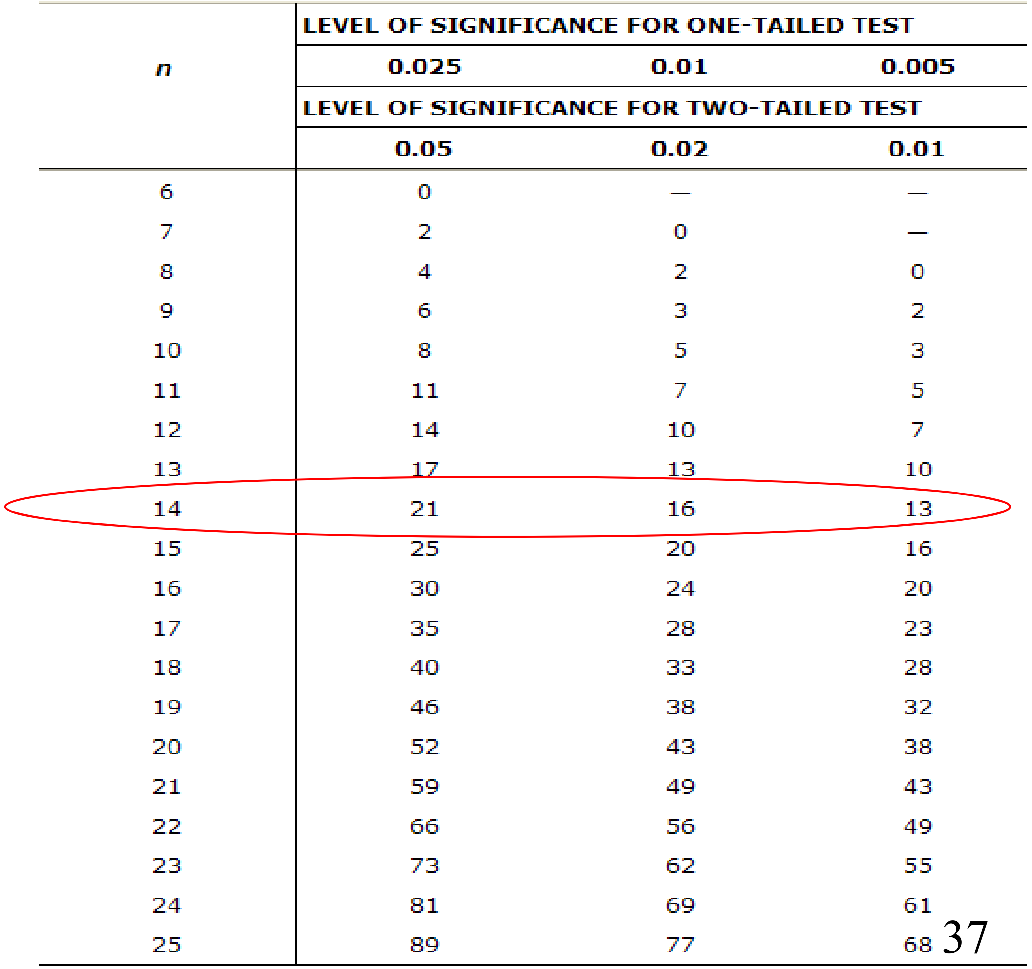
$6 + 14 + 11 + 13 +$

$8 + 10 + 1.5 = 93$

$R^- = 7 + 3.5 + 1.5 = 12$

$T = \text{Min}\ \{R^+, R^-\} = 12$

$\alpha = 0.05,\ N = 14 \quad \text{dif} = 21$

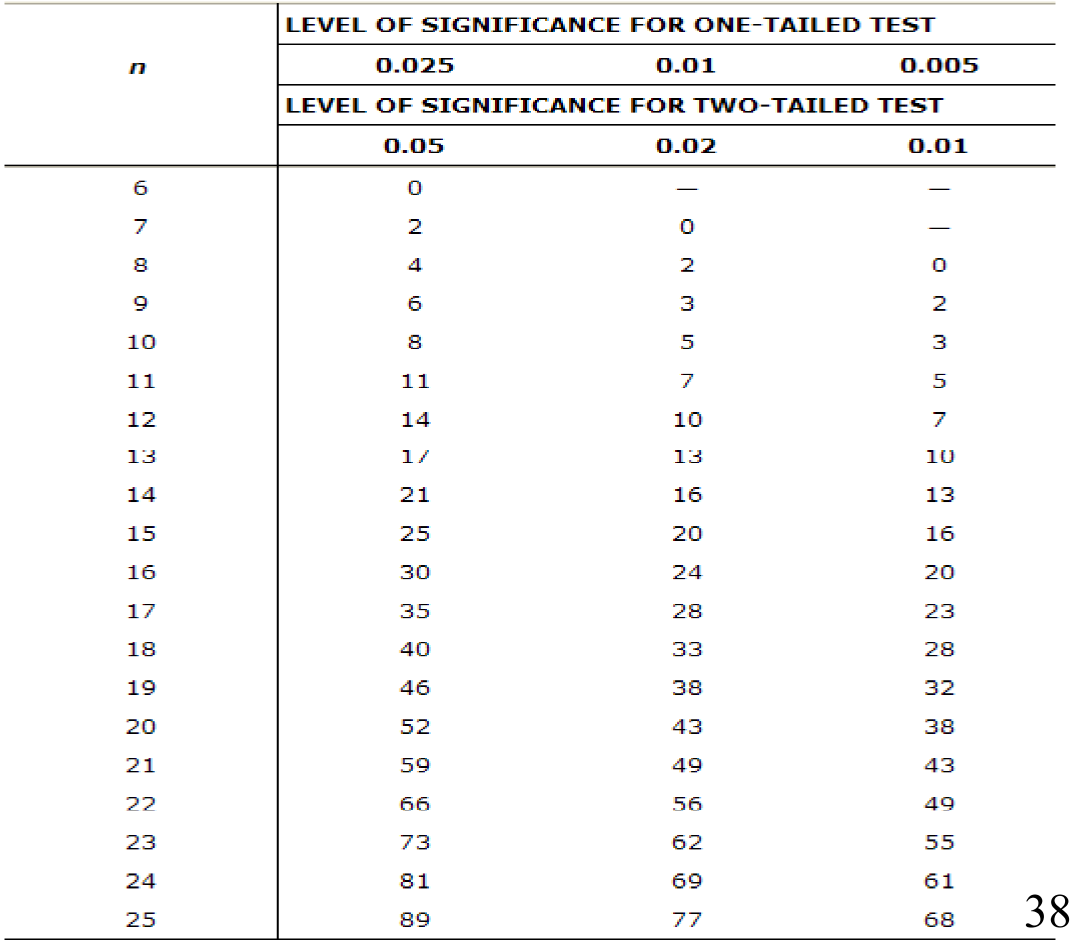


n	LEVEL OF SIGNIFICANCE FOR ONE-TAILED TEST		
	0.025	0.01	0.005
	LEVEL OF SIGNIFICANCE FOR TWO-TAILED TEST		
	0.05	0.02	0.01
6	0	—	—
7	2	0	—
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Example of the Wilcoxon Signed-Ranks Test

Critical value for T for N up to 25.

It T <= dif (table-value) then Reject the H_o



n	LEVEL OF SIGNIFICANCE FOR ONE-TAILED TEST		
	0.025	0.01	0.005
	LEVEL OF SIGNIFICANCE FOR TWO-TAILED TEST		
	0.05	0.02	0.01
6	0	—	—
7	2	0	—
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

For n ≤ 30: use T values (and refer to a Table B.12. Critical Values of the Wilcoxon T Distribution, Zar, App 101)

For n > 30: use z-scores (z is distributed approximately normally).

(and refer to the z-Table, Table B.2. Zar – Proportions of the Normal Curve (One-tailed), App 17)

Where,

$$z = \frac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$$

with α = 0.05, the null-hypothesis can be rejected if z is smaller than –1.96.

Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

The Wilcoxon signed ranks test is more sensible than the t-test. It assumes commensurability of differences, but only qualitatively: greater differences still count more, which is probably desired, but the absolute magnitudes are ignored.

From the statistical point of view, the test is safer since it does not assume normal distributions. Also, the outliers (exceptionally good/bad performances on a few data-sets/problems) have less effect on the Wilcoxon than on the t-test.

The Wilcoxon test assumes continuous differences, therefore they should not be rounded to one or two decimals, since this would decrease the power of the test due to a high number of ties.

Basic Non-Parametric Tests and Case Studies

- For Pairwise Comparisons
- For Multiple Comparisons involving a Control Method
- Evolutionary Algorithms: CEC’05 Special Session of Parameter Optimization

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Using Wilcoxon test for comparing multiple pairs of algorithms:

Wilcoxon’s test performs individual comparisons between two algorithms (pairwise comparisons). The *p-value in a pairwise comparison is independent from another* one. If we try to extract a conclusion involving more than one pairwise comparison in a Wilcoxon’s analysis, we will obtain an accumulated error coming from the combination of pairwise comparisons. In statistical terms, we are losing the control on the Family Wise Error Rate (FWER), defined as the probability of making one or more false discoveries among all the hypotheses when performing multiple pairwise tests.

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

When a p-value is considered in a multiple comparison, it reflects the probability error of a certain comparison, but it does not take into account the remaining comparisons belonging to the family.

If one is comparing k algorithms and in each comparison the level of significance is α, then in a single comparison the probability of not making a Type I error is $(1 - \alpha)$, then
the probability of not making a Type I error in the k-1 comparison is $(1- \alpha)\cdot(k-1)$. Then the probability of making one or more Type I error is $1 - (1- \alpha)\cdot(k-1)$.

For instance, if $\alpha = 0.05$ and $k = 10$, this is 0.37, which is rather high.

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Friedman's test: It is a non-parametric equivalent of the test of repeated-measures ANOVA. It computes the ranking of the observed results for algorithm (r_j for the algorithm j with k algorithms) for each function/algorithm, assigning to the best of them the ranking 1, and to the worst the ranking k.

Under the null hypothesis, formed from supposing that the results of the algorithms are equivalent and, therefore, their rankings are also similar, the Friedman statistic

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

is distributed according to con *k - 1* degrees of freedom, being , $R_j = \frac{1}{N}\sum_i r_i^j$ and *N* the number of functions/algorithms. (N > 10, k > 5)
(Table B.1. Critical Values of the Chi-Square Distribution, App. 12, Zar).

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Iman and Davenport's test: It is a metric derived from the Friedman's statistic given that this last metric produces a conservative undesirably effect. The statistic is:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

and it is distributed according to a F distribution with $k - 1$ and $(k - 1)(N - 1)$ degrees of freedom.

(Table B.4. Critical values of the F Distribution, App. 21, Zar).

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Example of the Friedman Test

The results obtained (performances) are arranged by a matrix of data with data sets in the rows and algorithms in the columns.

C4.5 with cf parameter is the version which optimizes AUC considering various levels of confidence for pruning a leaf.

dataset	C4.5	C4.5m	C4.5cf	C4.5cf,m
Adult	0.763	0.768	0.771	0.798
Breast	0.599	0.591	0.590	0.569
Wisconsin	0.954	0.971	0.968	0.967
Cmc	0.628	0.661	0.654	0.657
Ionosphere	0.882	0.888	0.886	0.898
Iris	0.936	0.931	0.916	0.931
Bupa	0.661	0.668	0.609	0.685
Lung	0.583	0.583	0.563	0.625
Lymphography	0.775	0.838	0.866	0.875
Mushroom	1.000	1.000	1.000	1.000
Tumor	0.940	0.962	0.965	0.962
Rheum	0.619	0.666	0.614	0.669
Voting	0.972	0.981	0.975	0.975
Wine	0.957	0.978	0.946	0.970

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Example of the Friedman Test

Rankings are assigned in increasing order from the best to the worst algorithm for each dataset/problem.

Ties in performance are computed by averaged rankings.

The most interesting datum for now is the *Average Rank* for each algorithm.

dataset	C4.5	C4.5m	C4.5cf	C4.5cf,m
Adult	4	3	2	1
Breast	1	2	3	4
Wisconsin	4	1	2	3
Cmc	4	1	3	2
Ionosphere	4	2	3	1
Iris	1	2.5	4	2.5
Bupa	3	2	4	1
Lung	2.5	2.5	4	1
Lymphography	4	3	2	1
Mushroom	2.5	2.5	2.5	2.5
Tumor	4	2.5	1	2.5
Rheum	3	2	4	1
Voting	4	1	2	3
Wine	3	1	4	2
Average Rank	**3.143**	**2.000**	**2.893**	**1.964**

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

	C4.5	C4.5m	C4.5cf	C4.5cf,m
Average Rank	**3.143**	**2.000**	**2.893**	**1.964**

Friedman's measure

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] =$$

$$= \frac{12 \cdot 14}{4 \cdot 5}\left[9.878 + 4.000 + 8.369 + 3.857 - \frac{4 \cdot 25}{4}\right] =$$

$$= 9.28$$

Observing the critical value, it can be concluded that it rejects the null hypothesis

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

	C4.5	C4.5m	C4.5cf	C4.5cf,m
Average Rank	3.143	2.000	2.893	1.964

Iman and Davenport’s measure

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} = \frac{13 \cdot 9.28}{13 \cdot 3 - 9.28} = 3.69$$

$F_F = 3.69$, $F(3,3x13) = 2.85$

Observing the critical value, it can be concluded that it rejects the null hypothesis

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

If the null hypothesis is rejected by Friedman or Iman-Davenport test, we can proceed with a post-hoc test:

The most frequent case is when we want to compare one algorithm (the proposal) with a set of algorithm. This type of comparison involves a CONTROL method, and it is usually denoted as a 1 x n comparison.

The simplest procedure in 1 x n comparisons is the Bonferroni-Dunn test. It adjusts the global level of significance by dividing it by (k – 1) in all cases, being k the number of algorithms.

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

However, a more general way to obtain the differences among algorithms is to obtain a statistic that follow a normal distribution. The test statistics for comparing the i-th algorithm with the j-th algorithm is computed by:

$$z = (R_i - R_j) \Bigg/ \sqrt{\frac{k(k+1)}{6N}}$$

The z value is used to find the corresponding probability from the table of normal distribution, which is then compared with an appropriate α.

In Bonferroni-Dunn, α is always divided by (k - 1) independently of the comparison, following a very conservative behavior. For this reason other procedures such as Holm’s or Hochberg’s are preferred.

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Holm's method: We dispose of a test that sequentially checks the hypothesis ordered according to their significance. We will denote the p values ordered: $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$.

Holm's method compares each p_i with $\alpha/(k-i)$ starting from the most significant p value. If p_1 Is below than $\alpha/(k-1)$, the corresponding hypothesis is rejected and it leaves us to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, we continue with the process. As soon as a certain hypothesis can not be rejected, all the remaining hypothesis are maintained as accepted.

The value of z is used for finding the corresponding probability from the table of the nomal distribution, which is compared with the corresponding value of α .
(Table B.2. Zar – Proportions of the Normal Curve (One-tailed), App 17)

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Hochberg's method: It is a step-up procedure that works in the opposite direction to Holm's method, comparing the largest *p* value with α, the next largest with α/2 and so forth until it encounters a hypothesis it can reject. All hypotheses with smaller p values are then rejected as well.

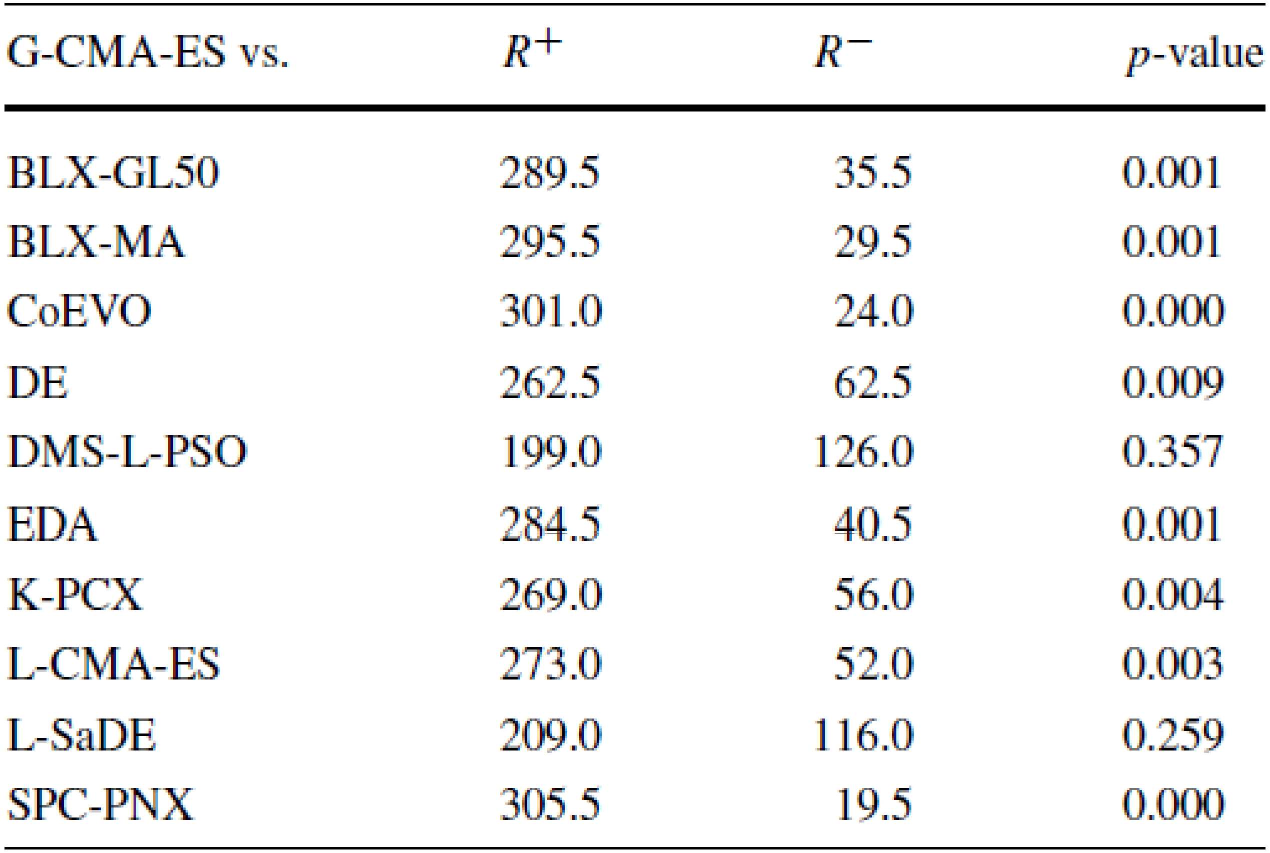
Hochberg's method is more powerful than Holm's although it may under some circumstances exceed the family-wise error.

Basic Non-Parametric Tests and Case Studies

- For Pairwise Comparisons
- For Multiple Comparisons involving a Control Method
- Evolutionary Algorithms: CEC’05 Special Session of Parameter Optimization

Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC’2005 Special Session of Parameter Optimization



G-CMA-ES vs.	R^+	R^-	p-value
BLX-GL50	289.5	35.5	0.001
BLX-MA	295.5	29.5	0.001
CoEVO	301.0	24.0	0.000
DE	262.5	62.5	0.009
DMS-L-PSO	199.0	126.0	0.357
EDA	284.5	40.5	0.001
K-PCX	269.0	56.0	0.004
L-CMA-ES	273.0	52.0	0.003
L-SaDE	209.0	116.0	0.259
SPC-PNX	305.5	19.5	0.000

G-CMAES versus the remaining algorithms.
P-value obtained through normal approximation

Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC'2005 Special Session of Parameter Optimization

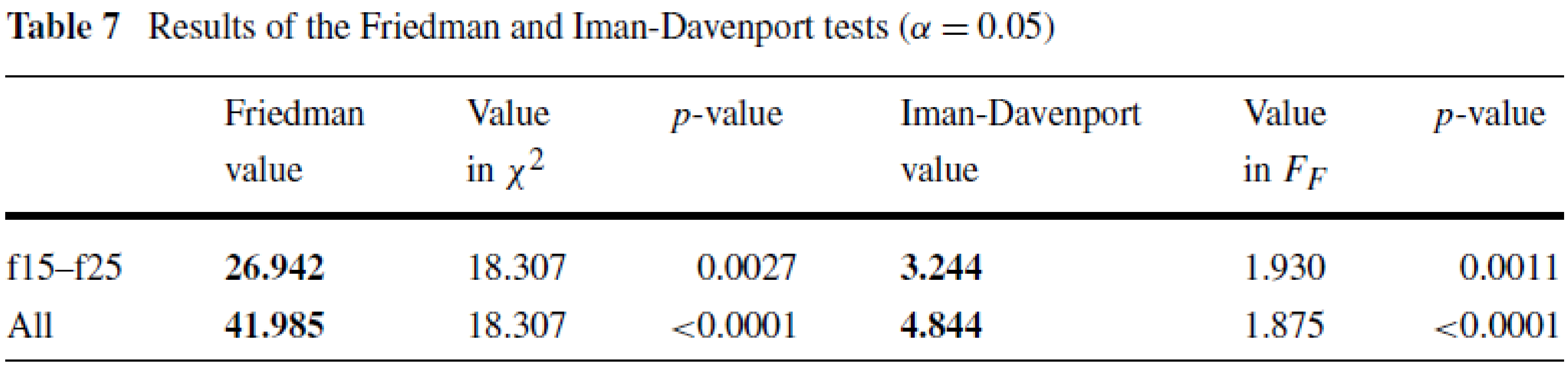
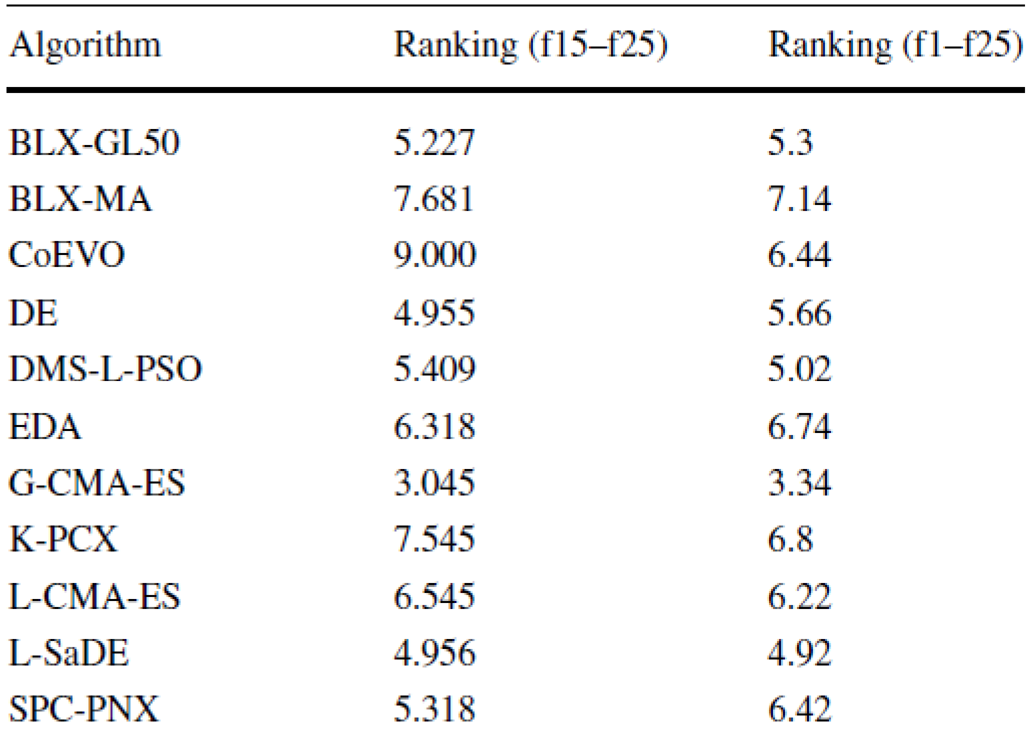


Table 7 Results of the Friedman and Iman-Davenport tests ($\alpha = 0.05$)

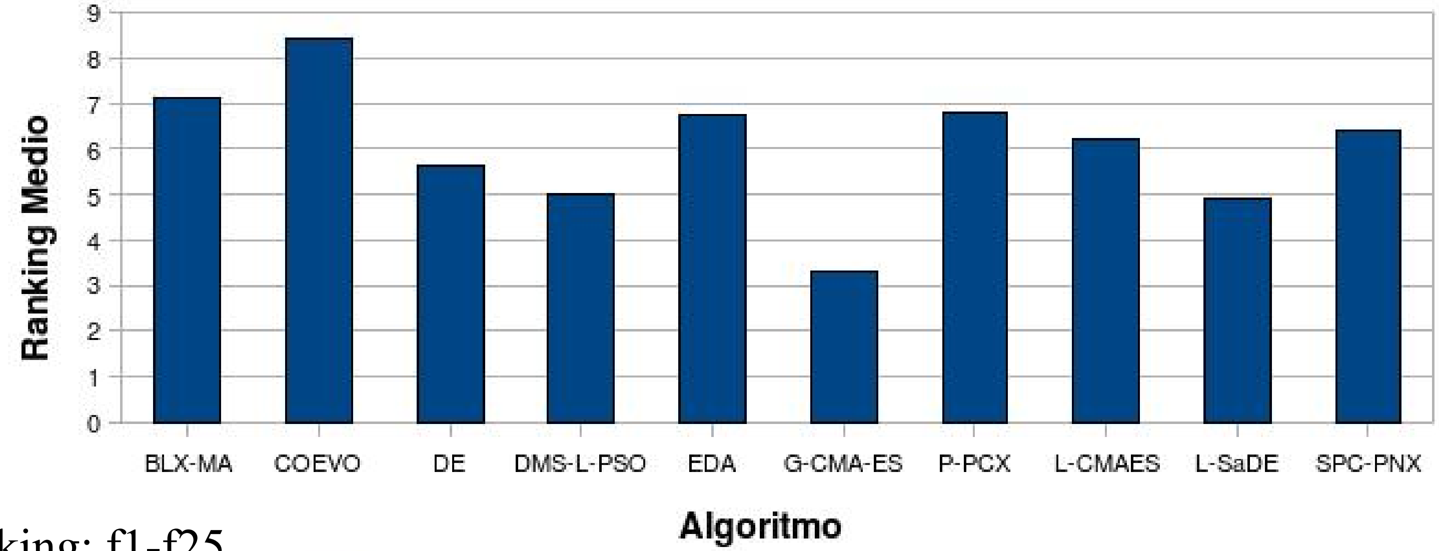
	Friedman value	Value in χ^2	p-value	Iman-Davenport value	Value in F_F	p-value
f15–f25	**26.942**	18.307	0.0027	**3.244**	1.930	0.0011
All	**41.985**	18.307	<0.0001	**4.844**	1.875	<0.0001



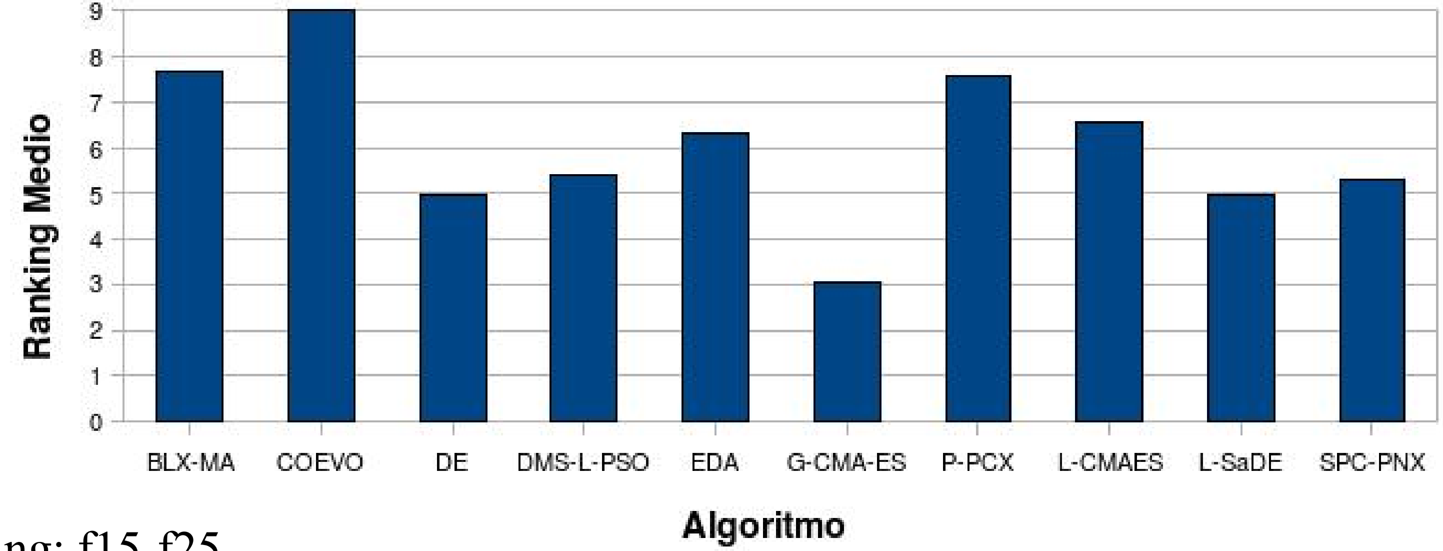
Algorithm	Ranking (f15–f25)	Ranking (f1–f25)
BLX-GL50	5.227	5.3
BLX-MA	7.681	7.14
CoEVO	9.000	6.44
DE	4.955	5.66
DMS-L-PSO	5.409	5.02
EDA	6.318	6.74
G-CMA-ES	3.045	3.34
K-PCX	7.545	6.8
L-CMA-ES	6.545	6.22
L-SaDE	4.956	4.92
SPC-PNX	5.318	6.42

Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC'2005 Special Session of Parameter Optimization



Ranking: f1-f25



Ranking: f15-f25

Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC’2005 Special Session of Parameter Optimization

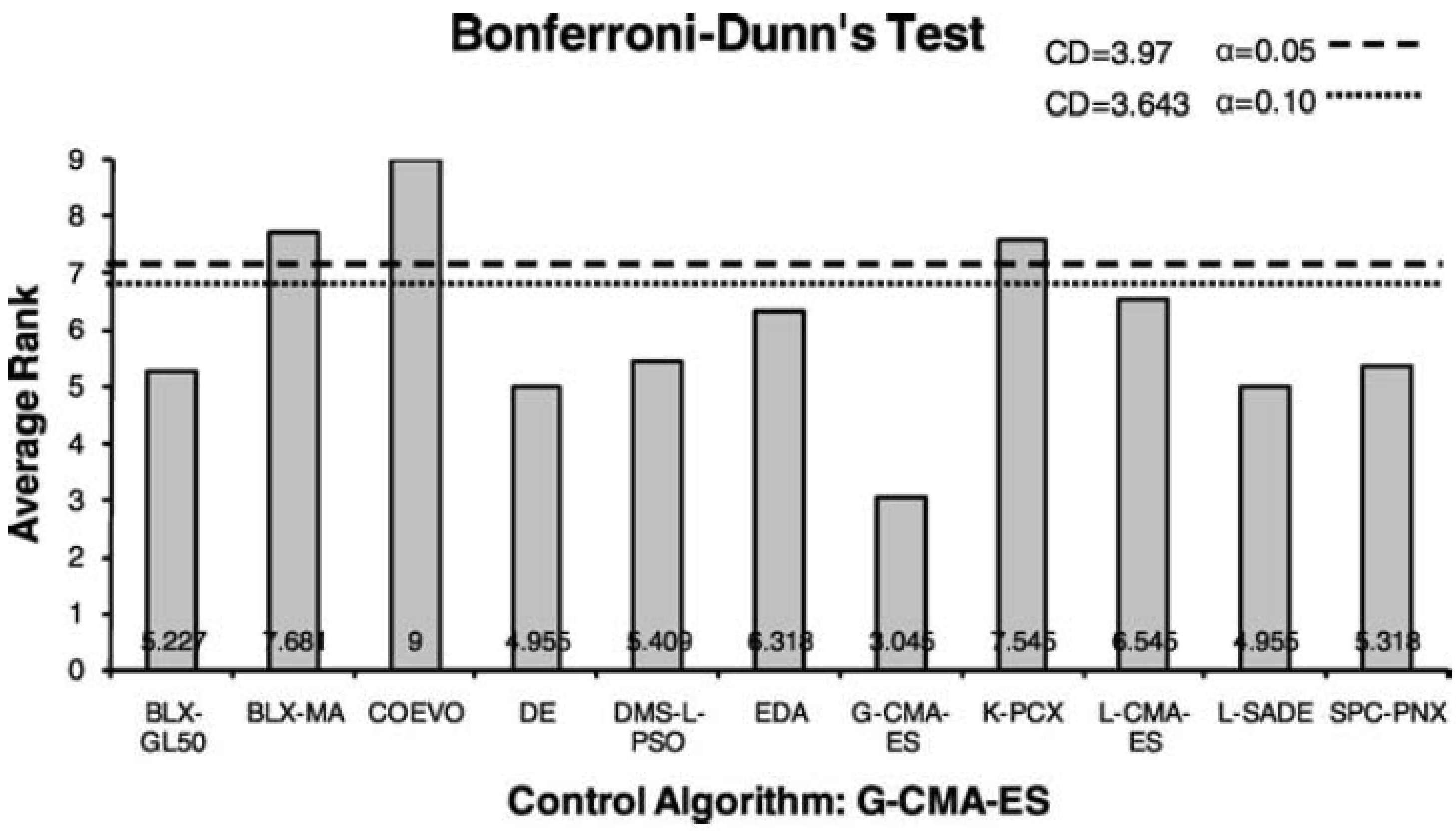
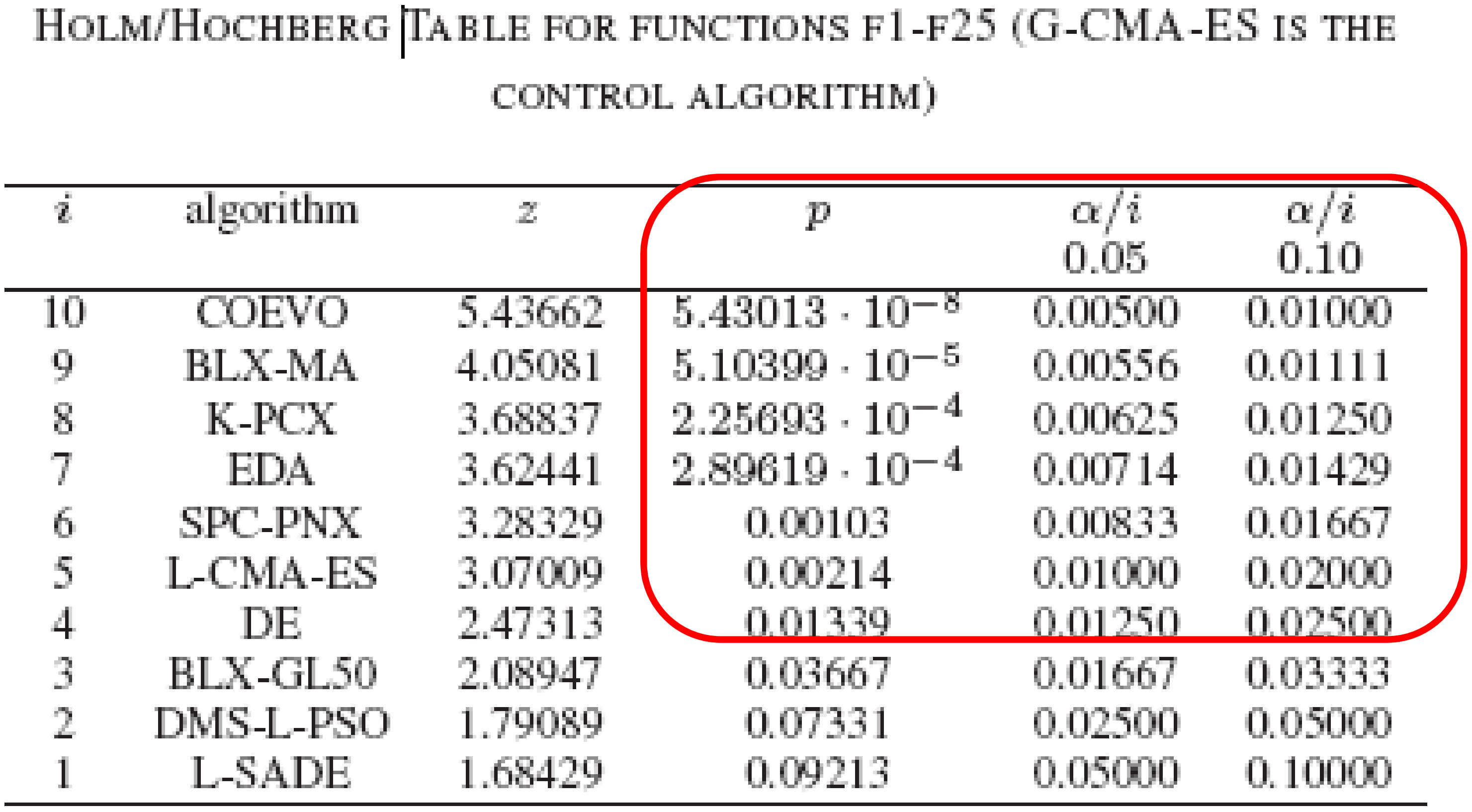


Fig. 6 Bonferroni-Dunn’s graphic corresponding to the results for f15–f25

Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC'2005 Special Session of Parameter Optimization

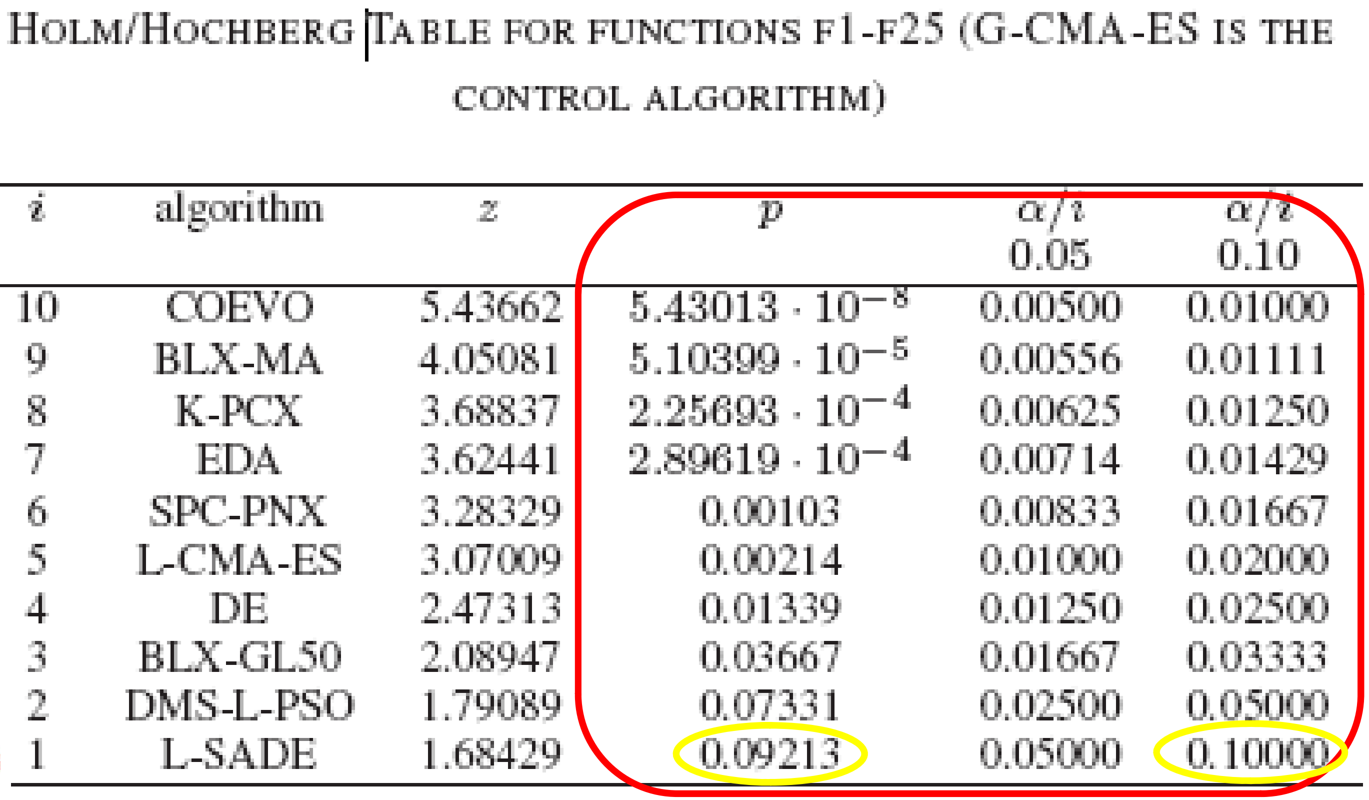


HOLM/HOCHBERG TABLE FOR FUNCTIONS F1-F25 (G-CMA-ES IS THE CONTROL ALGORITHM)

i	algorithm	z	p	α/i	α/i
				0.05	0.10
10	COEVO	5.43662	$5.43013 \cdot 10^{-8}$	0.00500	0.01000
9	BLX-MA	4.05081	$5.10399 \cdot 10^{-5}$	0.00556	0.01111
8	K-PCX	3.68837	$2.25693 \cdot 10^{-4}$	0.00625	0.01250
7	EDA	3.62441	$2.89619 \cdot 10^{-4}$	0.00714	0.01429
6	SPC-PNX	3.28329	0.00103	0.00833	0.01667
5	L-CMA-ES	3.07009	0.00214	0.01000	0.02000
4	DE	2.47313	0.01339	0.01250	0.02500
3	BLX-GL50	2.08947	0.03667	0.01667	0.03333
2	DMS-L-PSO	1.79089	0.07331	0.02500	0.05000
1	L-SADE	1.68429	0.09213	0.05000	0.10000

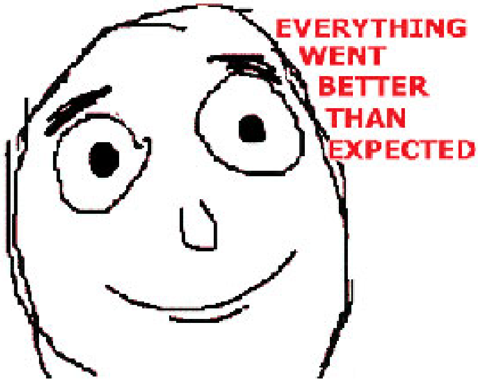
Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC'2005 Special Session of Parameter Optimization



HOLM/HOCHBERG TABLE FOR FUNCTIONS F1-F25 (G-CMA-ES IS THE CONTROL ALGORITHM)

i	algorithm	z	p	α/i 0.05	α/i 0.10
10	COEVO	5.43662	$5.43013 \cdot 10^{-8}$	0.00500	0.01000
9	BLX-MA	4.05081	$5.10399 \cdot 10^{-5}$	0.00556	0.01111
8	K-PCX	3.68837	$2.25693 \cdot 10^{-4}$	0.00625	0.01250
7	EDA	3.62441	$2.89619 \cdot 10^{-4}$	0.00714	0.01429
6	SPC-PNX	3.28329	0.00103	0.00833	0.01667
5	L-CMA-ES	3.07009	0.00214	0.01000	0.02000
4	DE	2.47313	0.01339	0.01250	0.02500
3	BLX-GL50	2.08947	0.03667	0.01667	0.03333
2	DMS-L-PSO	1.79089	0.07331	0.02500	0.05000
1	L-SADE	1.68429	0.09213	0.05000	0.10000



Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC'2005 Special Session of Parameter Optimization

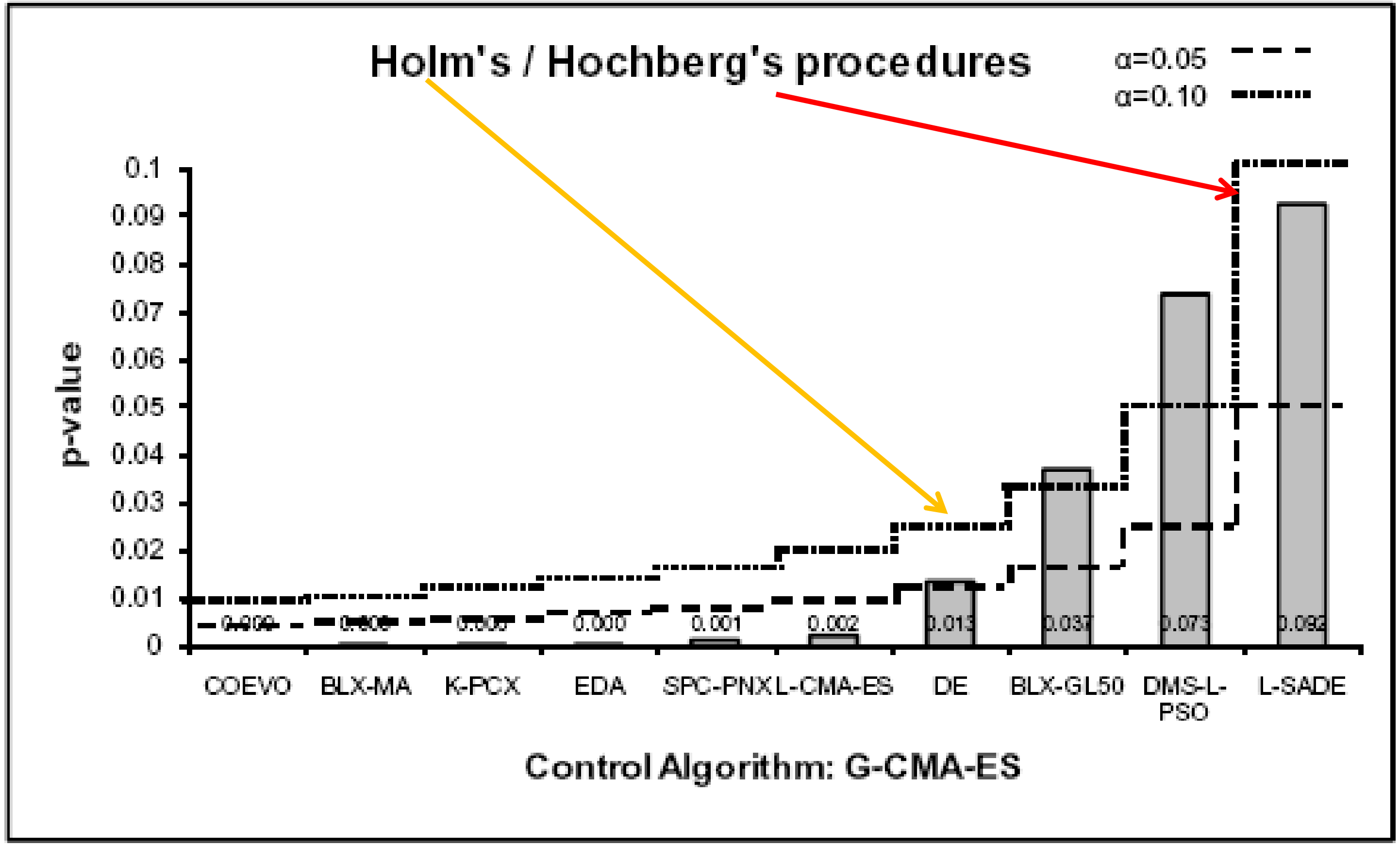


Fig. 11. Holm's/Hochberg's procedure for all functions (f1–f25).

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
  - Theoretical background
  - Checking the conditions in Parameter Optimization Experiments
- Basic non-parametric tests and case studies:
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Evolutionary Algorithms: CEC'05 Special Session on Parameter Optimization
- **Lessons Learned**
  - **Recommendations on the use of nonparametric tests**
  - **Frequent Questions**
- Books of Interest, References and Software

Lessons Learned

- Recommendations on the use of non-parametric tests
- Frequent questions

Lessons Learned

Recommendations on the Use of Non-Parametric Tests

Design of Experiments

They are not the objective of our talk, but they are two additional important questions:

- ❑ Benchmark functions/data sets … are very important.
- ❑ To compare with the state of the art is a necessity.

Lessons Learned

Recommendations on the Use of Non-Parametric Tests

What happens if I use a nonparametric test when the data is normal?

- It will work, but a parametric test would be more powerful, i.e., give a lower p value.

- If the data is not normal, then the nonparametric test is usually more powerful

- **Always look at the data first, then decide what test to use.**

Lessons Learned

Recommendations on the Use of Non-Parametric Tests

General

If we have a set of data sets/benchmark functions, we must apply a parametric test for each data set/benchmark function.

We only need to use a non-parametric test for comparing the algorithms on the whole set of benchmarks.

Lessons Learned

Recommendations on the Use of Non-Parametric Tests

Multiple comparison with a control (1)

❑ Holm’s procedure can always be considered better than Bonferroni-Dunn’s one, because it appropriately controls the FWER and it is more powerful than the Bonferroni-Dunn’s. We strongly recommend the use of Holm’s method in a rigorous comparison.

❑ Hochberg’s procedure is more powerful than Holm’s. The differences reported between it and Holm’s procedure are in practice rather small. We recommend the use of this test together with Holm’s method

Lessons Learned

Recommendations on the Use of Non-Parametric Tests

Multiple comparison with a control (2)

❑ The choice of any of the statistical procedures for conducting an experimental analysis should be justified by the researcher. The use of the most powerful procedures does not imply that the results obtained by his/her proposal will be better

Lessons Learned

- Recommendations on the use of non-parametric tests
- Frequent questions

Lessons Learned

Frequent Questions

❑ Can we analyze any performance measure?

❑ With non-parametric statistic, any unitary performance measure (associated to an only algorithm) with a pre-defined range of output can be analyzed. This range could be unlimited, allowing us to analyze time resources as example.

Lessons Learned

Frequent Questions

❑ Can we compare deterministic algorithms with stochastic ones?

❑ They allow us to compare both types of algorithms because they can be applied in multi-domain comparisons, where the sample of results is composed by a result that relates an algorithm and a domain of aplication (problem, function, data-set, …)

Lessons Learned

Frequent Questions

- How the average results should be obtained from each algorithm?

- This question does not concern to the use of non-parametric statistics, due to the fact that these tests require a result for each pair algorithm-domain. The obtaining of such result must be according to a standard procedure followed by all the algorithms in the comparison, such the case of validation techniques. Average results from various runs (at least 3) must be used for stochastic algorithms.

Lessons Learned

Frequent Questions

❑ What is the relationship between the number of algorithms and datasets/problems to do a correct statistical analysis?

❑ In multiple comparisons, the number of problems (data-sets) must be greater than the double of algorithms. With lesser data-sets, it is highly probable to not reject any null hyphotesis.

Lessons Learned

Frequent Questions

❑ Is there a maximum number of datasets/problems to be used?

❑ There not exists a theoretical threshold, although if the number of problems is very high in relation with the number of algorithms, the results trend to be inaccurate by the central limit theorem. For pairwise comparisons, such Wilcoxon's, a maximum of 30 problems is suggested. In multiple comparisons with a control, we should indicate as a rule of thumb that $n > 8 \cdot k$ could be excessive and results in no significant comparisons.

Lessons Learned

Frequent Questions

❑ The Wilcoxon test applied several times works better than a multiple comparison test such as Holm, Is it correct to be used in these cases?

❑ The Wilcoxon test can be applied according a multiple comparison scheme, but the results obtained cannot be considered into a family which control the FWER. Each time a new comparison is conducted, the level of significance established a priori can be overcome. For this reason, the multiple comparison tests exist.

Lessons Learned

Frequent Questions

❑ Can we use only the rankings obtained to justify the results?

❑ With the rankings values obtained by Friedman and derivatives we can establish a clear order in the algorithms and even to measure the differences among them. However, it cannot be concluded that one proposal is better than other until the hypothesis of comparison associated to them is rejected.

Lessons Learned

Frequent Questions

❑ Is it necessary to check the rejection of the null hypothesis of Friedman and derivatives before conducting a post-hoc analysis?

❑ It should be done, although by definition, it can be computed independently.

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
  - Theoretical background
  - Checking the conditions in Parameter Optimization Experiments
- Basic non-parametric tests and case studies:
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Evolutionary Algorithms: CEC’05 Special Session on Parameter Optimization
- Lessons Learned
  - Considerations on the use of nonparametric tests
  - Recommendations on the use of nonparametric tests
  - Frequent Questions
- **Books of Interest, References and Software**

Books of interest and References

JMLR
Expert Systems with Applications
An International Journal
Soft Computing
A Fusion of Foundations, Methodologies and Applications
HEURISTICS
INFORMATION SCIENCES

P1: S. García, F. Herrera, **An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons**. *Journal of Machine Learning Research 9 (2008) 2677-2694*

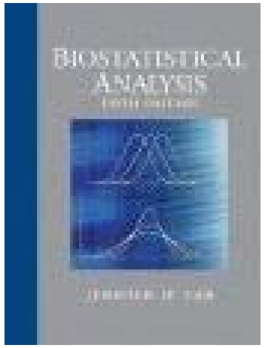
P2: J. Luengo, S. García, F. Herrera, **A Study on the Use of Statistical Tests for Experimentation with Neural Networks: Analysis of Parametric Test Conditions and Non-Parametric Tests**. *Expert Systems with Applications 36 (2009) 7798-7808 doi:10.1016/j.eswa.2008.11.041*.

P3: S. García, A. Fernández, J. Luengo, F. Herrera, **A Study of Statistical Techniques and Performance Measures for Genetics-Based Machine Learning: Accuracy and Interpretability**. *Soft Computing 13:10 (2009) 959-977, doi:10.1007/s00500-008-0392-y*.

P4: S. García, D. Molina, M. Lozano, 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, 15 (2009) 617-644. doi: 10.1007/s10732-008-9080-4*.

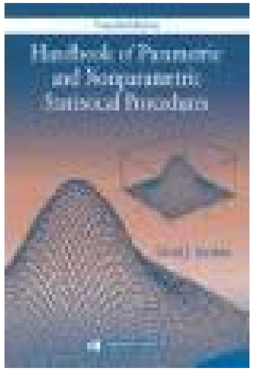
P5: S. García, A. Fernández, J. Luengo, F. Herrera, **Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental Analysis of Power**. *Information Sciences 180 (2010) 2044–2064. doi:10.1016/j.ins.2009.12.010*.

Books of interest and References



J.H. Zar, Biostatistical Analyhsis, Prentice Hall, 1999.

D. Sheskin. Handbook of parametric and nonparametric statistical procedures. Chapman & Hall/CRC, 2007.



Demsar, J., Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. Vol. 7. pp. 1–30. 2006.

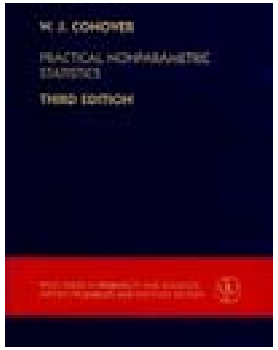
Books of interest and References

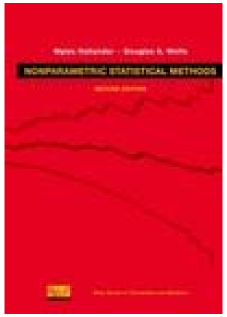
W.W. Daniel. Applied Nonparametric Statistics.

Houghton Mifflin Harcourt. (1990)

W.J. Conover. Practical Nonparametric Statistics.

Wiley. (1998)





M. Hollander and D.A. Wolfe. Nonparametric Statistical Methods.

Wiley-Interscience. (1999)

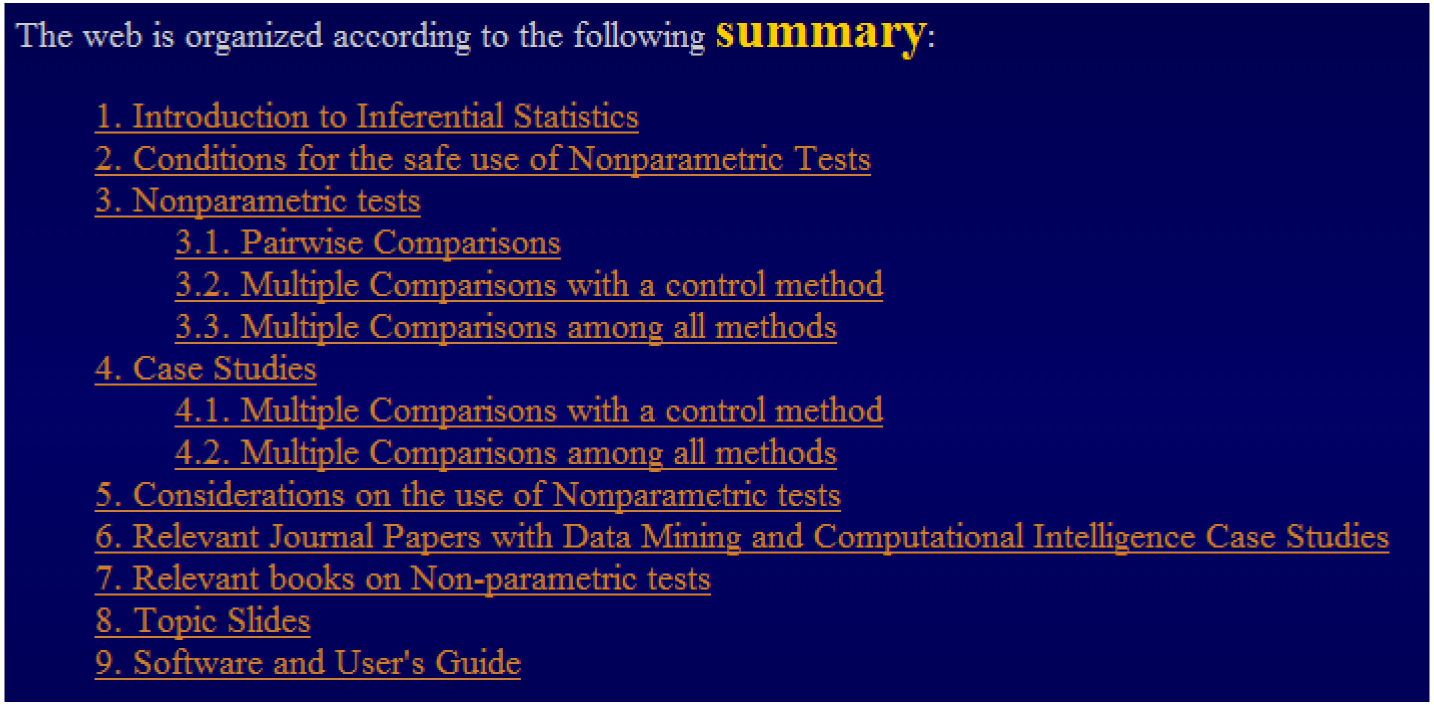
J.J. Higgins. Introduction to Modern Nonparametric

Statistics. Duxbury Press. (2003).

Books of interest and References

Website **http://sci2s.ugr.es/sicidm/**

SCI2S Thematic Public Websites: Statistical Inference in Computational Intelligence and Data Mining



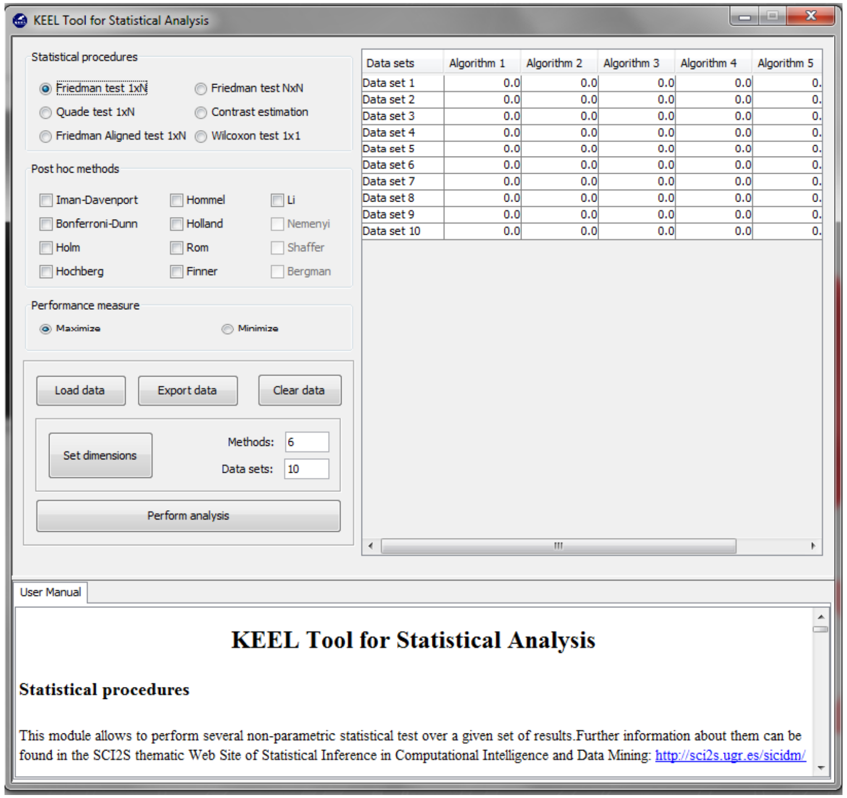
The web is organized according to the following **summary**:

1. Introduction to Inferential Statistics
2. Conditions for the safe use of Nonparametric Tests
3. Nonparametric tests
    - 3.1. Pairwise Comparisons
    - 3.2. Multiple Comparisons with a control method
    - 3.3. Multiple Comparisons among all methods
4. Case Studies
    - 4.1. Multiple Comparisons with a control method
    - 4.2. Multiple Comparisons among all methods
5. Considerations on the use of Nonparametric tests
6. Relevant Journal Papers with Data Mining and Computational Intelligence Case Studies
7. Relevant books on Non-parametric tests
8. Topic Slides
9. Software and User's Guide

Software

Software for conducting nonparametric statistical analysis

http://www.keel.es/



KEEL Tool for Statistical Analysis

Statistical procedures

● Friedman test 1xN ○ Friedman test NxN
○ Quade test 1xN ○ Contrast estimation
○ Friedman Aligned test 1xN ○ Wilcoxon test 1x1

Post hoc methods

□ Iman-Davenport □ Hommel □ Li
□ Bonferroni-Dunn □ Holland □ Nemenyi
□ Holm □ Rom □ Shaffer
□ Hochberg □ Finner □ Bergman

Performance measure

● Maximize ○ Minimize

Methods: 6
Data sets: 10

Data sets	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4	Algorithm 5
Data set 1	0.0	0.0	0.0	0.0	0.
Data set 2	0.0	0.0	0.0	0.0	0.
Data set 3	0.0	0.0	0.0	0.0	0.
Data set 4	0.0	0.0	0.0	0.0	0.
Data set 5	0.0	0.0	0.0	0.0	0.
Data set 6	0.0	0.0	0.0	0.0	0.
Data set 7	0.0	0.0	0.0	0.0	0.
Data set 8	0.0	0.0	0.0	0.0	0.
Data set 9	0.0	0.0	0.0	0.0	0.
Data set 10	0.0	0.0	0.0	0.0	0.

KEEL Tool for Statistical Analysis

Statistical procedures

This module allows to perform several non-parametric statistical test over a given set of results.Further information about them can be found in the SCI2S thematic Web Site of Statistical Inference in Computational Intelligence and Data Mining: http://sci2s.ugr.es/sicidm/

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

How must I conduct statistical comparisons in my Experimental Study? On the use of Nonparametric Tests and Case Studies.

Thanks!!! ?