INIT/AERFAI Summer School on MACHINE LEARNING

Benicàssim, June 24-28, 2013

Statistical Analysis of Experiments

Salvador García, Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI²S)
Dept. of Computer Science and A.I.
University of Granada, Spain

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

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

In this talk

We focus on the use of statistical tests for analyzing the results obtained in a design of experiments within the fields of 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 on Data Mining or Computational Intelligence (among other fields).

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

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

You cannot show the superiority of your method without statistical tests.

Experimental papers without statistics tests may be rejected

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Motivation

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

Example for classification

Large Variations in Accuracies of Different Classifiers

	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 in the fields of Data Mining and Computational Intelligence.

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 nonparametrics tests depending on the situation will be given.

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
- Basic non-parametric tests and case studies
- Advanced non-parametric tests and case studies
- Lessons Learned
- Books of Interest and References
- Software

OUTLINE (I)

- **Introduction to Inferential Statistics**
- **Conditions for the safe use of parametric tests**
  - Theoretical background
  - Checking the conditions in Data Mining Experiments
  - Checking the conditions in Parameter Optimization Experiments
- **Basic non-parametric tests and case studies**
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Data Mining: Neural Networks and Genetic Learning
  - Evolutionary Algorithms: CEC’05 Special Session on Parameter Optimization

OUTLINE (II)

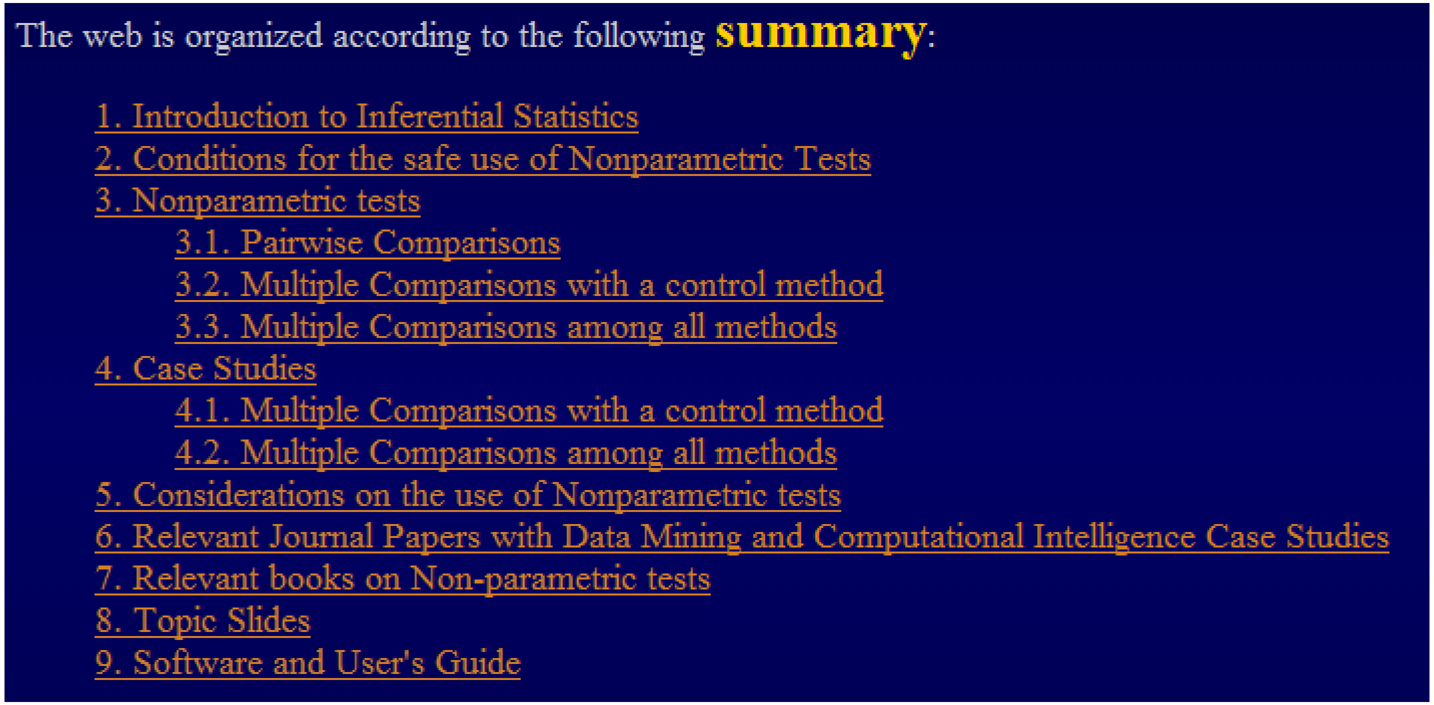
- **Advanced non-parametric tests and case studies**
  - For Multiple Comparisons involving control method
  - Post-hoc Procedures
  - Adjusted p-values
  - Detecting all pairwise differences in a multiple comparison
- **Lessons Learned**
  - Considerations on the use of nonparametric tests
  - Recommendations on the use of nonparametric tests
  - Frequent Questions
- **Books of Interest and References**
- **Software**

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

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

OUTLINE (I)

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
  - Theoretical background
  - Checking the conditions in Data Mining Experiments
  - Checking the conditions in Parameter Optimization Experiments
- Basic non-parametric tests and case studies
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Data Mining: Neural Networks and Genetic Learning
  - Evolutionary Algorithms: CEC'05 Special Session on Parameter Optimization

Introduction to Inferential Statistics

Inferential Statistics

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

For example: Comparing two or various sets of experiments/results 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.

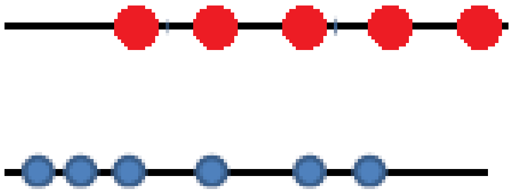
Hypothesis testing is procedure in which sample data are employed to evaluate a hypothesis.

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



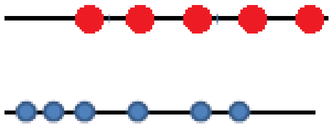
Can I consider a hypothesis for these data?

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 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.

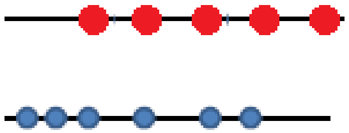
Introduction to Inferential Statistics

Examples of Null-Hypothesis

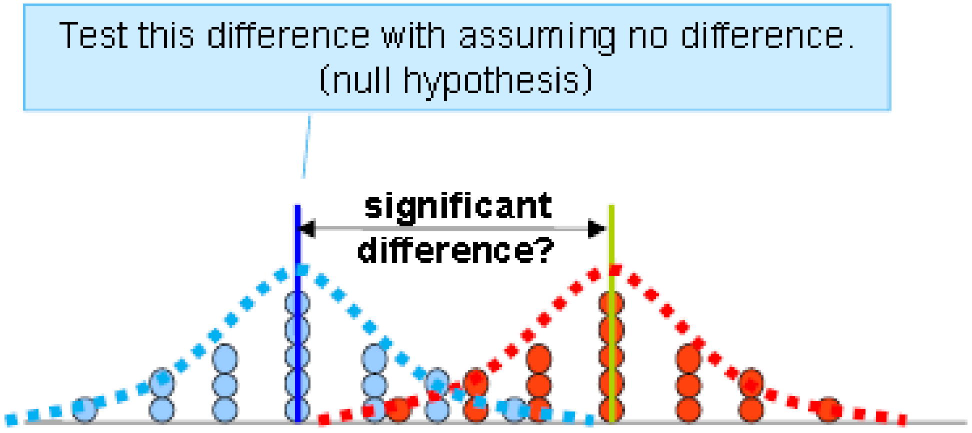
H_o: The 2 samples come from populations with the same distributions.

or,

median of population 1 = median of population 2



(generalization with n samples)



Test this difference with assuming no difference. (null hypothesis)

significant difference?

Introduction to Inferential Statistics

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).

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.**

Introduction to Inferential Statistics

Significance level α

- **Important to 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**

p-value

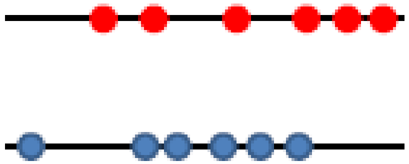
- Instead of stipulating a priori level of significance α (alpha), one could calculate the smallest level of significance that results in the rejection of the null hypothesis.

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

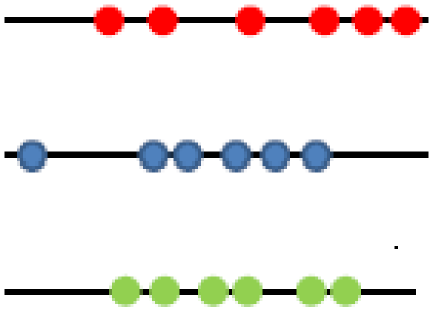
- **It does it without commiting to a particular level of significance.**

Introduction to Inferential Statistics

- **Compare two variables**



- **If more than two variables**



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, …)

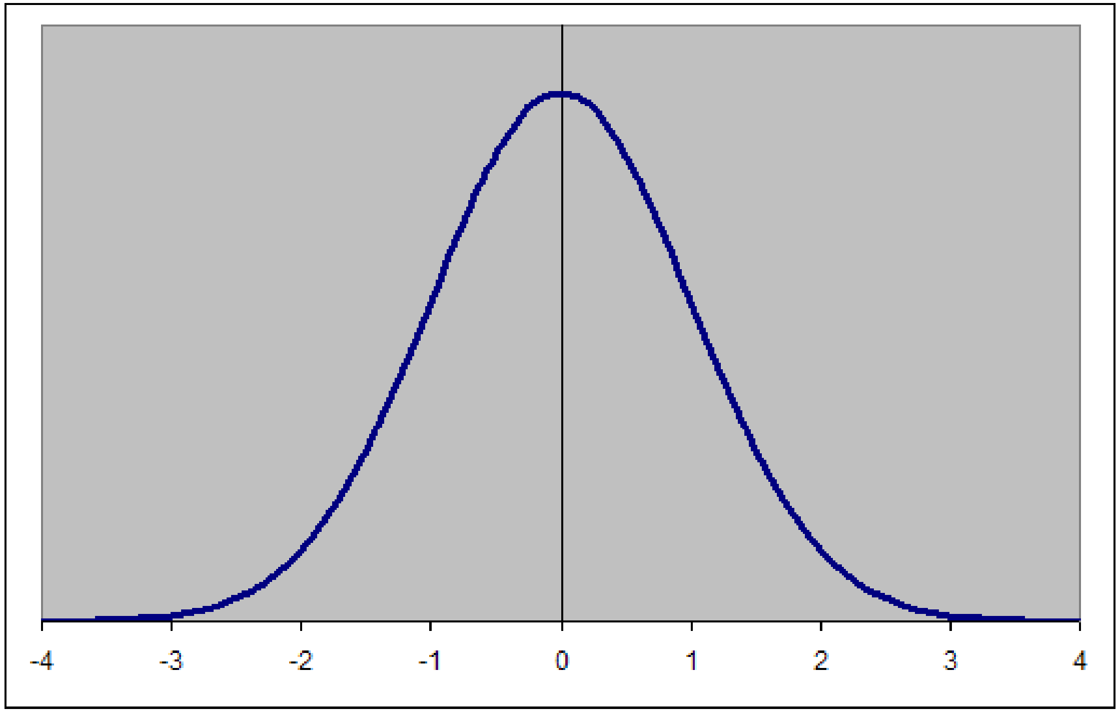
- The observations must be independent
- **Normality:** The observations must be drawn from normally distributed populations
- **Homoscedasticity:** These populations must have the same variances

Introduction to Inferential Statistics

Normality Tip

If a histogram representing your data looks like this,

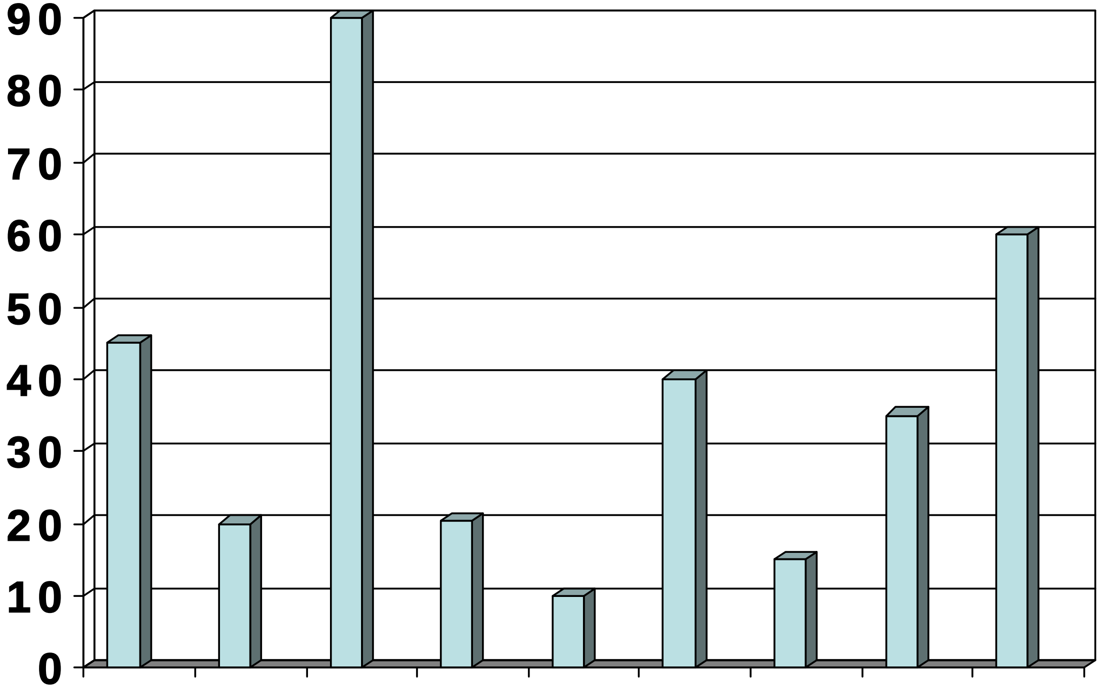
you can conduct a parametric test!



Introduction to Inferential Statistics

Otherwise, don’t conduct a parametric test!

The conclusions could be erroneous



Histogram

Introduction to Inferential Statistics

Nonparametric Assumptions

(t-test, ANOVA, …)

- 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
- Basic non-parametric tests and case studies
- Advanced non-parametric tests and case studies
- Lessons Learned
- Books of Interest and References
- Software

OUTLINE (I)

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
  - Theoretical background
  - Checking the conditions in Data Mining Experiments
  - Checking the conditions in Parameter Optimization Experiments
- Basic non-parametric tests and case studies
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Data Mining: Neural Networks and Genetic Learning
  - Evolutionary Algorithms: CEC'05 Special Session on Parameter Optimization

Conditions for the safe use of parametric tests

- Theoretical background
- Checking the conditions in Data Mining Experiments
- 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.**

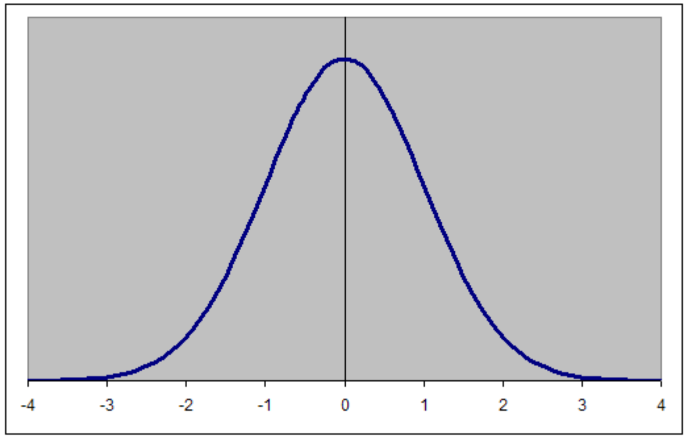
Conditions for the Safe Use of Parametric Tests

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.

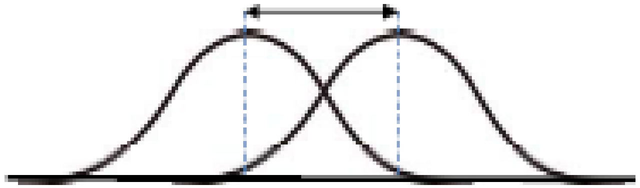
Conditions for the Safe Use of Parametric Tests

Theoretical Background

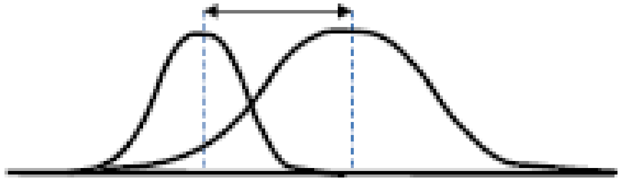
Heteroscedasticity: This property indicates the existence of a violation of the hypothesis of equality of variances.

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

Two sample assuming equal variances



Two sample assuming the unequal variances

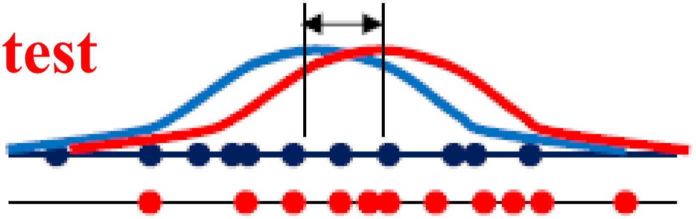


Conditions for the Safe Use of Parametric Tests

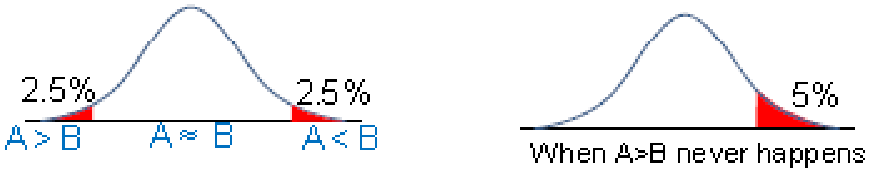
Theoretical Background

On the parametric tests

T-test

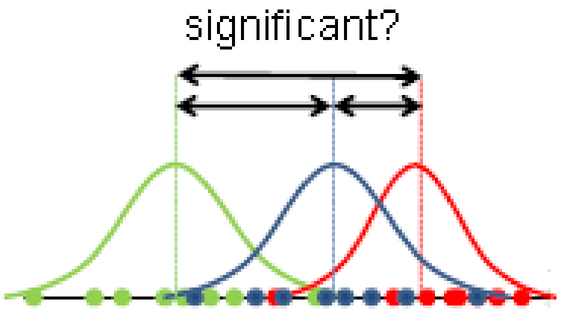


When *p*-value is less than 0.01 or 0.05, we assume that there is significant difference with the level of significance of ($p < 0.01$) or ($p < 0.05$).



When A>B never happens, you may use a one-tail test.

ANOVA Analysis



If normality and equal variances are not guaranteed, use non-parametric tests.

Conditions for the Safe Use of Parametric Tests

Theoretical Background

If X has a standard normal distribution, i.e. $X \sim N(0,1)$,

$$P(X > 1.96) = 0.025,$$

$$P(X < 1.96) = 0.975,$$

and as the normal distribution is symmetric,

$$P(-1.96 < X < 1.96) = 0.95.$$

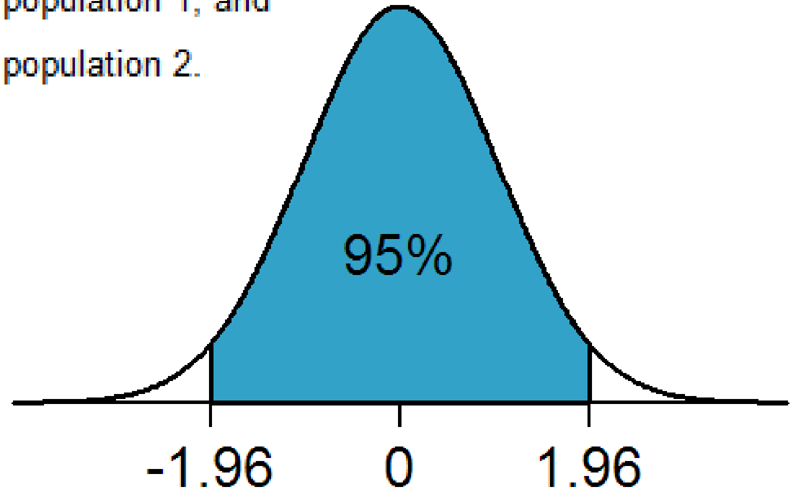
H_0: $\mu_1 = \mu_2$

where:

H_0 = the null hypothesis

μ_1 = the mean of population 1, and

μ_2 = the mean of population 2.



1.96: "normal score" or "Z score"

1.96 is the approximate value of the 97.5 percentile point of the normal distribution used in probability and statistics. 95% of the area under a normal curve lies within roughly 1.96 standard deviations of the mean, and due to the central limit theorem, this number is therefore used in the construction of approximate 95% confidence intervals.

Conditions for the safe use of parametric tests

- Theoretical background
- Checking the conditions in Data Mining Experiments
- Checking the conditions in Parameter Optimization Experiments

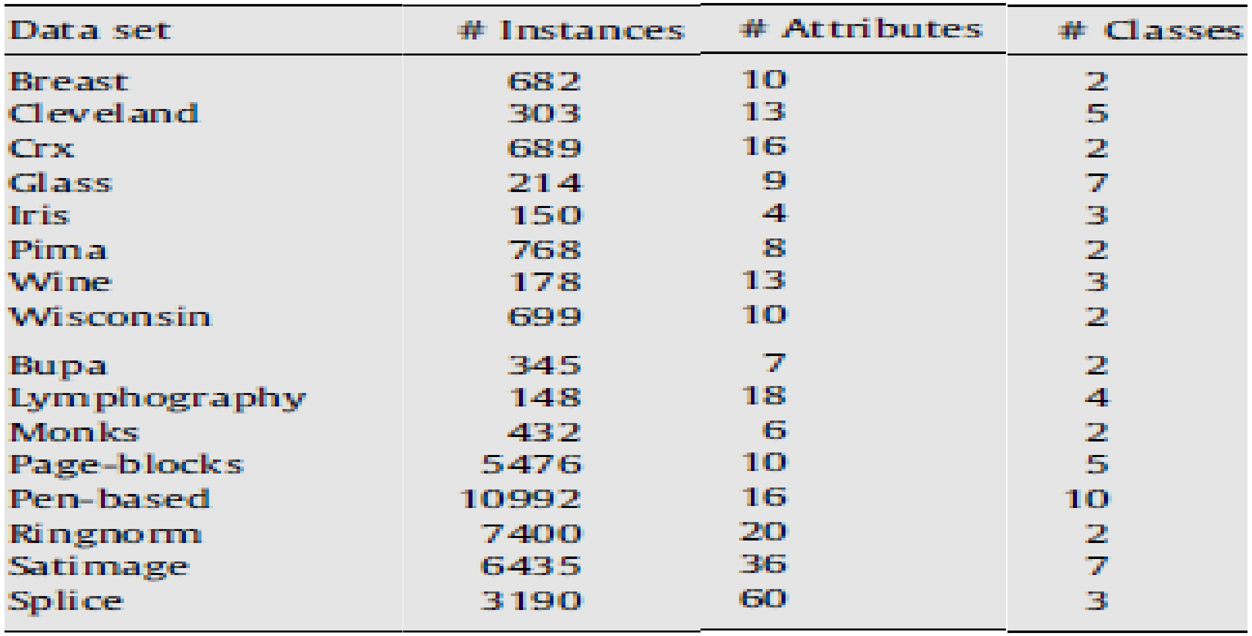
Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

FIRST CASE STUDY: Neural networks models:

MLP, RBFN (3 versions), LQV

Hold-Out Validation (HOV), 10FCV and 5x2CV (5 runs each one)



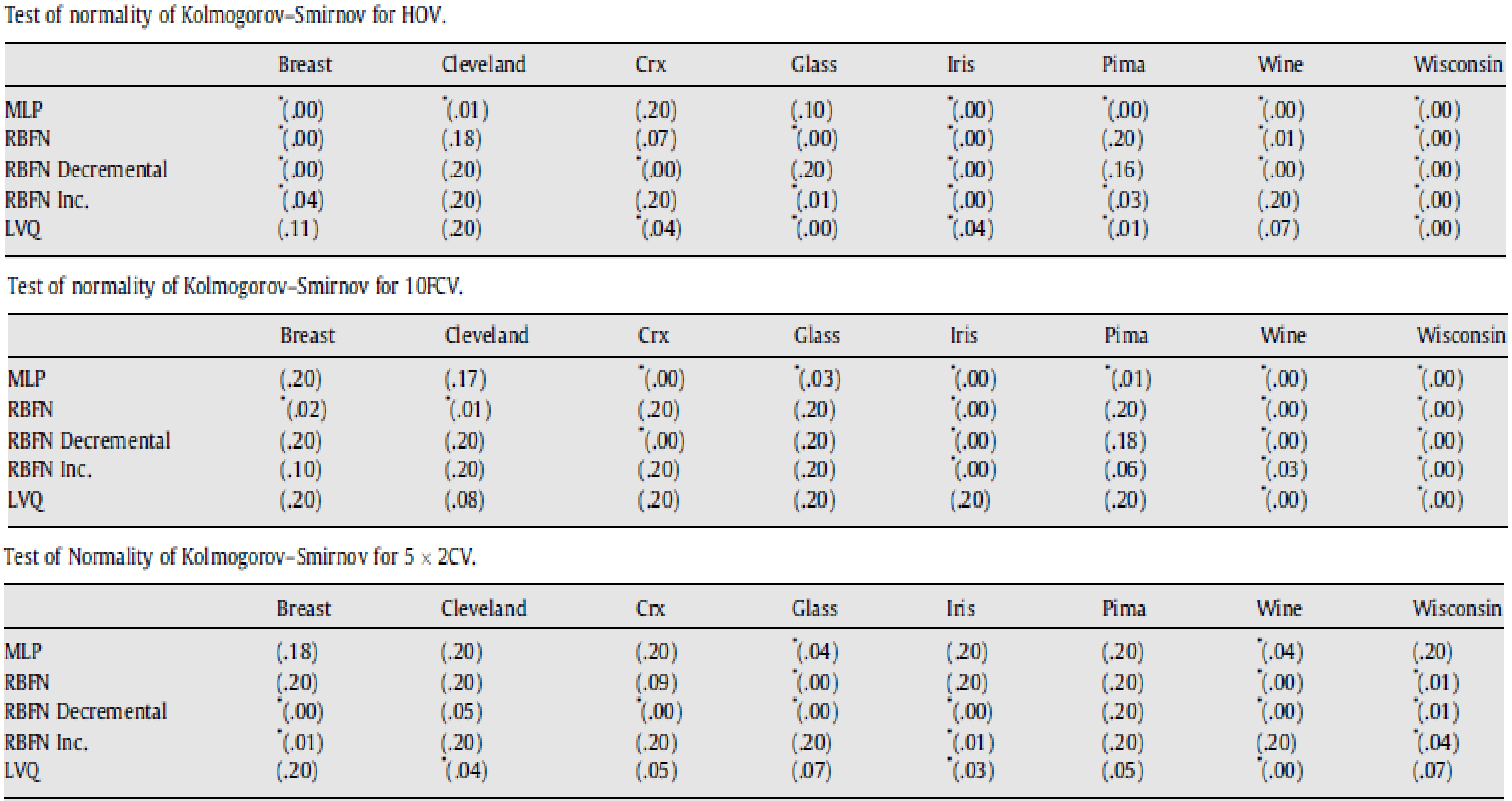
Data set	# Instances	# Attributes	# Classes
Breast	682	10	2
Cleveland	303	13	5
Crx	689	16	2
Glass	214	9	7
Iris	150	4	3
Pima	768	8	2
Wine	178	13	3
Wisconsin	699	10	2
Bupa	345	7	2
Lymphography	148	18	4
Monks	432	6	2
Page-blocks	5476	10	5
Pen-based	10992	16	10
Ringnorm	7400	20	2
Satimage	6435	36	7
Splice	3190	60	3

Refernce: 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*

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

TABLE 1. Kolmogorov-Smirnov test



Test of normality of Kolmogorov-Smirnov for HOV.

	Breast	Cleveland	Crx	Glass	Iris	Pima	Wine	Wisconsin
MLP	*(.00)	*(.01)	(.20)	(.10)	*(.00)	*(.00)	*(.00)	*(.00)
RBFN	*(.00)	(.18)	(.07)	*(.00)	*(.00)	(.20)	*(.01)	*(.00)
RBFN Decremental	*(.00)	(.20)	*(.00)	(.20)	*(.00)	(.16)	*(.00)	*(.00)
RBFN Inc.	*(.04)	(.20)	(.20)	*(.01)	*(.00)	*(.03)	(.20)	*(.00)
LVQ	(.11)	(.20)	*(.04)	*(.00)	*(.04)	*(.01)	(.07)	*(.00)

Test of normality of Kolmogorov-Smirnov for 10FCV.

	Breast	Cleveland	Crx	Glass	Iris	Pima	Wine	Wisconsin
MLP	(.20)	(.17)	*(.00)	*(.03)	*(.00)	*(.01)	*(.00)	*(.00)
RBFN	*(.02)	*(.01)	(.20)	(.20)	*(.00)	(.20)	*(.00)	*(.00)
RBFN Decremental	(.20)	(.20)	*(.00)	(.20)	*(.00)	(.18)	*(.00)	*(.00)
RBFN Inc.	(.10)	(.20)	(.20)	(.20)	*(.00)	(.06)	*(.03)	*(.00)
LVQ	(.20)	(.08)	(.20)	(.20)	(.20)	(.20)	*(.00)	*(.00)

Test of Normality of Kolmogorov-Smirnov for 5 × 2CV.

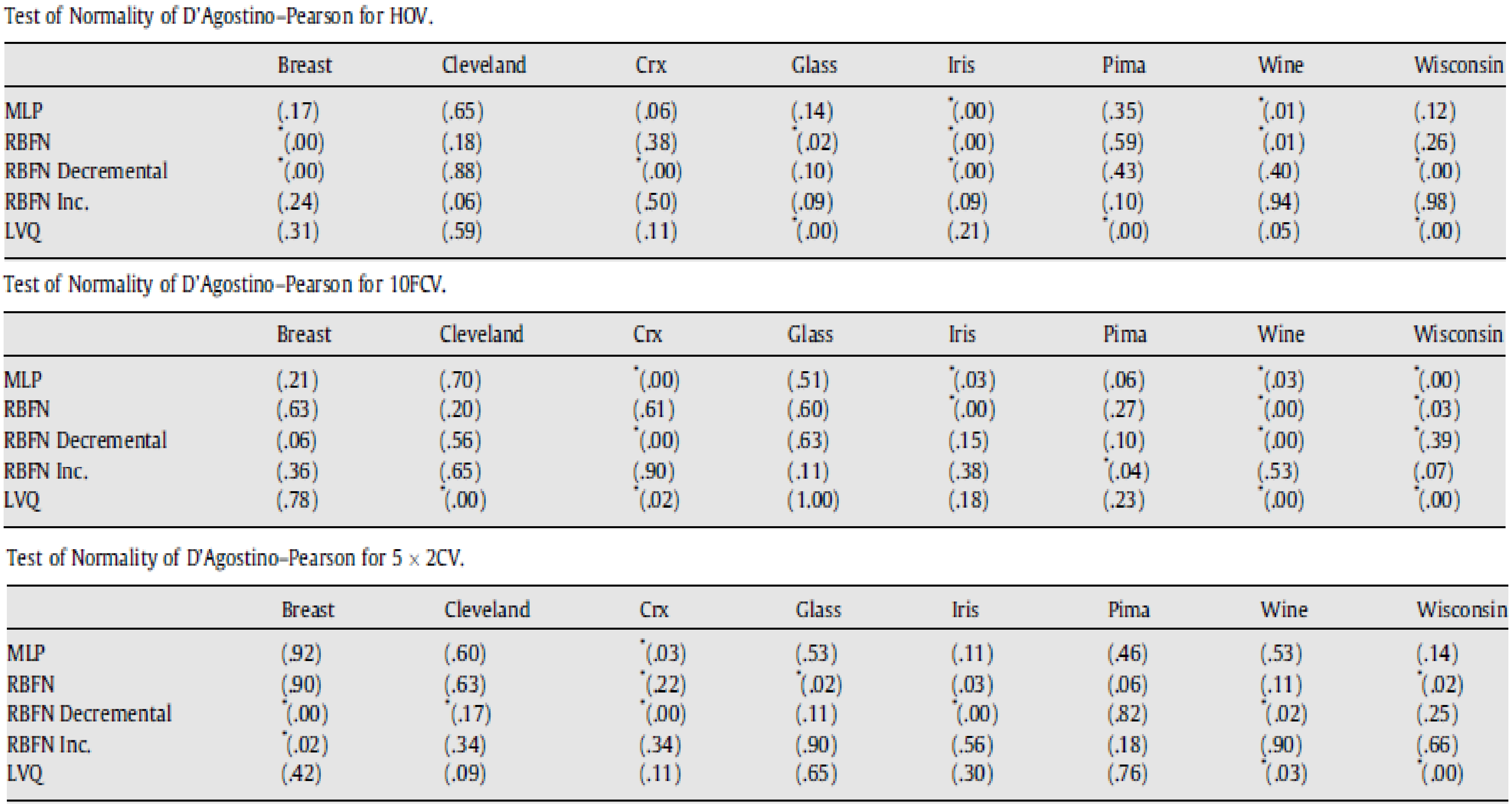
	Breast	Cleveland	Crx	Glass	Iris	Pima	Wine	Wisconsin
MLP	(.18)	(.20)	(.20)	*(.04)	(.20)	(.20)	*(.04)	(.20)
RBFN	(.20)	(.20)	(.09)	*(.00)	(.20)	(.20)	*(.00)	*(.01)
RBFN Decremental	*(.00)	(.05)	*(.00)	*(.00)	*(.00)	(.20)	*(.00)	*(.01)
RBFN Inc.	*(.01)	(.20)	(.20)	(.20)	*(.01)	(.20)	(.20)	*(.04)
LVQ	(.20)	*(.04)	(.05)	(.07)	*(.03)	(.05)	*(.00)	(.07)

a **p-value** smaller than 0.05 indicates that you can reject the **null-hypothesis**

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

TABLE 2. Comparison among validations



Test of Normality of D'Agostino-Pearson for HOV.

	Breast	Cleveland	Crx	Glass	Iris	Pima	Wine	Wisconsin
MLP	(.17)	(.65)	(.06)	(.14)	*(.00)	(.35)	*(.01)	(.12)
RBFN	*(.00)	(.18)	(.38)	*(.02)	*(.00)	(.59)	*(.01)	(.26)
RBFN Decremental	*(.00)	(.88)	*(.00)	(.10)	*(.00)	(.43)	(.40)	*(.00)
RBFN Inc.	(.24)	(.06)	(.50)	(.09)	(.09)	(.10)	(.94)	(.98)
LVQ	(.31)	(.59)	(.11)	*(.00)	(.21)	*(.00)	(.05)	*(.00)

Test of Normality of D'Agostino-Pearson for 10FCV.

	Breast	Cleveland	Crx	Glass	Iris	Pima	Wine	Wisconsin
MLP	(.21)	(.70)	*(.00)	(.51)	*(.03)	(.06)	*(.03)	*(.00)
RBFN	(.63)	(.20)	(.61)	(.60)	*(.00)	(.27)	*(.00)	*(.03)
RBFN Decremental	(.06)	(.56)	*(.00)	(.63)	(.15)	(.10)	*(.00)	*(.39)
RBFN Inc.	(.36)	(.65)	(.90)	(.11)	(.38)	*(.04)	(.53)	(.07)
LVQ	(.78)	*(.00)	*(.02)	(1.00)	(.18)	(.23)	*(.00)	*(.00)

Test of Normality of D'Agostino-Pearson for 5 × 2CV.

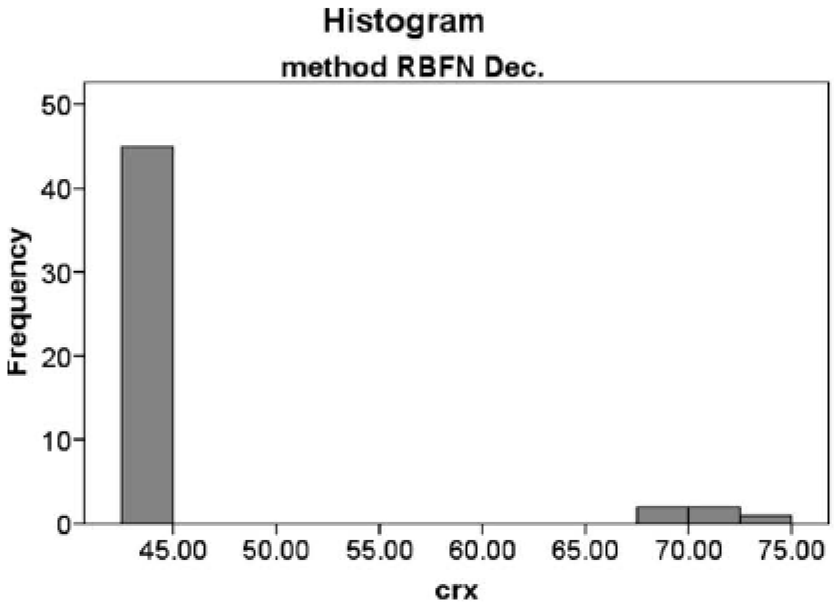
	Breast	Cleveland	Crx	Glass	Iris	Pima	Wine	Wisconsin
MLP	(.92)	(.60)	*(.03)	(.53)	(.11)	(.46)	(.53)	(.14)
RBFN	(.90)	(.63)	*(.22)	*(.02)	(.03)	(.06)	(.11)	*(.02)
RBFN Decremental	*(.00)	*(.17)	*(.00)	(.11)	*(.00)	(.82)	*(.02)	(.25)
RBFN Inc.	*(.02)	(.34)	(.34)	(.90)	(.56)	(.18)	(.90)	(.66)
LVQ	(.42)	(.09)	(.11)	(.65)	(.30)	(.76)	*(.03)	*(.00)

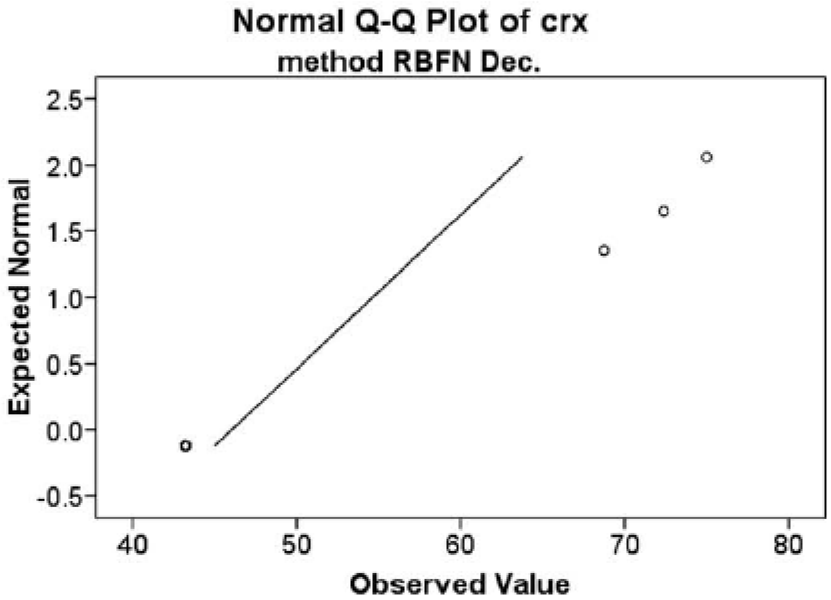
a **p-value** smaller than 0.05 indicates that you can reject the **null-hypothesis**

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

Histograms and Q-Q Grapics



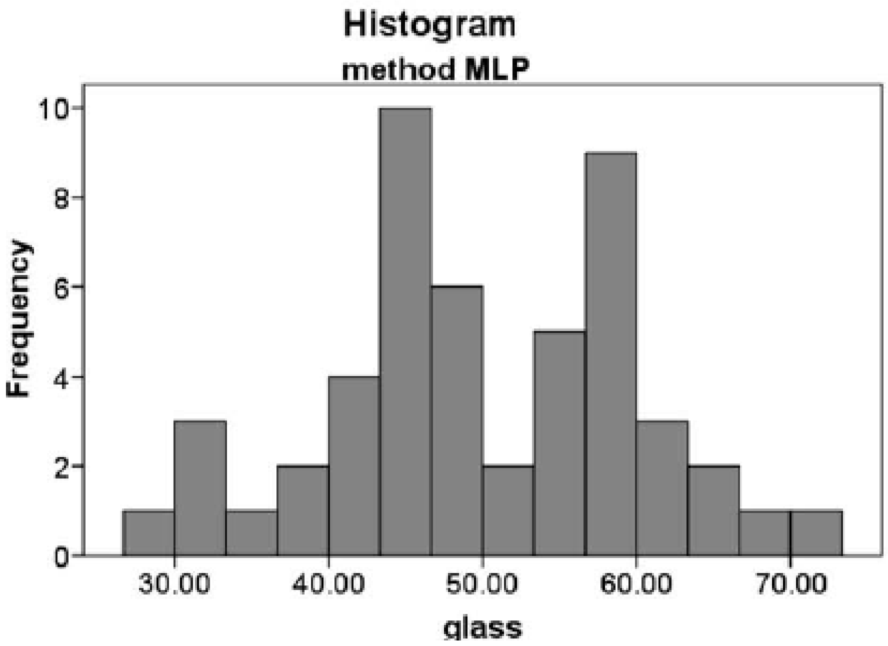


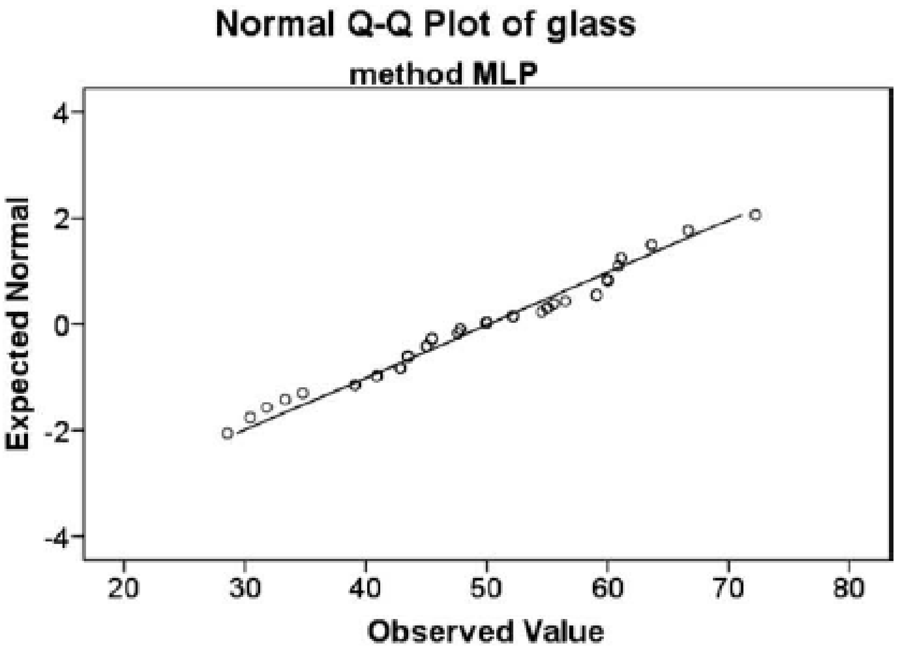
* A Q-Q graphic represents a confrontation between the quartiles from data observed and those from the normal distributions. Absolute lack of normality.

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

Histograms and Q-Q Grapics



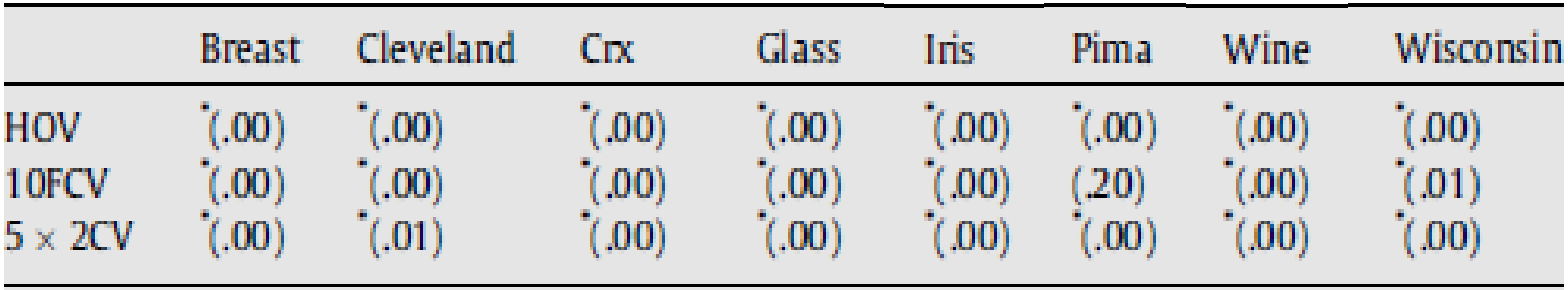


Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

TABLE 3. Test of HETEROSCEDASTICITY OF LEVENE

(BASED ON MEANS)



	Breast	Cleveland	Crx	Glass	Iris	Pima	Wine	Wisconsin
HOV	*(.00)	*(.00)	*(.00)	*(.00)	*(.00)	*(.00)	*(.00)	*(.00)
10FCV	*(.00)	*(.00)	*(.00)	*(.00)	*(.00)	(.20)	*(.00)	*(.01)
5 × 2CV	*(.00)	*(.01)	*(.00)	*(.00)	*(.00)	*(.00)	*(.00)	*(.00)

Table 3 shows the results by applying Levene's tests, where the symbol "*" indicates that the variances of the distributions of the different algorithms for a certain function are not homogeneities (we reject the null hypothesis).

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

SECOND CASE STUDY: Genetics-Based Machine Learning

- We have chosen four Genetic Interval Rule Based Algorithms:
  - *Pittsburgh Genetic Interval Rule Learning Algorithm.*
  - *XCS Algorithm.*
  - *GASSIST Algorithm.*
  - *HIDER Algorithm.*
- GBML will be analyzed by two performance measures: *Accuracy* and *Cohen's kappa*.
- **How we state which is the best?**

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

Experimental Study

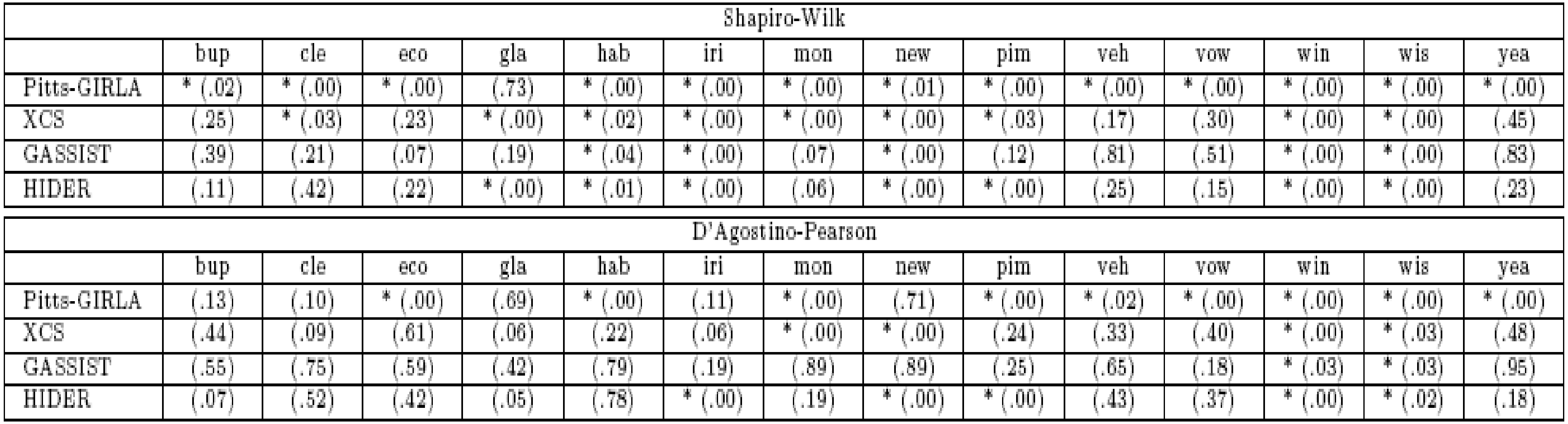
- We have selected 14 data sets from UCI repository.

Data set	#Ex.	#Atts.	#C.
bupa (bup)	345	6	2
cleveland (cle)	297	13	5
ecoli (eco)	336	7	8
glass (gla)	214	9	7
haberman (hab)	306	3	2
iris (iri)	150	4	3
monk-2 (mon)	432	6	2
new-Thyroid (new)	215	5	3
pima (pim)	768	8	2
vehicle (veh)	846	18	4
vowel (vow)	988	13	11
wine (win)	178	13	3
wisconsin (wis)	683	9	2
yeast (yea)	1484	8	10

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

TABLE I. Normality condition in accuracy



Shapiro-Wilk														
	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Pitts-GIRLA	* (.02)	* (.00)	* (.00)	(.73)	* (.00)	* (.00)	* (.00)	* (.01)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)
XCS	(.25)	* (.03)	(.23)	* (.00)	* (.02)	* (.00)	* (.00)	* (.00)	* (.03)	(.17)	(.30)	* (.00)	* (.00)	(.45)
GASSIST	(.39)	(.21)	(.07)	(.19)	* (.04)	* (.00)	(.07)	* (.00)	(.12)	(.81)	(.51)	* (.00)	* (.00)	(.83)
HIDER	(.11)	(.42)	(.22)	* (.00)	* (.01)	* (.00)	(.06)	* (.00)	* (.00)	(.25)	(.15)	* (.00)	* (.00)	(.23)

D'Agostino-Pearson														
	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Pitts-GIRLA	(.13)	(.10)	* (.00)	(.69)	* (.00)	(.11)	* (.00)	(.71)	* (.00)	* (.02)	* (.00)	* (.00)	* (.00)	* (.00)
XCS	(.44)	(.09)	(.61)	(.06)	(.22)	(.06)	* (.00)	* (.00)	(.24)	(.33)	(.40)	* (.00)	* (.03)	(.48)
GASSIST	(.55)	(.75)	(.59)	(.42)	(.79)	(.19)	(.89)	(.89)	(.25)	(.65)	(.18)	* (.03)	* (.03)	(.95)
HIDER	(.07)	(.52)	(.42)	(.05)	(.78)	* (.00)	(.19)	* (.00)	* (.00)	(.43)	(.37)	* (.00)	* (.02)	(.18)

a value smaller than 0.05 indicates that you can reject the **null-hypothesis** (i.e. the normality condition is not satisfied) and it is noted with "*"

S. García, A. Fernandez, 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.*

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

GBML Case of Study: some facts

- Conditions needed for the application of parametric tests are not fulfilled in some cases.
  - The size of the sample should be enough (50)

- One main factor: the nature of the problem
- Graphically, we can use Q-Q graphics and histograms to see the normality

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

Analyzing parametric tests

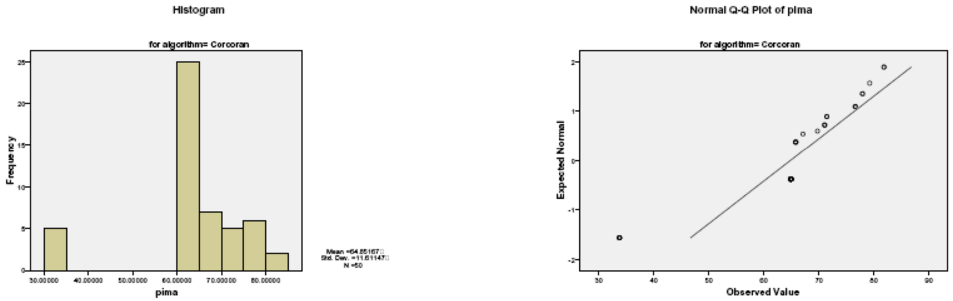


Figure 1: Results of Pitts-GIRLA over pima data set in 10fcv: Histogram and Q-Q Graphic.

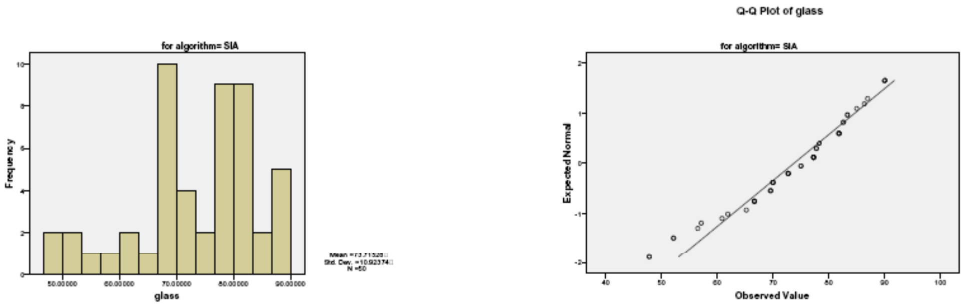


Figure 2: Results of SIA over glass data set in 10fcv: Histogram and Q-Q Graphic.

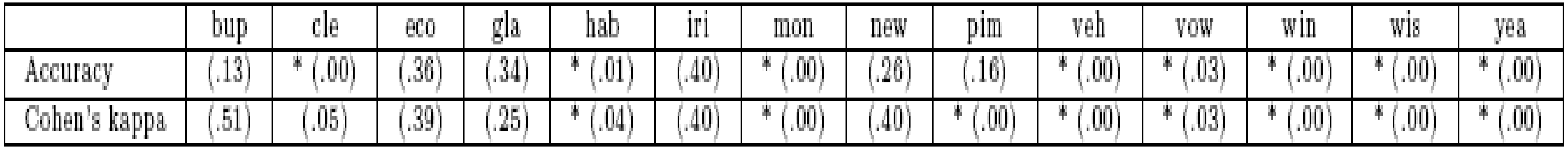
* A Q-Q graphic represents a confrontation between the quartiles from data observed and those from the normal distributions.

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

TABLE 2. Test of HETEROSCEDASTICITY OF LEVENE

(BASED ON MEANS)



	bup	cle	eco	gla	hab	iri	mon	new	pim	veh	vow	win	wis	yea
Accuracy	(.13)	* (.00)	(.36)	(.34)	* (.01)	(.40)	* (.00)	(.26)	(.16)	* (.00)	* (.03)	* (.00)	* (.00)	* (.00)
Cohen's kappa	(.51)	(.05)	(.39)	(.25)	* (.04)	(.40)	* (.00)	(.40)	* (.00)	* (.00)	* (.03)	* (.00)	* (.00)	* (.00)

Table 2 shows the results by applying Levene's tests, where the symbol "*" indicates that the variances of the distributions of the different algorithms for a certain function are not homogeneities (we reject the null hypothesis).

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Data Mining Experiments

NN and GBML do not verify parametric conditions.

Similar studies can be presented with other learning algorithms.

Conditions for the safe use of parametric tests

- Theoretical background
- Checking the conditions in Data Mining Experiments
- Checking the conditions in Parameter Optimization Experiments

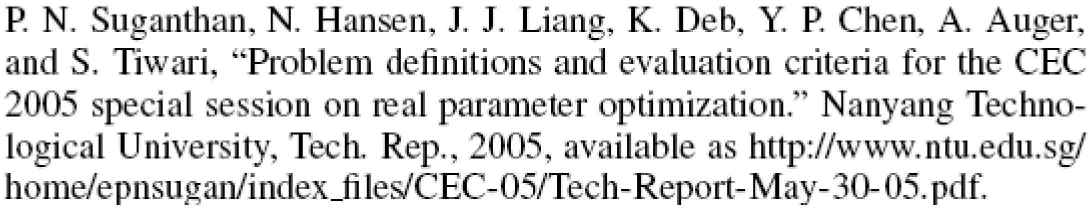
Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

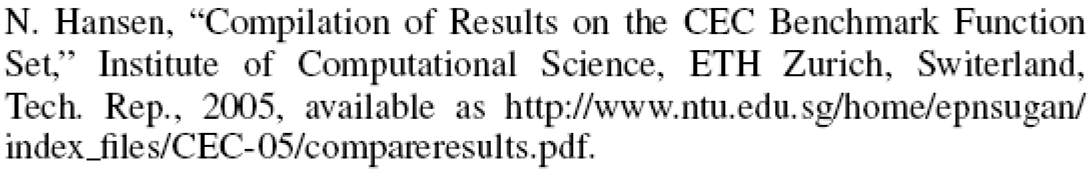
Special Session on Real-Parameter Optimization at CEC-05, Edinburgh, UK, 2-5 Sept. 2005

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*, 13:10 (2009) 959-977.

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

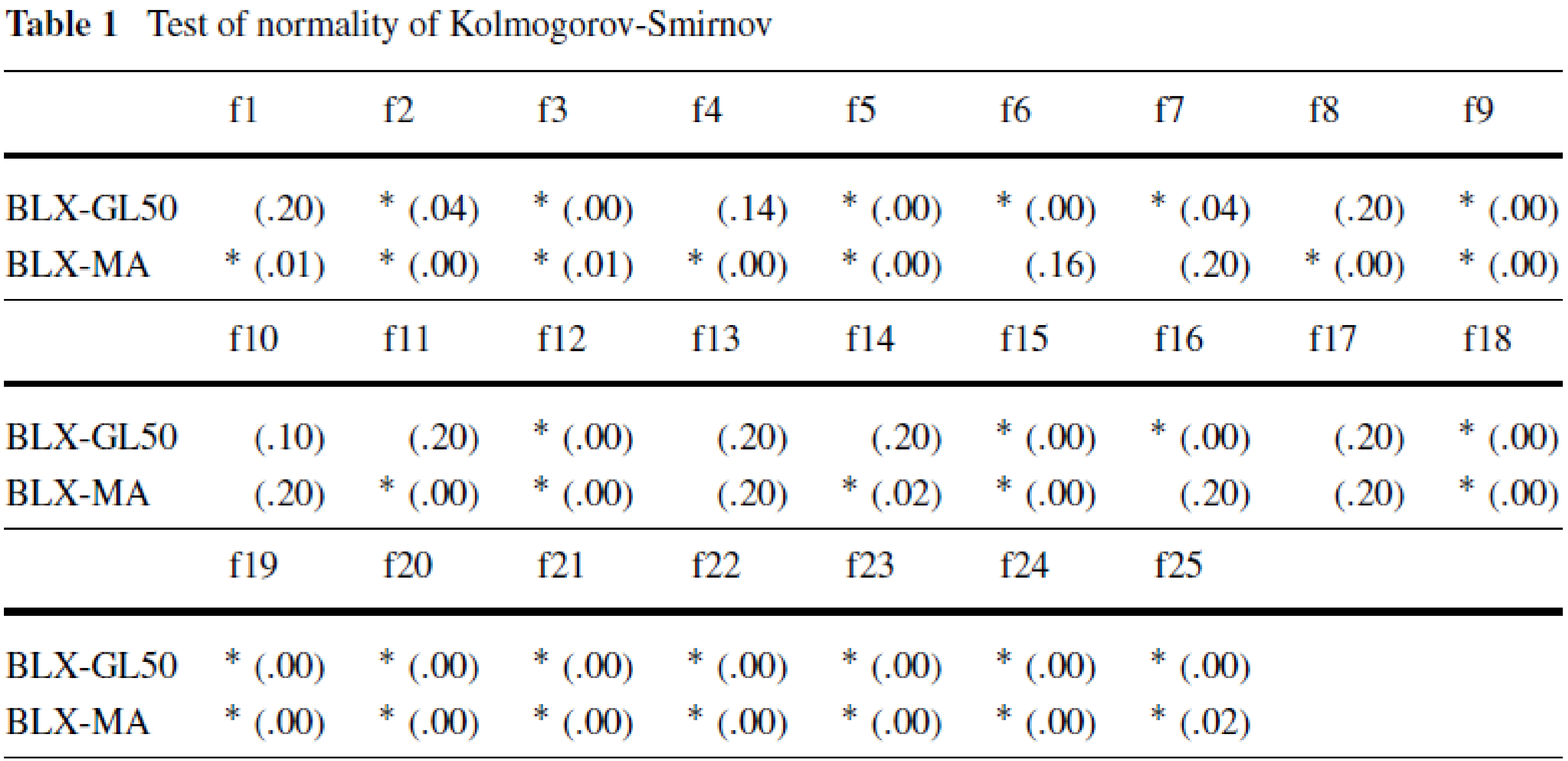


Table 1 Test of normality of Kolmogorov-Smirnov

	f1	f2	f3	f4	f5	f6	f7	f8	f9
BLX-GL50	(.20)	* (.04)	* (.00)	(.14)	* (.00)	* (.00)	* (.04)	(.20)	* (.00)
BLX-MA	* (.01)	* (.00)	* (.01)	* (.00)	* (.00)	(.16)	(.20)	* (.00)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
BLX-GL50	(.10)	(.20)	* (.00)	(.20)	(.20)	* (.00)	* (.00)	(.20)	* (.00)
BLX-MA	(.20)	* (.00)	* (.00)	(.20)	* (.02)	* (.00)	(.20)	(.20)	* (.00)
	f19	f20	f21	f22	f23	f24	f25		
BLX-GL50	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)		
BLX-MA	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.02)		

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

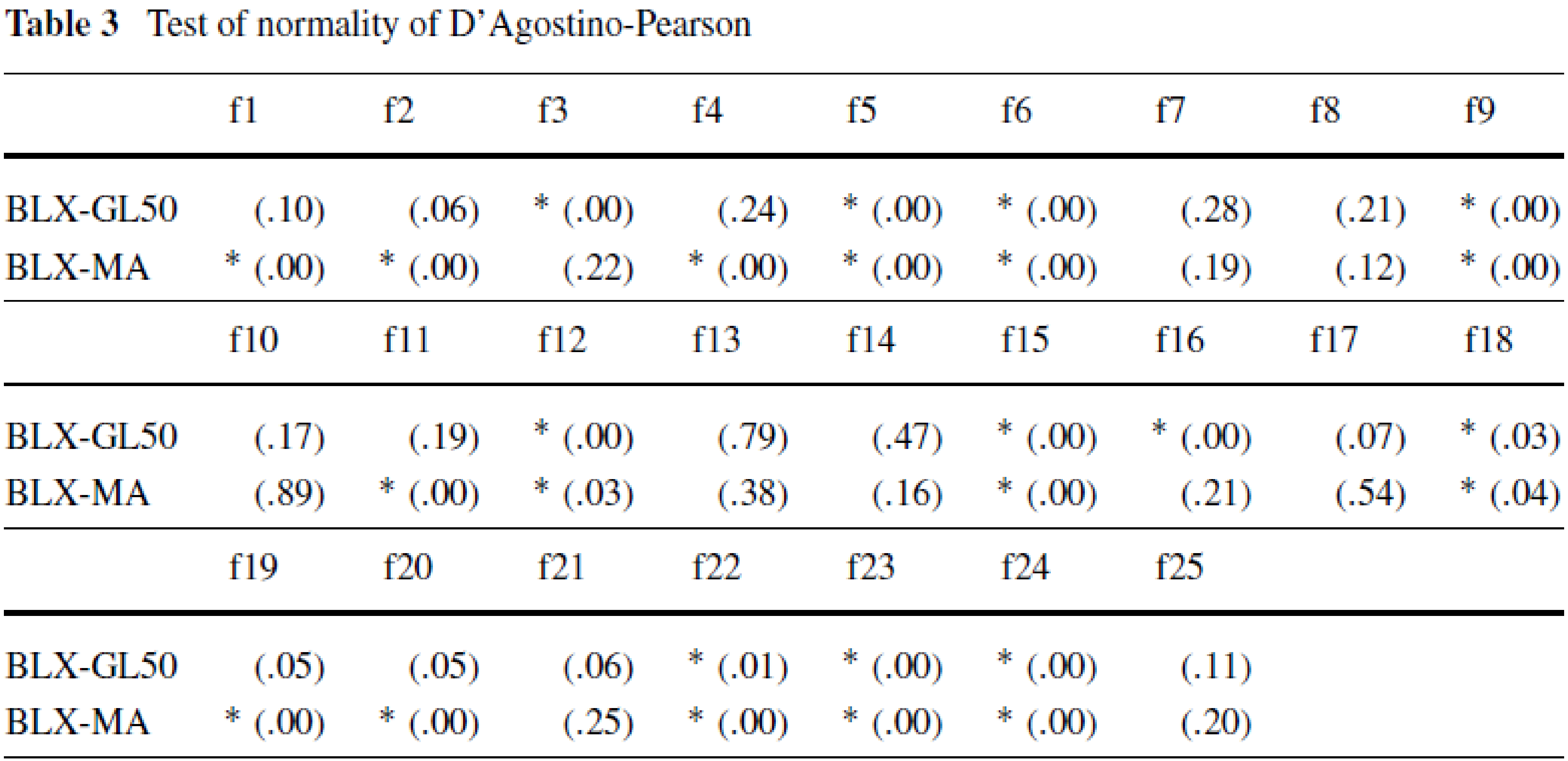


Table 3 Test of normality of D'Agostino-Pearson

	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

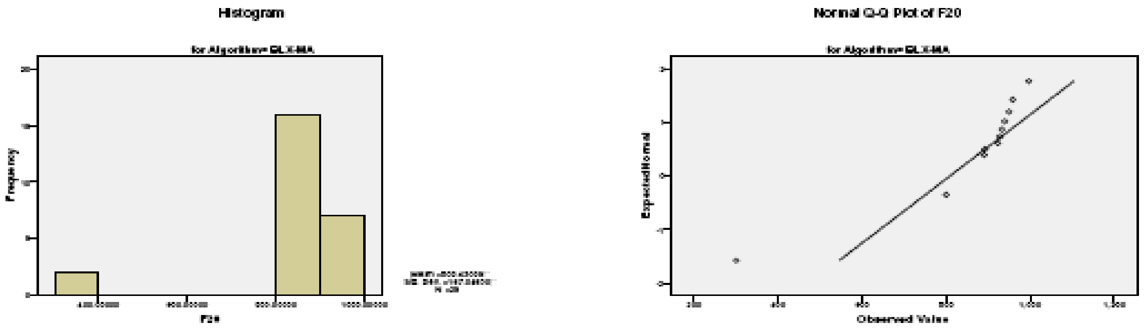


Figure 1: Example of non-normal distribution: Function f20 and BLX-GL50 algorithm: Histogram and Q-Q Graphic.

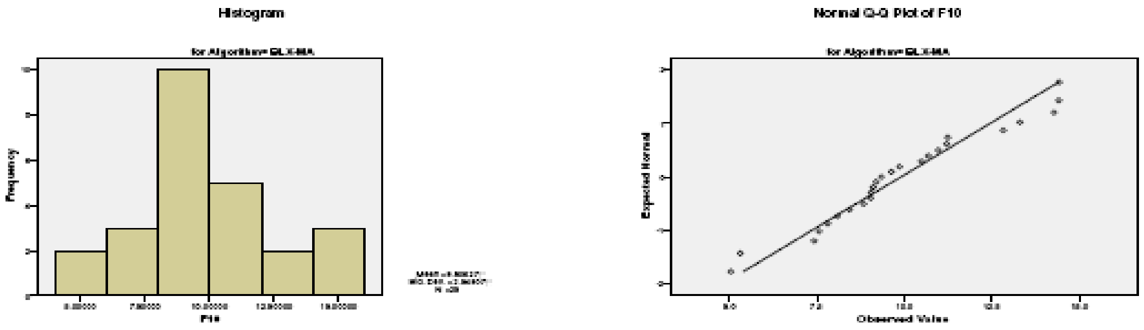


Figure 2: Example of normal distribution: Function f10 and BLX-MA algorithm: Histogram and Q-Q Graphic.

Conditions for the Safe Use of Parametric Tests

Checking the Conditions in Parameter Optimization Experiments

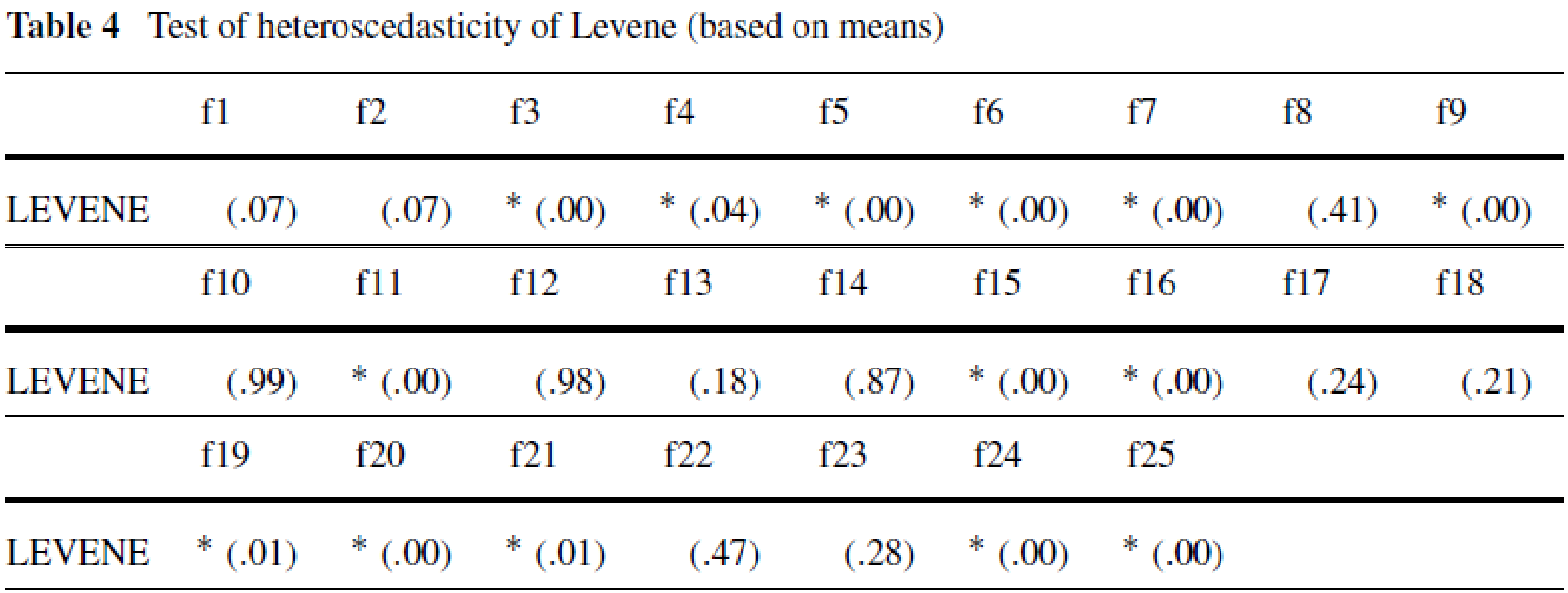


Table 4 Test of heteroscedasticity of Levene (based on means)

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

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
- Basic non-parametric tests and case studies
- Advanced non-parametric tests and case studies
- Lessons Learned
- Books of Interest and References
- Software

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

OUTLINE (I)

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
  - Theoretical background
  - Checking the conditions in Data Mining Experiments
  - Checking the conditions in Parameter Optimization Experiments
- **Basic non-parametric tests and case studies**
  - For Pairwise Comparisons
  - For Multiple Comparisons involving control method
  - Data Mining: Neural Networks and Genetic Learning
  - Evolutionary Algorithms: CEC'05 Special Session on Parameter Optimization

Basic Non-Parametric Tests and Case Studies

- For Pairwise Comparisons
- For Multiple Comparisons involving a Control Method
- Data Mining: Neural Networks and Genetic Learning
- Evolutionary Algorithms: CEC’05 Special Session of Parameter Optimization

Pairwise Comparisons involve Two-Sample Tests

When comparing means of two samples to make inferences about differences between two populations, there are 4 main tests that could be used:

	Unpaired data	Paired data
Parametric test	Independent-Samples T-Test	Paired-Samples T-Test
Non-parametric test	Mann-Whitney U test (or Wilcoxon rank-sum test)	Wilcoxon Signed-Ranks test (*Also, Sign test*)

Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

Pairwise Comparisons involve Two-Sample Tests

unpaired data
(independent)

group A	group B
4.23	2.51
3.21	3.3
3.63	3.75
4.42	3.22
4.08	3.99
3.98	3.65

paired data
(related)

initial data #	conventional	proposed
1	4.23	2.51
2	3.21	3.30
3	3.63	3.75
4	4.42	3.22
5	4.08	3.99
6	3.98	3.65

nake inferences about
ıre 4 main tests that could

	Paired data
s	Paired-Samples T-Test
t	Wilcoxon Signed-Ranks test (*Also, Sign test*)

Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

(1) Sign Test

significance test between the # of winnings and losses

(2) Wilcoxon's Signed Ranks Test

significance test using both the # of winnings and losses and the level of winnings/losses

data of 2 groups		# of winnings and losses		the level of winnings/losses
173	174	-	+	-1
143	137	+	-	+6
158	151	+	-	+7
156	143	+	-	+13
176	180	-	+	-4
165	162	+	-	+3

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.

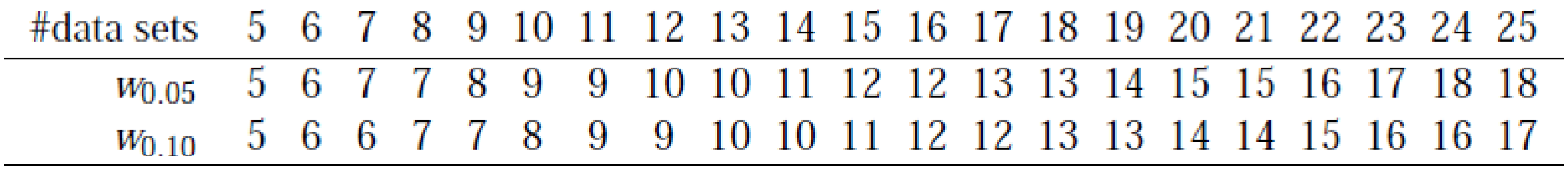
For a greater number of datasets/problems, the number of wins is under the null-hypothesis distributed according to $N(N/2, \sqrt{N}/2)$.

The Sign Test

1. Calculate the # of winnings and losses by comparing runs with the same initial data.

2. Check a sign test table to show significance of two methods.

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

The number of wins are distributed following a binomial distribution.

Basic Non-Parametric Tests and Case Studies

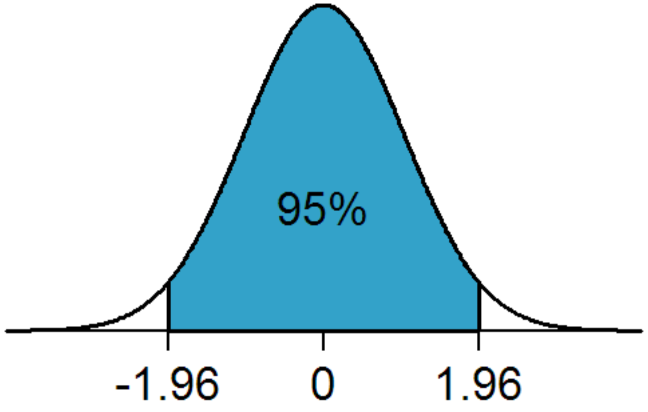
For Pairwise Comparisons

The Sign Test

For a greater number of datasets/problems, the number of wins is under the null-hypothesis distributed according to $N(N/2, \sqrt{N}/2)$.

Thus, if an algorithm obtains a number of wins which is at least

N/2 + $1.96\sqrt{N}/2$ the algorithm is significantly better with α=0.05. Tieds are split between the two algorithms. If they are an odd number, one is ignored.



Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

Example of the Sign Test (Demsar 2006, JMLR)

dataset	C4.5	C4.5m	Sign
Adult	0.763	0.768	+
Breast	0.599	0.591	-
Wisconsin	0.954	0.971	+
Cmc	0.628	0.661	+
Ionosphere	0.882	0.888	+
Iris	0.936	0.931	-
Bupa	0.661	0.668	+
Lung	0.583	0.583	=
Lymphography	0.775	0.838	+
Mushroom	1.000	1.000	=
Tumor	0.940	0.962	+
Rheum	0.619	0.666	+
Voting	0.972	0.981	+
Wine	0.957	0.978	+

Classification problem with 14 datasets.

C4.5 standard vs C4.5 with *m* parameter (minimum number of examples for creating a leaf) tuned for AUC measure.

Number of wins of C4.5m = 10

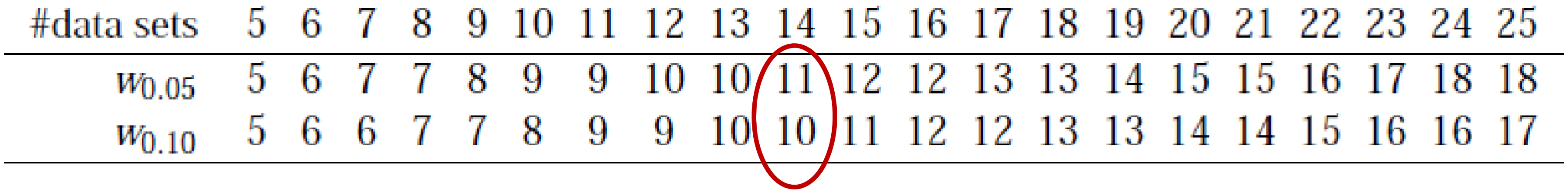
Number of loses of C4.5m = 2

Number of ties = 2

Moreover, one tie is added in the wins count. No. of wins = 11.

Example of Sign Test

According to the previous Table, this difference is significant with α = 0.05.



#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

This test does not assume any commensurability of scores or differences nor does it assume normal distributions and is thus applicable to any data. On the other hand, it is much weaker than the Wilcoxon signed-ranks test because it will not reject the null-hypothesis unless one algorithm almost always outperforms the other.

Exercise of Sign Test

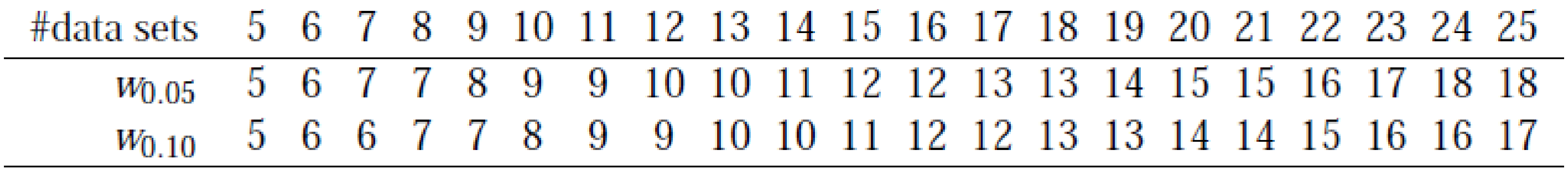
Check the significance of:

16 vs. 4

14 vs. 1

9 vs. 3

18 vs. 5



#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

Procedure for the Wilcoxon Signed-Ranks Test

1. For each pair of data, calculate the difference. Keep track of the sign (+ve or –ve).

2. Temporarily ignoring the sign of the difference, rank the absolute values of the difference. When the differences have the same value, assign them the mean of the ranks involved in the tie.

3. Consider the sign of the differences again and ADD up the ranks of all the positive differences and all the negative differences (R^+, R^-). Ranks of difference equal to 0 are split evenly among the sums; if there is an odd number of them, one is ignored.

Basic Non-Parametric Tests and Case Studies

For Pairwise Comparisons

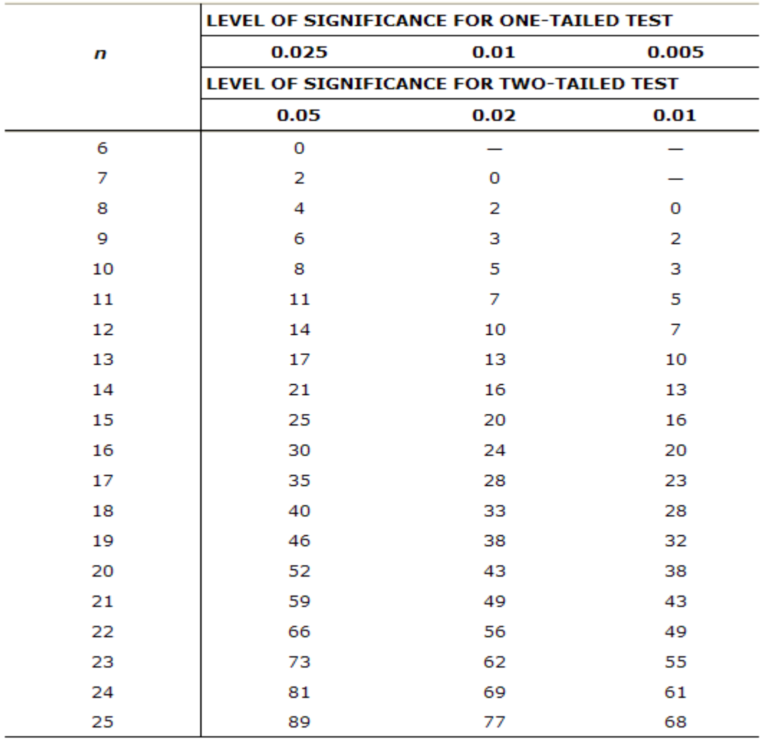
Procedure for the Wilcoxon Signed-Ranks Test

4. Let T be the **smaller** of the sums of positive and negative differences. T = Min $\{R^+, R^-\}$.

Use an appropriate Statistical Table or computer to determine the test statistic, critical region or p-values.

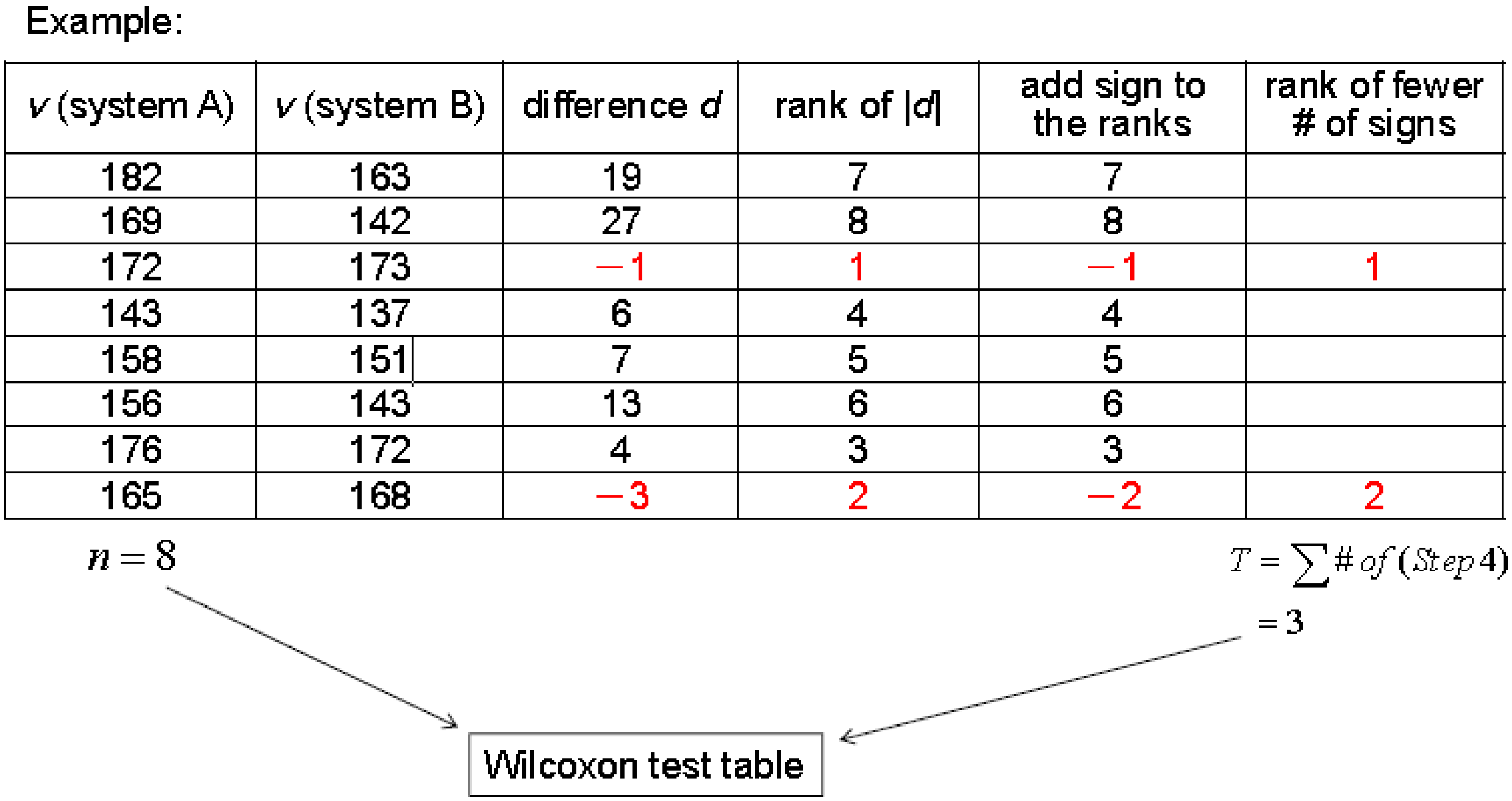
5. Reject the H_o if test statistic ≤ critical value, or if p ≤ α (alpha).

6. Report Test results.



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



Example:

v (system A)	*v* (system B)	difference *d*	rank of |*d*|	add sign to the ranks	rank of fewer # of signs
182	163	19	7	7	
169	142	27	8	8	
172	173	−1	1	−1	1
143	137	6	4	4	
158	151	7	5	5	
156	143	13	6	6	
176	172	4	3	3	
165	168	−3	2	−2	2

$n = 8$

$T = \sum \# \, of \, (Step \, 4)$

$= 3$

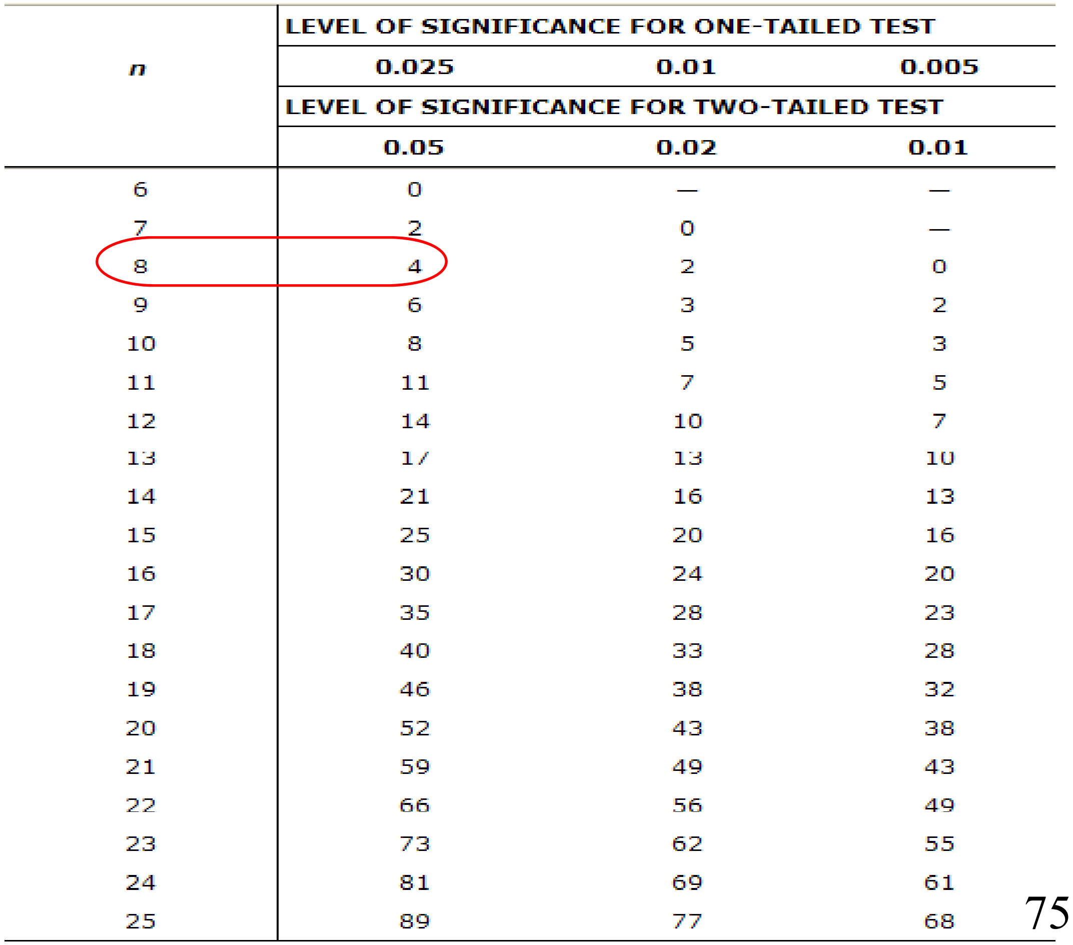
Wilcoxon test table

Example of the Wilcoxon Signed-Ranks Test

n = 8

T= 3

α= 0.05, dif = 4



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

(Demsar 2006, JMLR)

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

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

$6+ 14+ 11 + 13 +$

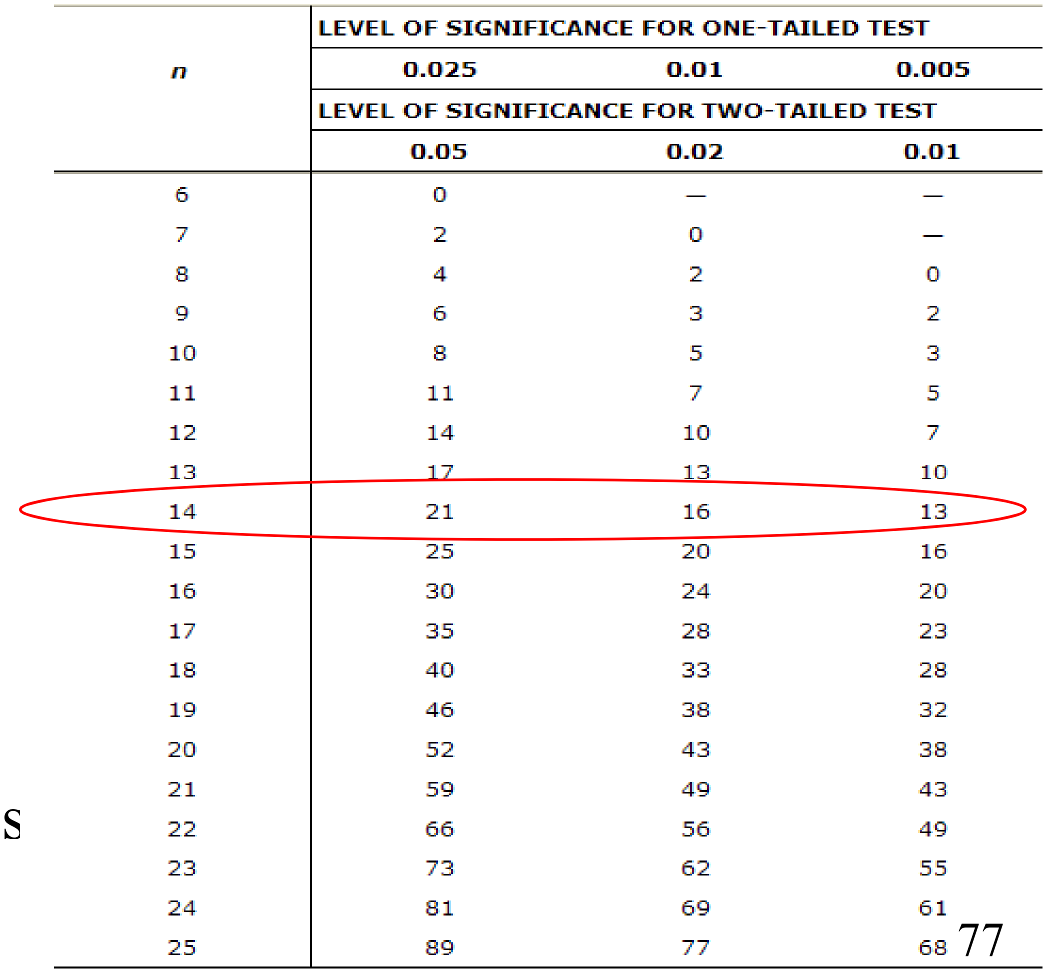
$8 + 10 + 1.5 = 93$

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

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

$\alpha = 0.05$, $N = 14$ dif $= 21$

We reject the null-hypothesis

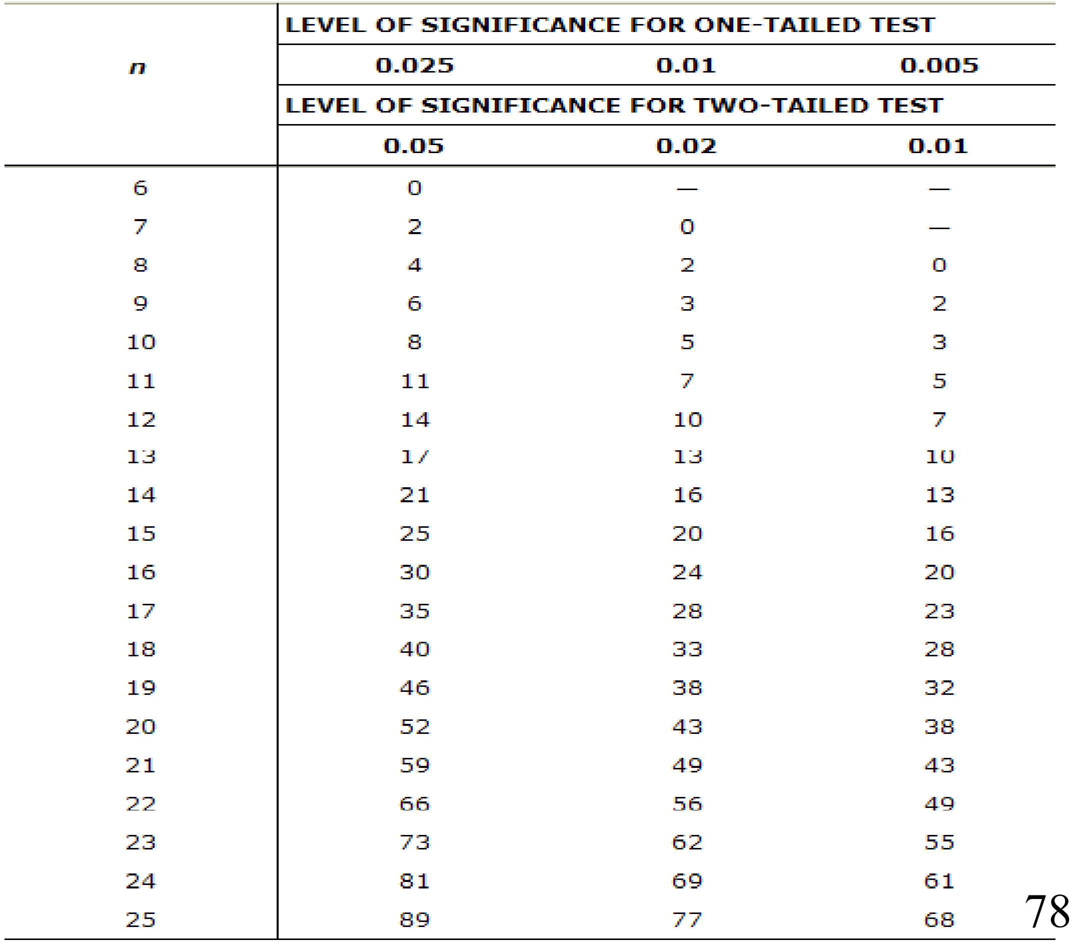


n	LEVEL OF SIGNIFICANCE FOR ONE-TAILED TEST		
	0.025	0.01	0.005
	LEVEL OF SIGNIFICANCE FOR TWO-TAILED TEST		
	0.05	0.02	0.01
6	0	—	—
7	2	0	—
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66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

Example of the Wilcoxon Signed-Ranks Test

Critical values for T for N up to 25.

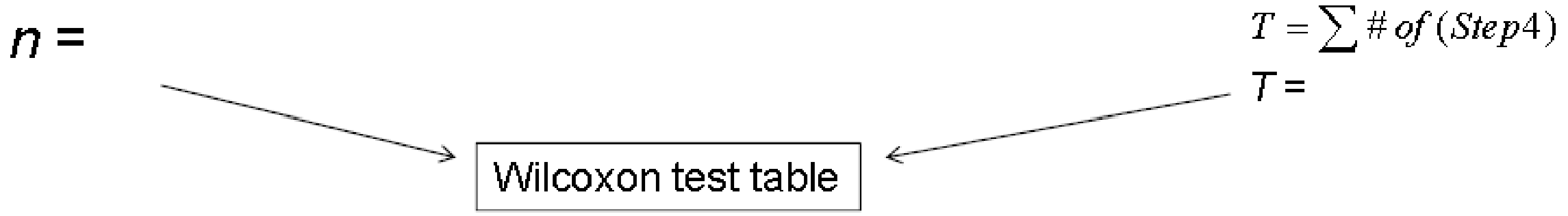
It T <= dif (table-value) then Reject the H_o



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

Exercise 1: Wilcoxon Signed-Ranks Test

v (system A)	*v* (system B)	difference *d*	rank of |*d*|	add sign to the ranks	rank of fewer # of signs
182	163				
169	142				
173	172				
143	137				
158	151				
156	143				
176	172				
165	168				



$n =$

$T = \sum \# of\,(Step4)$

$T =$

Wilcoxon test table

Exercise 1: Wilcoxon Signed-Ranks Test

v (system A)	*v* (system B)	difference *d*	rank of |*d*|	add sign to the ranks	rank of fewer # of signs
182	163	19	7	7	
169	142	27	8	8	
173	172	1	1	1	
143	137	6	4	4	
158	151	7	5	5	
156	143	13	6	6	
176	172	4	3	3	
165	168	-3	2	-2	2

$n = 8$

$T = \sum \# of\,(Step4)$

$T = 2$

Wilcoxon test table

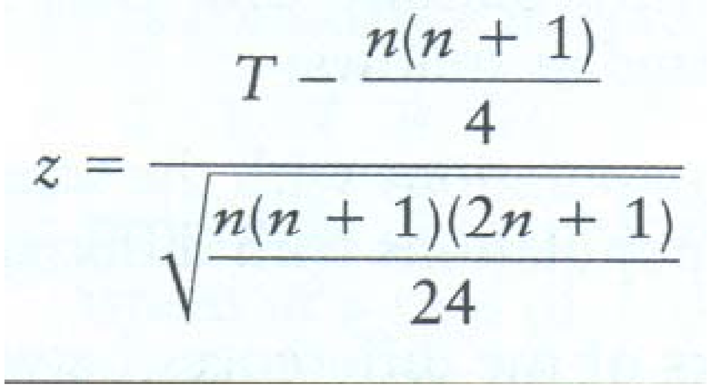
Basic Non-Parametric Tests and Case Studies

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)



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

with $\alpha = 0.05$, 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

Wilcoxon Signed-Ranks Test in SPSS

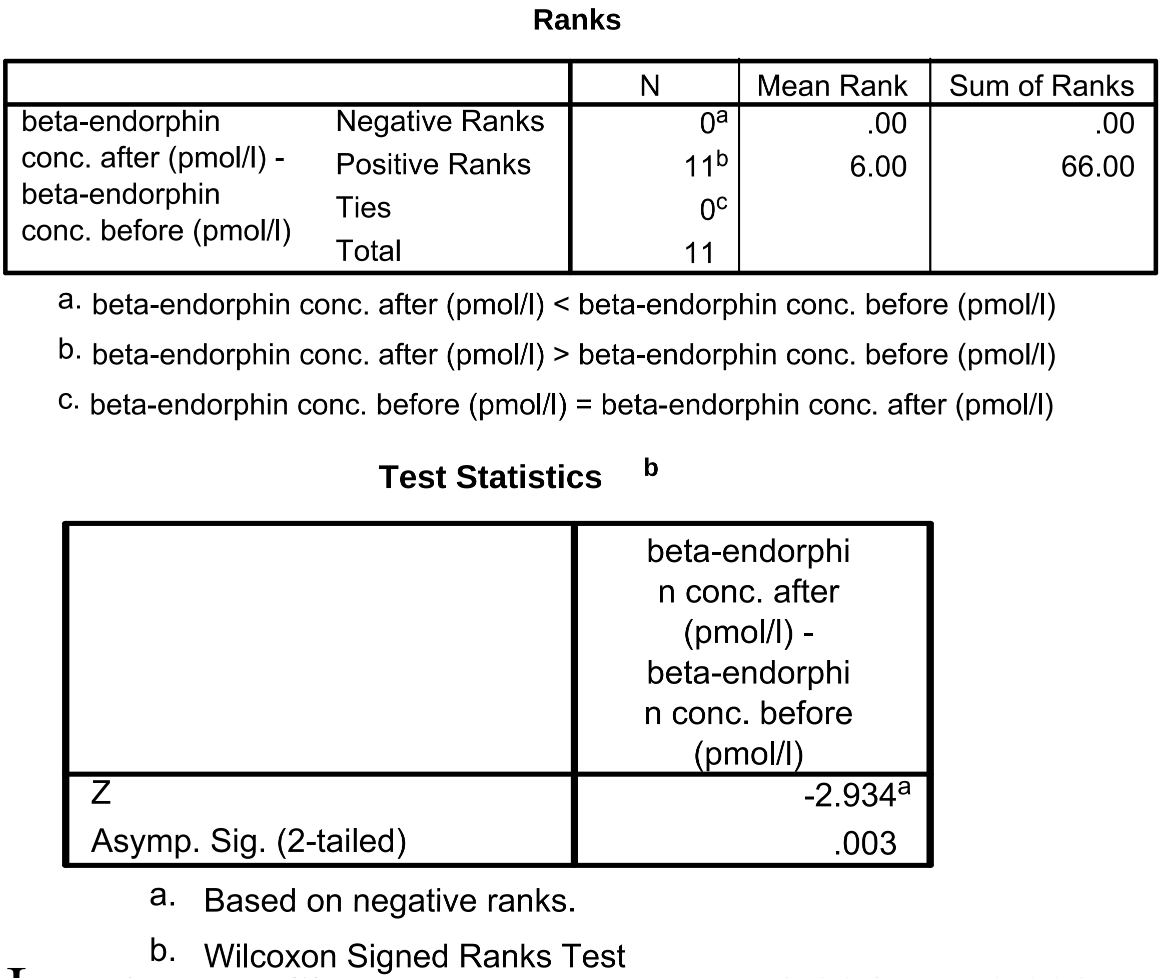
Analyze →Nonparametric Tests → 2 Related Samples Tests

- Select pair(s) of variables

- Select Wilcoxon

Wilcoxon Signed-Ranks Test in SPSS

OUTPUT



Ranks

		N	Mean Rank	Sum of Ranks
beta-endorphin conc. after (pmol/l) - beta-endorphin conc. before (pmol/l)	Negative Ranks	0[a]	.00	.00
	Positive Ranks	11[b]	6.00	66.00
	Ties	0[c]		
	Total	11		

a. beta-endorphin conc. after (pmol/l) < beta-endorphin conc. before (pmol/l)

b. beta-endorphin conc. after (pmol/l) > beta-endorphin conc. before (pmol/l)

c. beta-endorphin conc. before (pmol/l) = beta-endorphin conc. after (pmol/l)

Test Statistics[b]

	beta-endorphin conc. after (pmol/l) - beta-endorphin conc. before (pmol/l)
Z	-2.934[a]
Asymp. Sig. (2-tailed)	.003

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Conclude: Reject H_o (Wilcoxon Signed-Ranks test, Z = -2.934, p = 0.003, n = 11, 0).

Basic Non-Parametric Tests and Case Studies

- For Pairwise Comparisons
- For Multiple Comparisons involving a Control Method
- Data Mining: Neural Networks and Genetic Learning
- Evolutionary Algorithms: CEC'05 Special Session of Parameter Optimization

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 – α), then
the probability of not making a Type I error in the k-1 comparison is (1- α)·(k-1). Then the probability of making one or more Type I error is 1 - (1- α)·(k-1).

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

Using Wilcoxon test for comparing multiple pairs of algorithms:

The true statistical signification for the pairwise comparison test is given by:

$$
\begin{aligned}
p = P(Reject\ H_0|H_0\ true) = \\
= 1 - P(Accept\ H_0|H_0\ true) = \\
= 1 - P(Accept\ A_k = A_i, i = 1, \ldots, k-1|H_0\ true) = \\
= 1 - \prod_{i=1}^{k-1} P(Accept\_A_k = A_i|H_0\ true) = \\
= 1 - \prod_{i=1}^{k-1} [1 - P(Reject\ A_k = A_i|H_0\ true)] = \\
= 1 - \prod_{i=1}^{k-1} (1 - p_{H_i})
\end{aligned}
$$

Using Multiple Comparison Procedures:

Making pairwise comparisons allows us to conduct this analysis, but the experiment wise error can not be previously controlled. Furthermore, a pairwise comparison is not influenced by any external factor, whereas in a multiple comparison, the set of algorithms chosen can determine the results of the analysis.

Multiple comparison procedures are designed for allowing us to fix the FWER before performing the analysis and for taking into account all the influences that can exist within the set of results for each algorithm.

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Multiple Comparison Procedures:

Parametric	Nonparametric
ANOVA	Friedman’s test Iman-Davenport’s test
Turkey, Dunnet, …	Bonferroni-Dunn’s test Holm’s method Hochberg’s method

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 *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).

Example of the Friedman Test

(ex.) Comparison of recognition rates.

benchmark tasks	methods			
	a	*b*	*c*	*d*
A	0.92	0.75	0.65	0.81
B	0.48	0.45	0.41	0.52
C	0.56	0.41	0.47	0.50
D	0.61	0.50	0.56	0.54

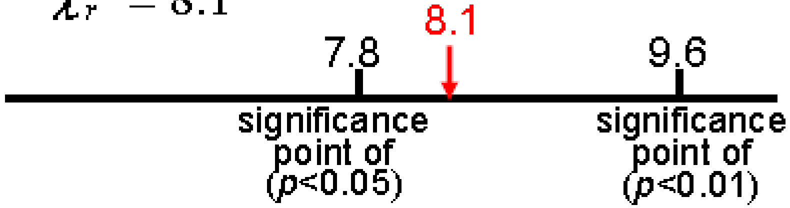
Rankings are assigned in decreasing order

benchmark tasks	method			
	a	b	c	d
A	4	2	1	3
B	3	2	1	4
C	4	1	2	3
D	4	1	3	2
Σ	15	6	7	12

# of data (n = 4)

# of methods (k = 4)

$\chi_r^2 = 8.1$


8.1
7.8
9.6
significance point of (p<0.05)
significance point of (p<0.01)

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Example of the Friedman Test

(Demsar 2006, JMLR)

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

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

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

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

Iman and Davenport's measure

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

$F_F = 3.69$, $F(3,3\text{x}13) = 2.85$

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 og algorithms.

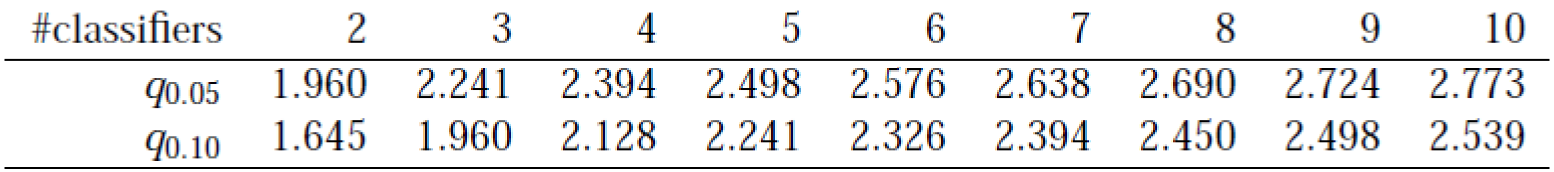
Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

The performance of two algorithms is significantly different if the corresponding average ranks differ by at least the critical difference:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

If the CD is greater than the values presented in the following Table, we can conclude that both algorithms have differences in performance:



#classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.241	2.394	2.498	2.576	2.638	2.690	2.724	2.773
$q_{0.10}$	1.645	1.960	2.128	2.241	2.326	2.394	2.450	2.498	2.539

(b) Critical values for the two-tailed Bonferroni-Dunn test; the number of classifiers include the control classifier.

Basic Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Considering the example of the four versions of C4.5, we have (C4.5cf,m as control):

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

$$CD_{\alpha=0.05} = 2.394\sqrt{\frac{4 \cdot 5}{6 \cdot 14}} = 1.16$$

$$CD_{\alpha=0.1} = 2.128\sqrt{\frac{4 \cdot 5}{6 \cdot 14}} = 1.038$$

With α=0.05, C4.5cf,m performs better than C4.5.

With α=0.1, C4.5cf,m also performs better than C4.5.

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:

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

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

Holm's method: SE = √(4.5/6.14) = 0.488.

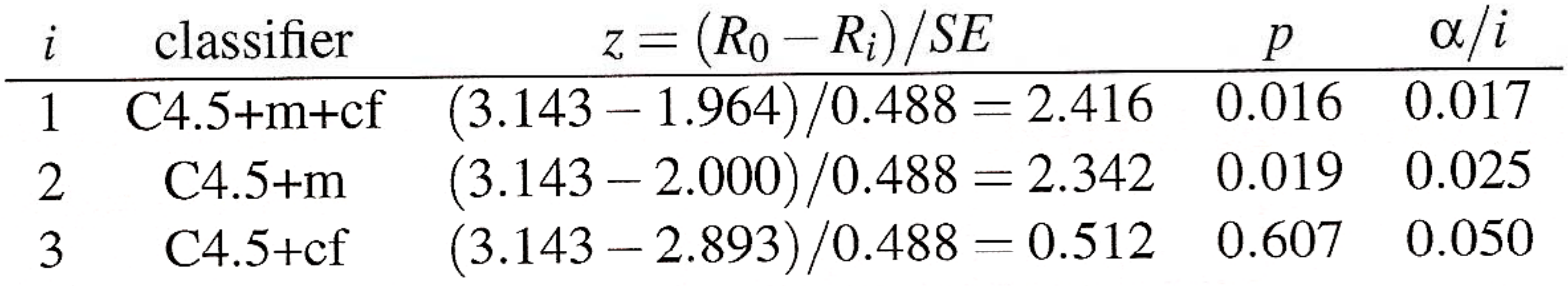
$$z = (R_i - R_j) \Bigg/ \sqrt{\frac{k(k+1)}{6N}}\,.$$

p-values are

0.016 (C4.5+m+cf)

0.019 (C4.5+m)

0.607 (C4.5+cf)



i	classifier	$z = (R_0 - R_i)/SE$	p	α/i
1	C4.5+m+cf	$(3.143 - 1.964)/0.488 = 2.416$	0.016	0.017
2	C4.5+m	$(3.143 - 2.000)/0.488 = 2.342$	0.019	0.025
3	C4.5+cf	$(3.143 - 2.893)/0.488 = 0.512$	0.607	0.050

The first one is rejected (0.016 < 0.017)

The second one is rejected (0.019 < 0.025),

The third one can not be rejected (0.607 > 0.05)

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
- Data Mining: Neural Networks and Genetic Learning
- Evolutionary Algorithms: CEC'05 Special Session of Parameter Optimization

Basic Non-Parametric Tests and Case Studies

Data Mining: Neural Networks and Genetic Learning

Wilcoxon Signed-Ranks Test for Paired Samples

Wilcoxon's test applied over the all possible comparisons between the algorithms in accuracy

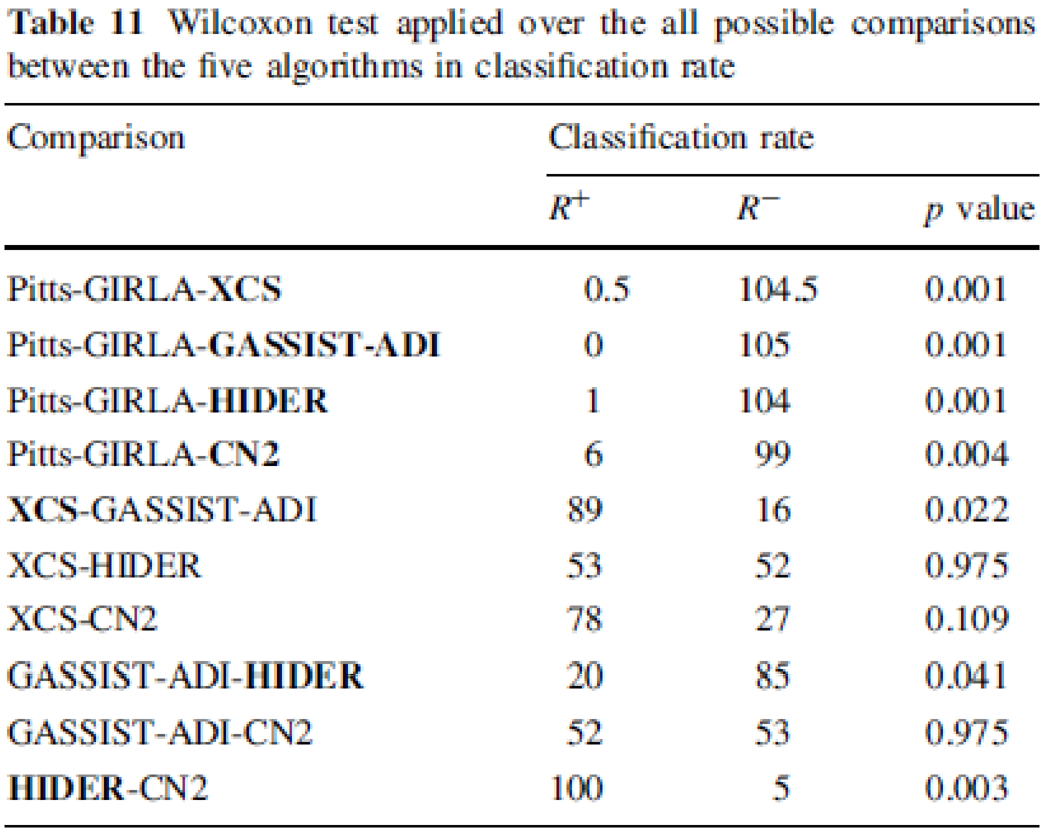


Table 11 Wilcoxon test applied over the all possible comparisons between the five algorithms in classification rate

Comparison	Classification rate		
	R^+	R^-	*p* value
Pitts-GIRLA-**XCS**	0.5	104.5	0.001
Pitts-GIRLA-**GASSIST-ADI**	0	105	0.001
Pitts-GIRLA-**HIDER**	1	104	0.001
Pitts-GIRLA-**CN2**	6	99	0.004
XCS-GASSIST-ADI	89	16	0.022
XCS-HIDER	53	52	0.975
XCS-CN2	78	27	0.109
GASSIST-ADI-**HIDER**	20	85	0.041
GASSIST-ADI-CN2	52	53	0.975
HIDER-CN2	100	5	0.003

We stress in bold **the winner algorithm in each row when**

the *p-value associated is below 0.05*

Wilcoxon Signed-Ranks Test for Paired Samples

Wilcoxon's test applied over the all possible comparisons between the algorithms in kappa rate

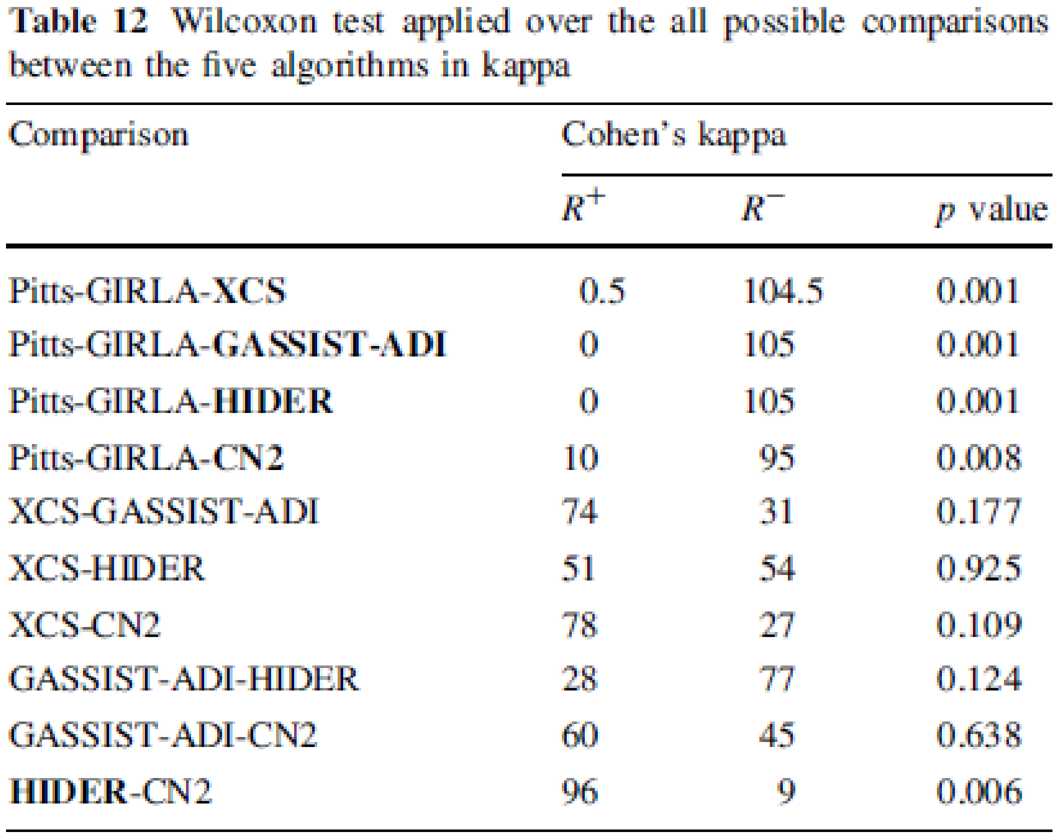


Table 12 Wilcoxon test applied over the all possible comparisons between the five algorithms in kappa

Comparison	Cohen's kappa		
	R^+	R^-	*p* value
Pitts-GIRLA-**XCS**	0.5	104.5	0.001
Pitts-GIRLA-**GASSIST-ADI**	0	105	0.001
Pitts-GIRLA-**HIDER**	0	105	0.001
Pitts-GIRLA-**CN2**	10	95	0.008
XCS-GASSIST-ADI	74	31	0.177
XCS-HIDER	51	54	0.925
XCS-CN2	78	27	0.109
GASSIST-ADI-HIDER	28	77	0.124
GASSIST-ADI-CN2	60	45	0.638
HIDER-CN2	96	9	0.006

We stress in bold **the winner algorithm in each row when**

the *p-value associated is below 0.05*

Basic Non-Parametric Tests and Case Studies

Data Mining: Neural Networks and Genetic Learning

Results of applying Friedman’s and Iman-Davenport’s test with level of significance $\alpha \leq 0.05$ to the GBMLs

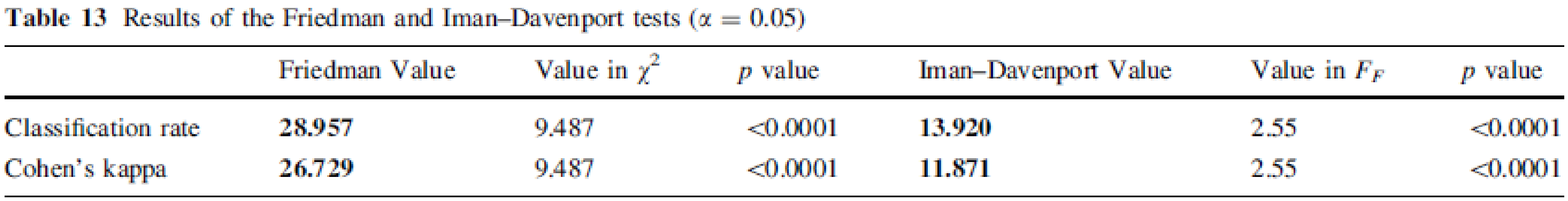


Table 13 Results of the Friedman and Iman–Davenport tests ($\alpha = 0.05$)

	Friedman Value	Value in χ^2	p value	Iman–Davenport Value	Value in F_F	p value
Classification rate	**28.957**	9.487	<0.0001	**13.920**	2.55	<0.0001
Cohen’s kappa	**26.729**	9.487	<0.0001	**11.871**	2.55	<0.0001

- The statistics of Friedman and Iman-Davenport are clearly greater than their associated critical values
  - There are significant differences among the observed results
- Next step: apply **post-hoc** test and find what algorithms partners' average results are dissimilar

Basic Non-Parametric Tests and Case Studies

Data Mining: Neural Networks and Genetic Learning

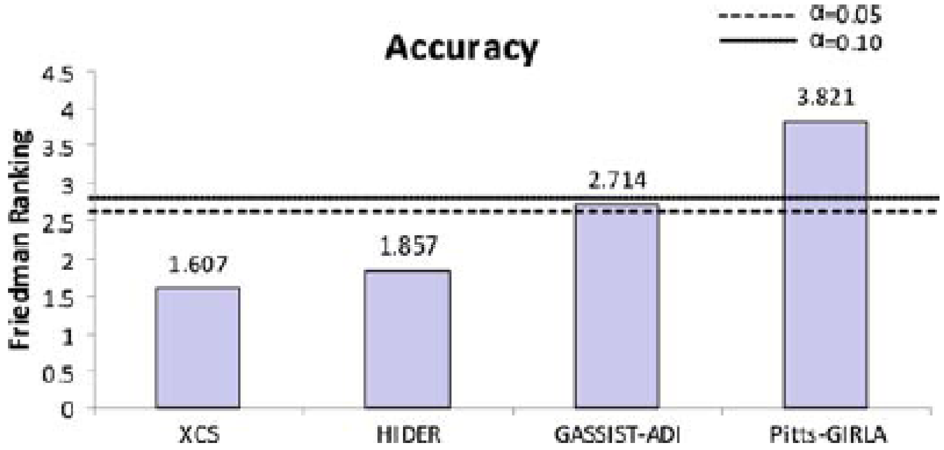


Fig. 3 Bonferroni–Dunn graphic for classification rate

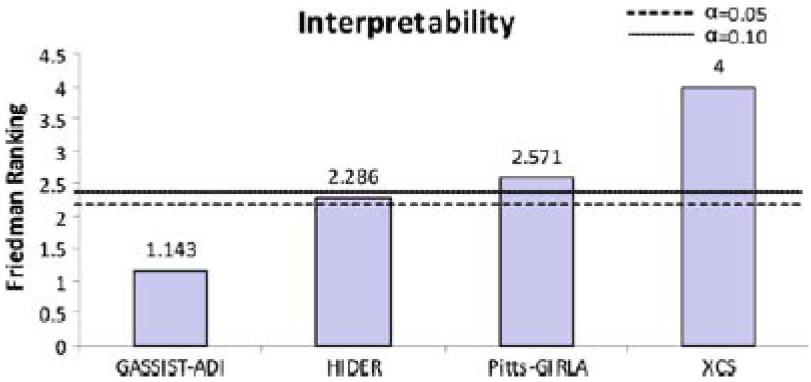


Fig. 5 Bonferroni–Dunn graphic measuring interpretability

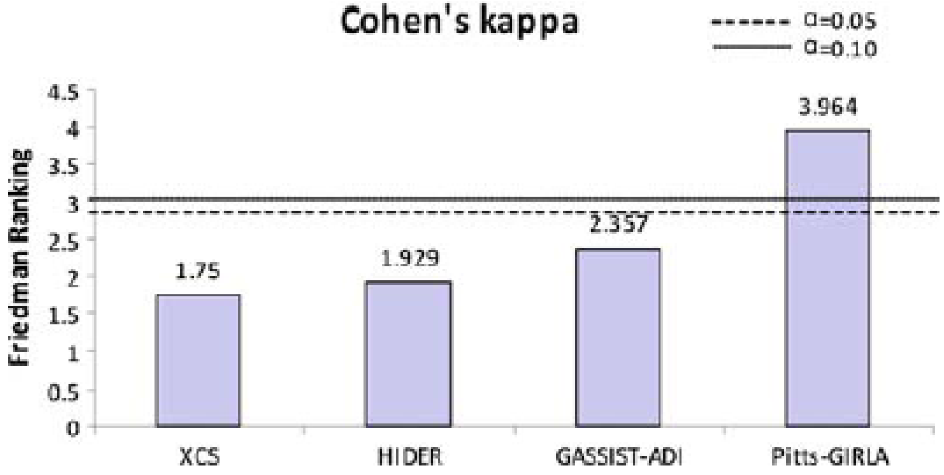


Fig. 4 Bonferroni–Dunn graphic for kappa

Basic Non-Parametric Tests and Case Studies

Data Mining: Neural Networks and Genetic Learning

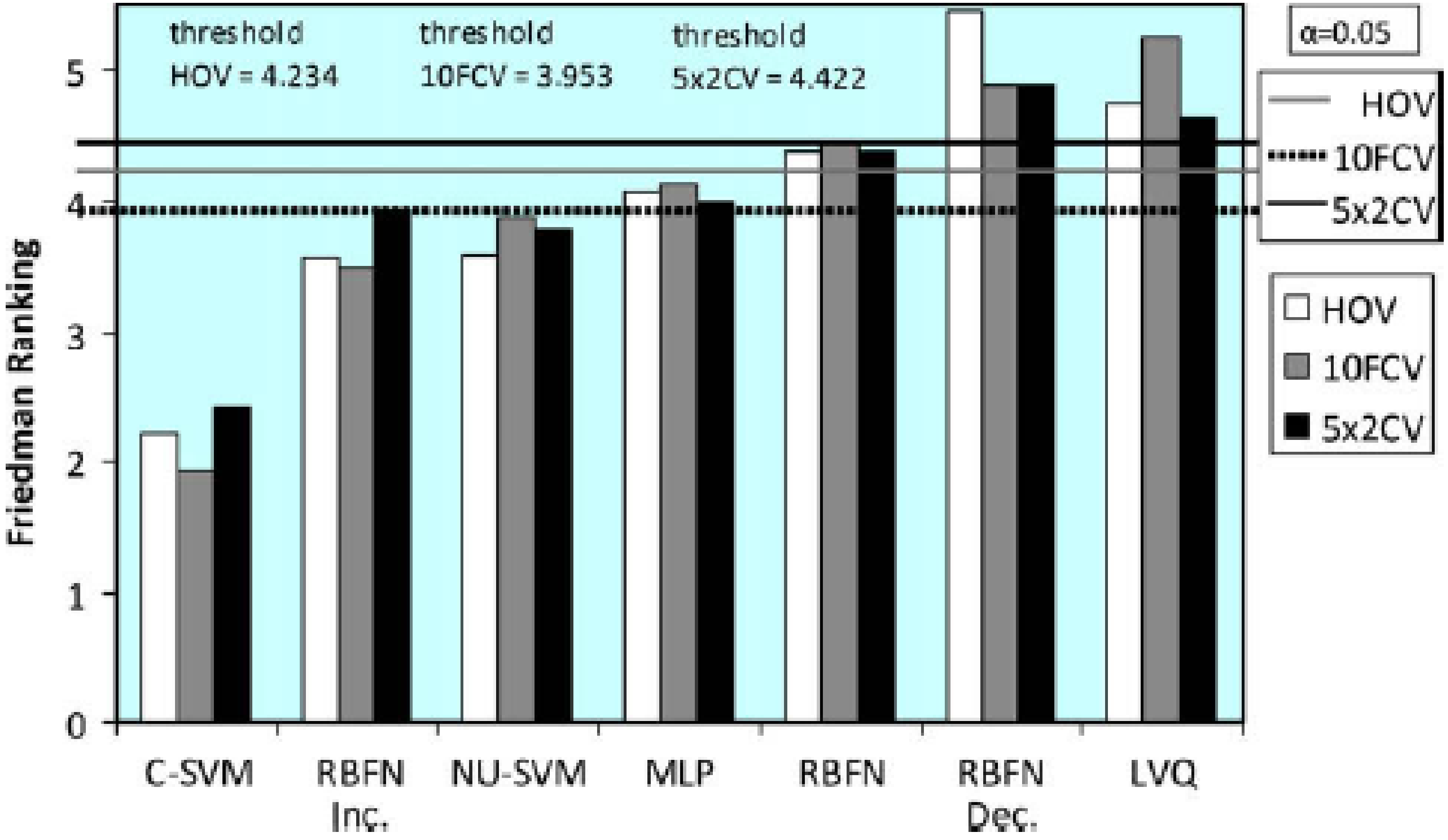


Fig. 5. Bonferroni–Dunn graphic for all validations.

Basic Non-Parametric Tests and Case Studies

- For Pairwise Comparisons
- For Multiple Comparisons involving a Control Method
- Data Mining: Neural Networks and Genetic Learning
- 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

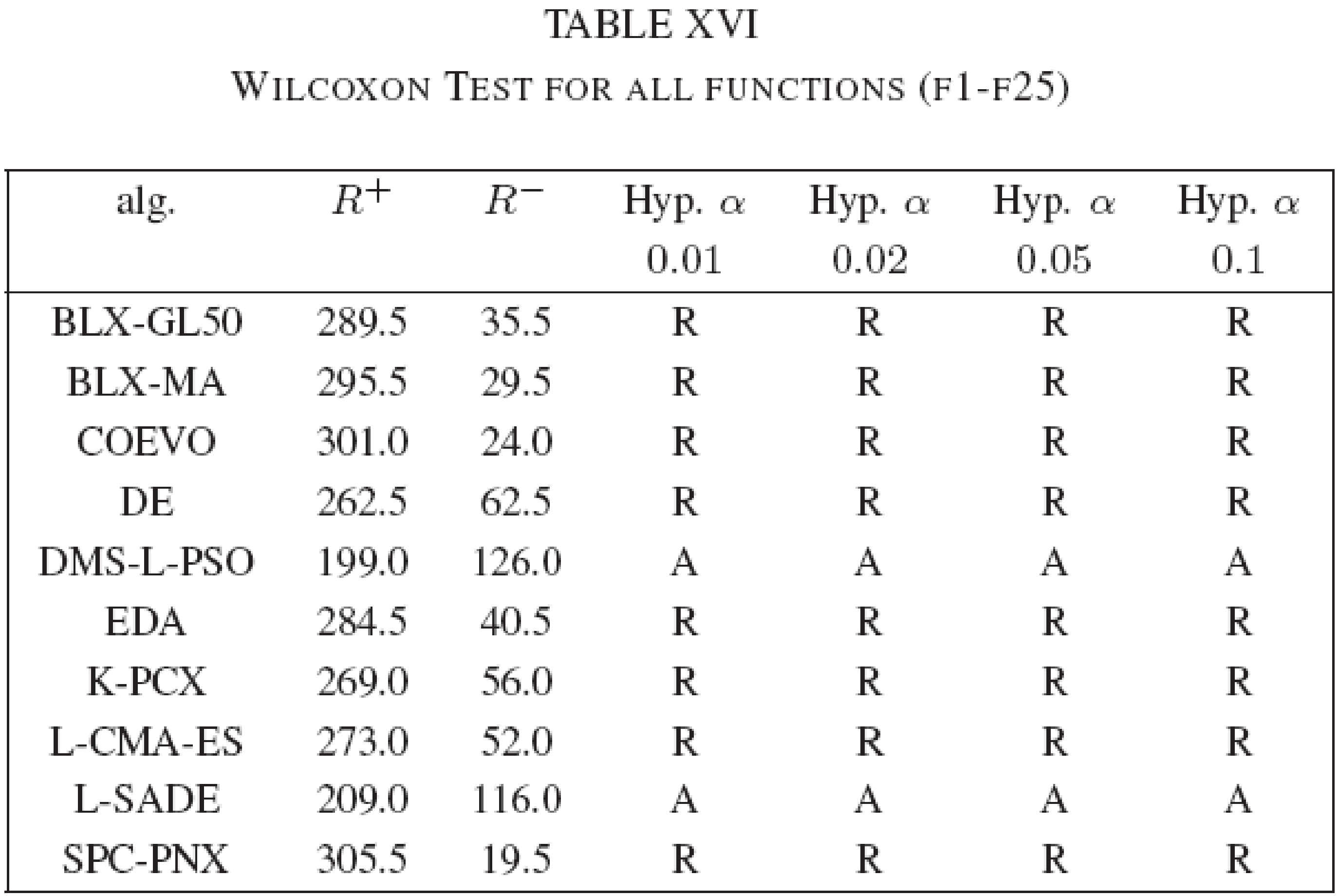


TABLE XVI

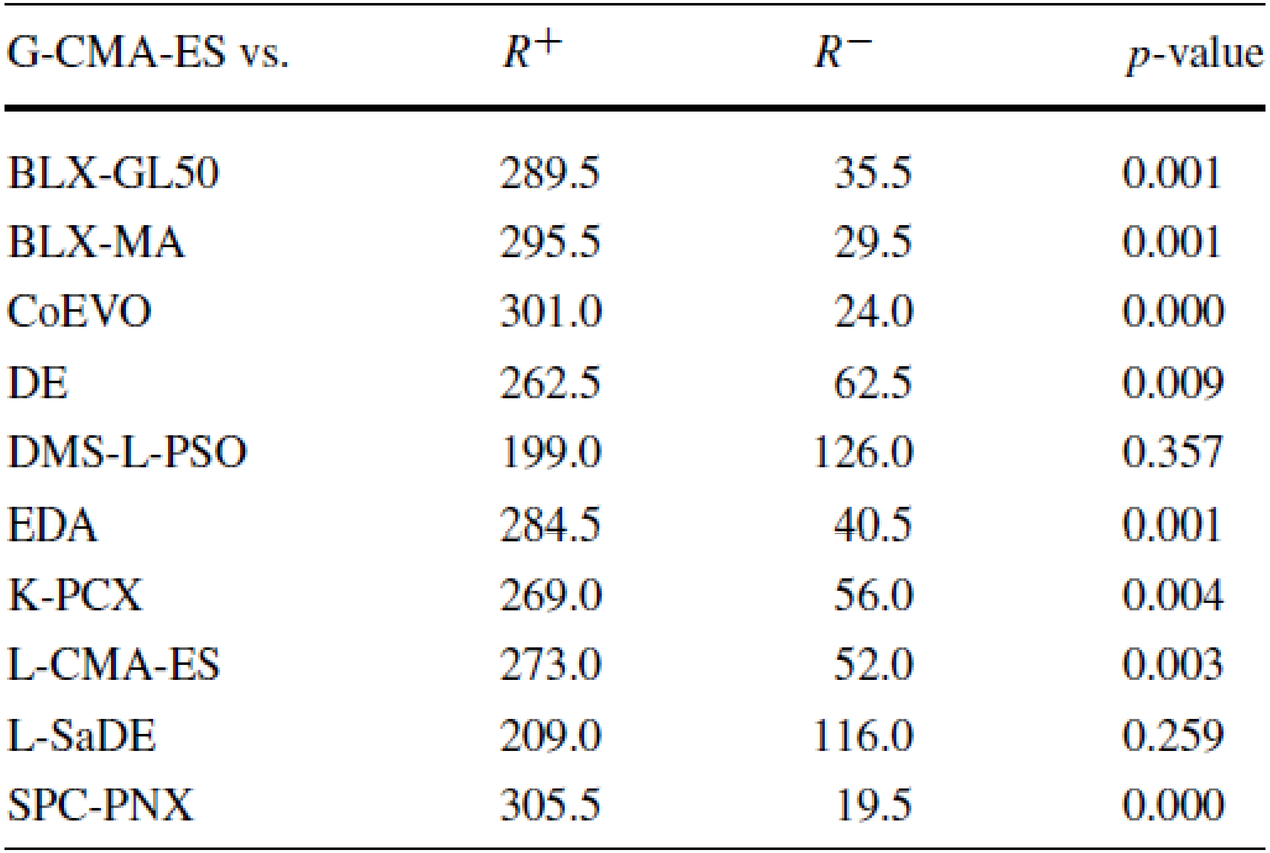
WILCOXON TEST FOR ALL FUNCTIONS (F1-F25)

alg.	R^+	R^-	Hyp. α 0.01	Hyp. α 0.02	Hyp. α 0.05	Hyp. α 0.1
BLX-GL50	289.5	35.5	R	R	R	R
BLX-MA	295.5	29.5	R	R	R	R
COEVO	301.0	24.0	R	R	R	R
DE	262.5	62.5	R	R	R	R
DMS-L-PSO	199.0	126.0	A	A	A	A
EDA	284.5	40.5	R	R	R	R
K-PCX	269.0	56.0	R	R	R	R
L-CMA-ES	273.0	52.0	R	R	R	R
L-SADE	209.0	116.0	A	A	A	A
SPC-PNX	305.5	19.5	R	R	R	R

G-CMAES versus the remaining algorithms.
The critical values are: 68, 76, 89 and 100 (0.01, 0.02, 0.05, 0.1)

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

Example on the use of Wilcoxon's test combined for multiple comparisons

$$
\begin{aligned}
p = P(Reject\ H_0|H_0\ true) = \\
= 1 - P(Accept\ H_0|H_0\ true) = \\
= 1 - P(Accept\ A_k = A_i, i = 1, \ldots, k-1|H_0\ true) = \\
= 1 - \prod_{i=1}^{k-1} P(Accept\_A_k = A_i|H_0\ true) = \\
= 1 - \prod_{i=1}^{k-1} [1 - P(Reject\ A_k = A_i|H_0\ true)] = \\
= 1 - \prod_{i=1}^{k-1} (1 - p_{H_i})
\end{aligned}
$$

$$
\begin{aligned}
p = 1 - (1-0.001)(1-0.001)(1-0.000)(1-0.009)(1-0.357) \\
(1-0.001)(1-0.004)(1-0.003)(1-0.259)(1-0.000) = 0.467
\end{aligned}
$$

Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC'2005 Special Session of Parameter Optimization

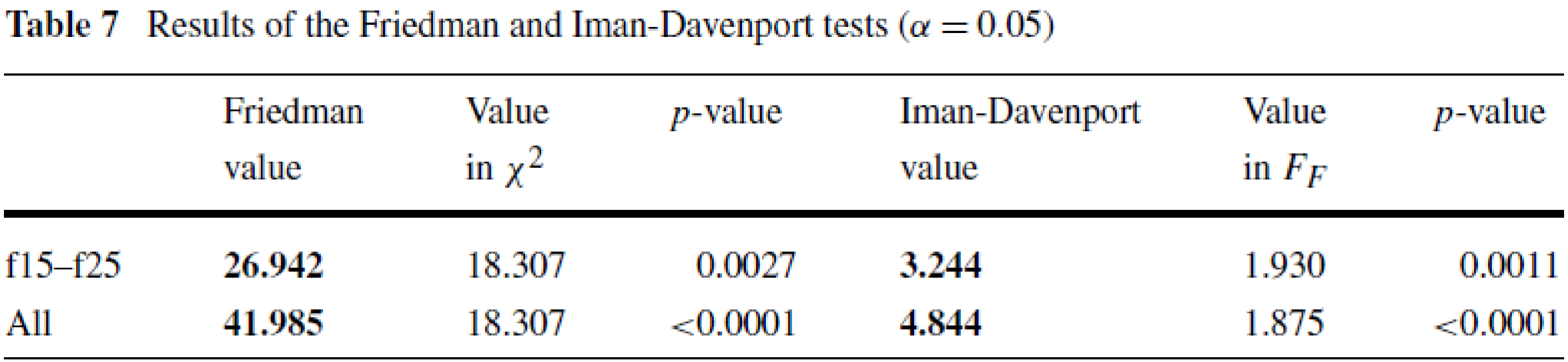
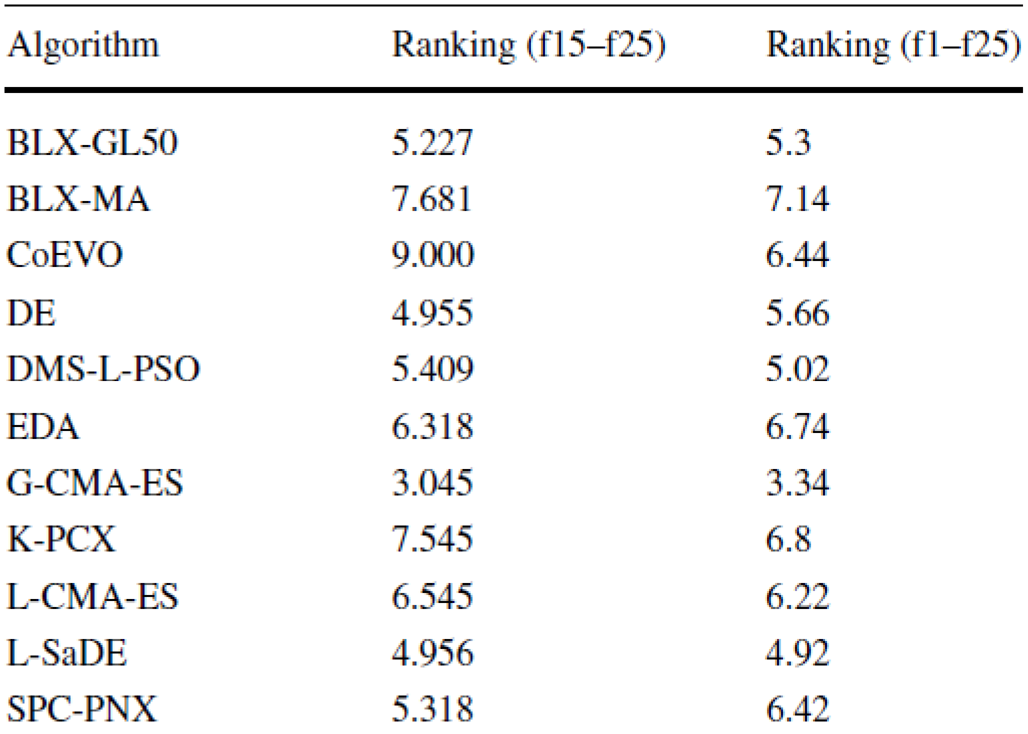


Table 7 Results of the Friedman and Iman-Davenport tests ($\alpha = 0.05$)

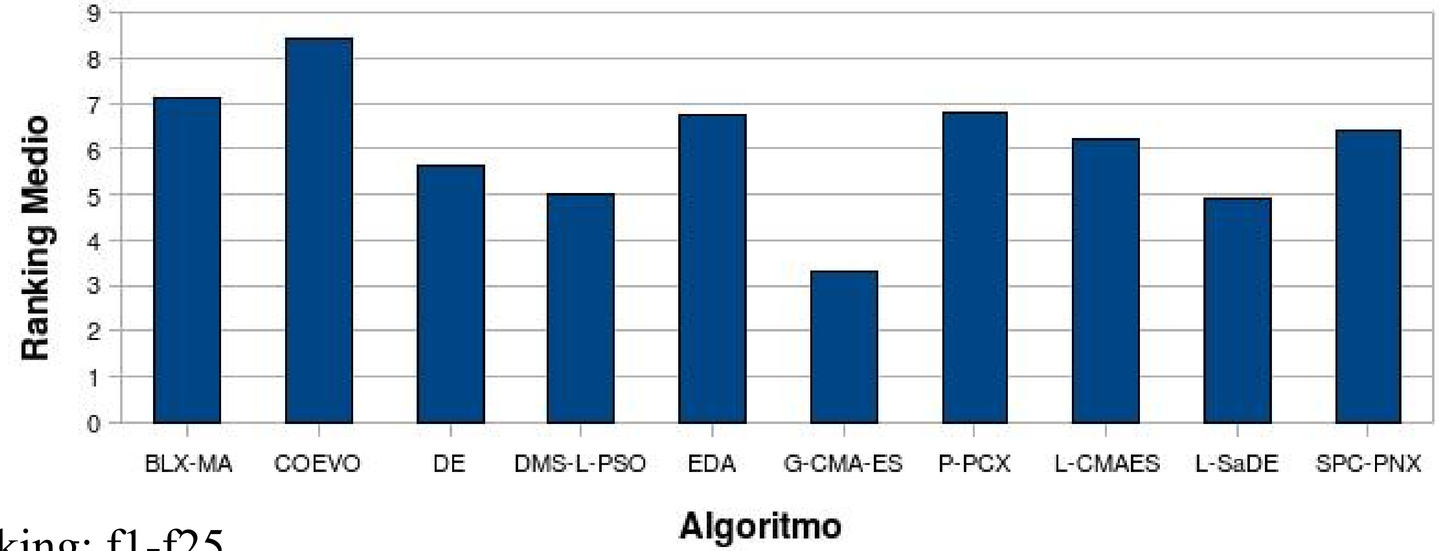
	Friedman value	Value in χ^2	p-value	Iman-Davenport value	Value in F_F	p-value
f15–f25	**26.942**	18.307	0.0027	**3.244**	1.930	0.0011
All	**41.985**	18.307	<0.0001	**4.844**	1.875	<0.0001



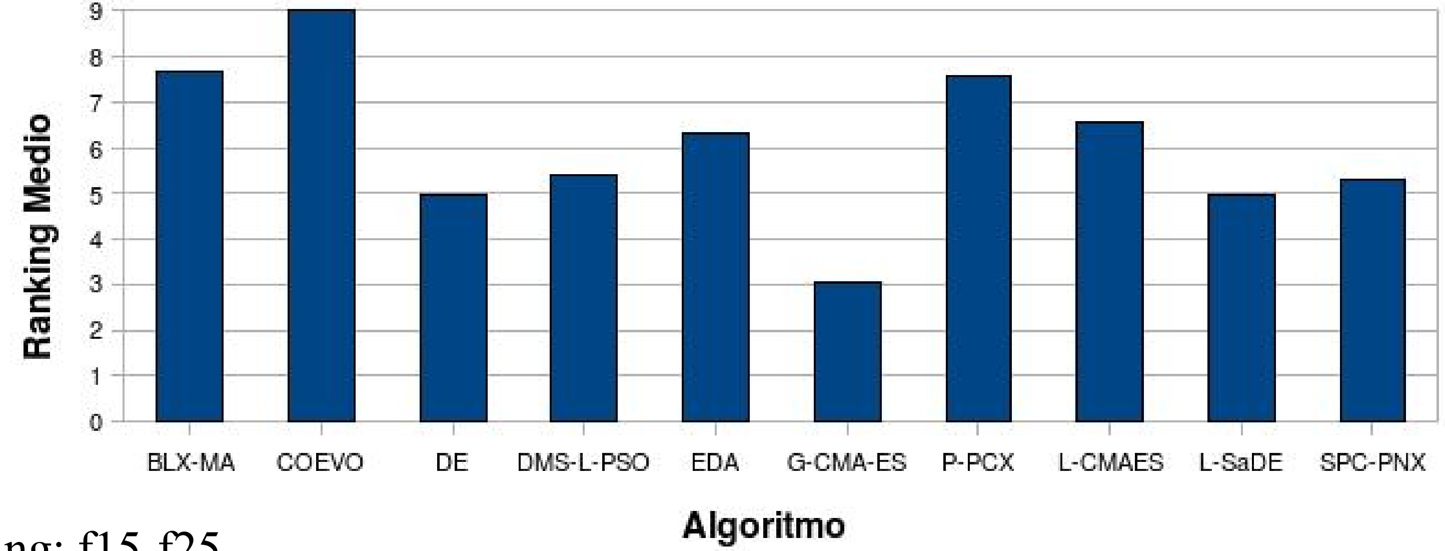
Algorithm	Ranking (f15–f25)	Ranking (f1–f25)
BLX-GL50	5.227	5.3
BLX-MA	7.681	7.14
CoEVO	9.000	6.44
DE	4.955	5.66
DMS-L-PSO	5.409	5.02
EDA	6.318	6.74
G-CMA-ES	3.045	3.34
K-PCX	7.545	6.8
L-CMA-ES	6.545	6.22
L-SaDE	4.956	4.92
SPC-PNX	5.318	6.42

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

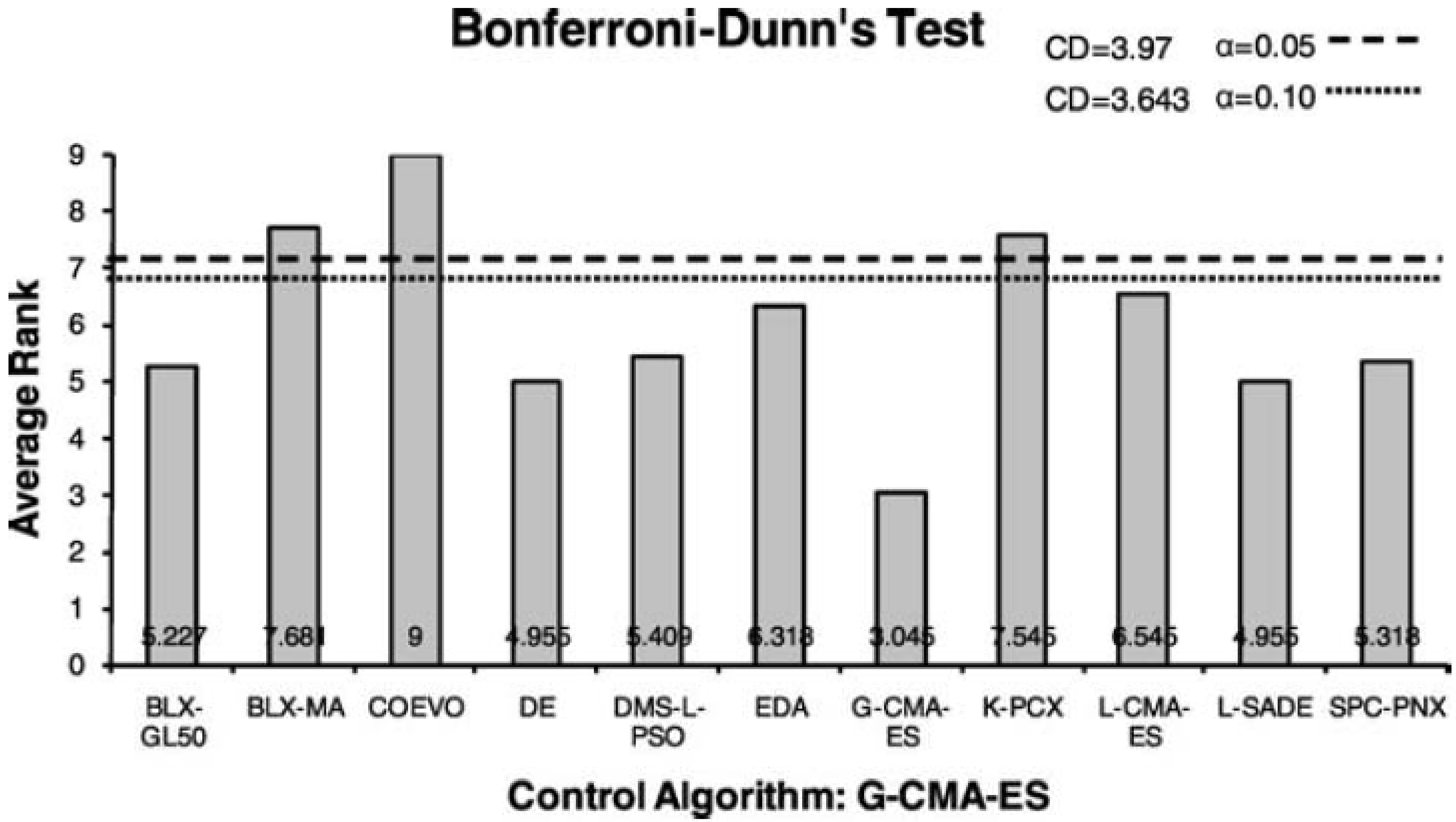
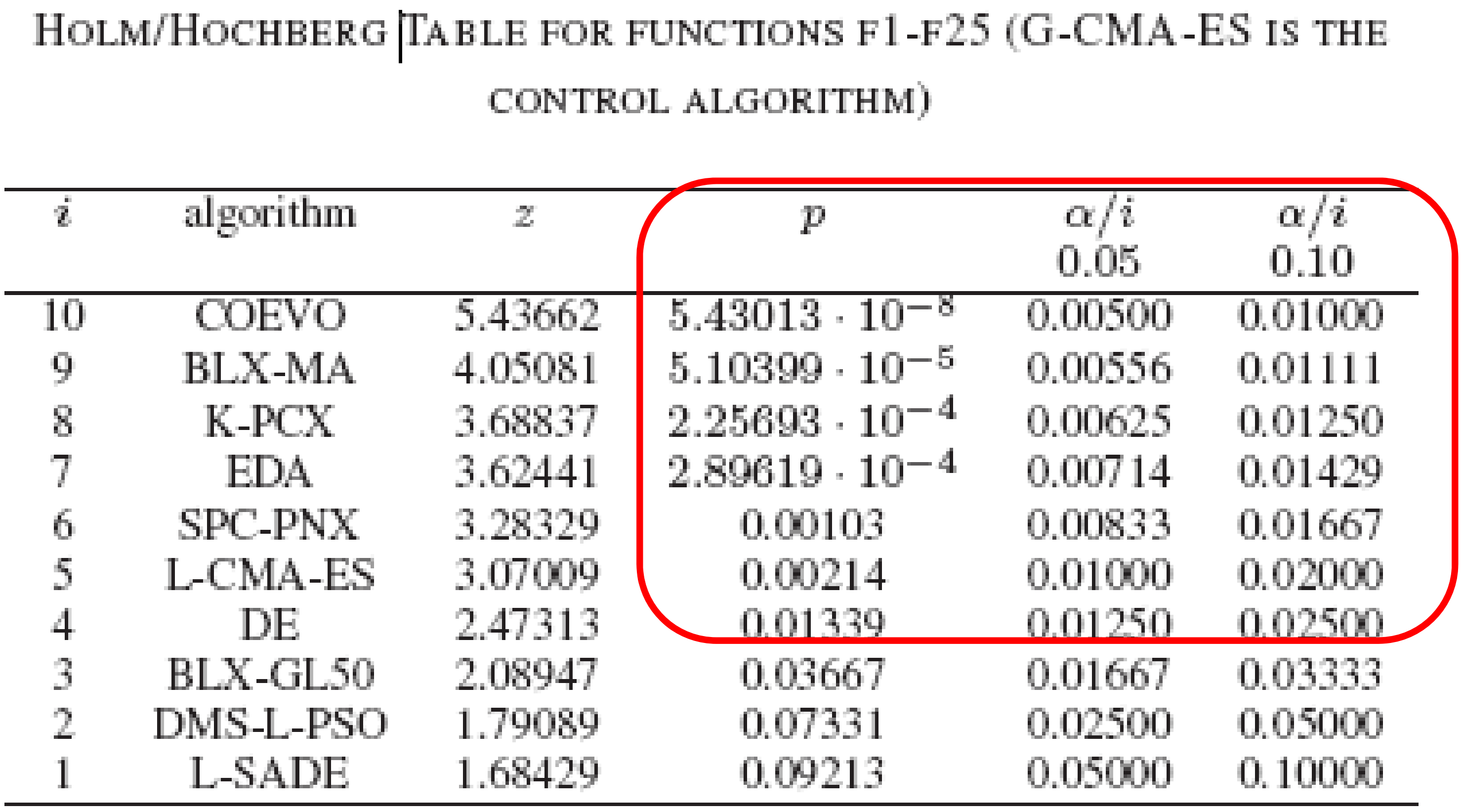


Fig. 6 Bonferroni-Dunn's graphic corresponding to the results for f15–f25

Basic Non-Parametric Tests and Case Studies

Evolutionary Algorithms: CEC'2005 Special Session of Parameter Optimization

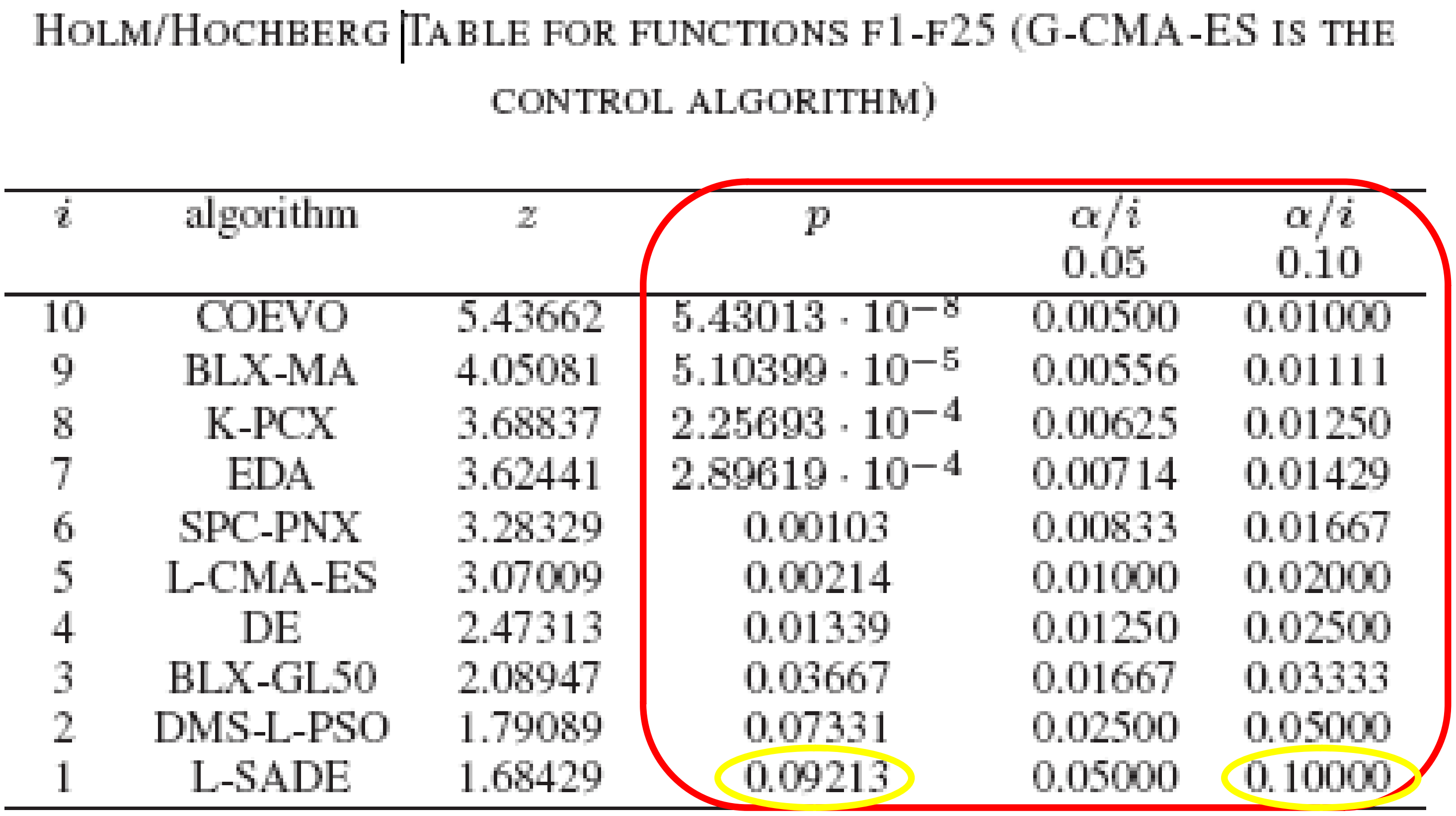


HOLM/HOCHBERG TABLE FOR FUNCTIONS F1-F25 (G-CMA-ES IS THE CONTROL ALGORITHM)

i	algorithm	z	p	α/i 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



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

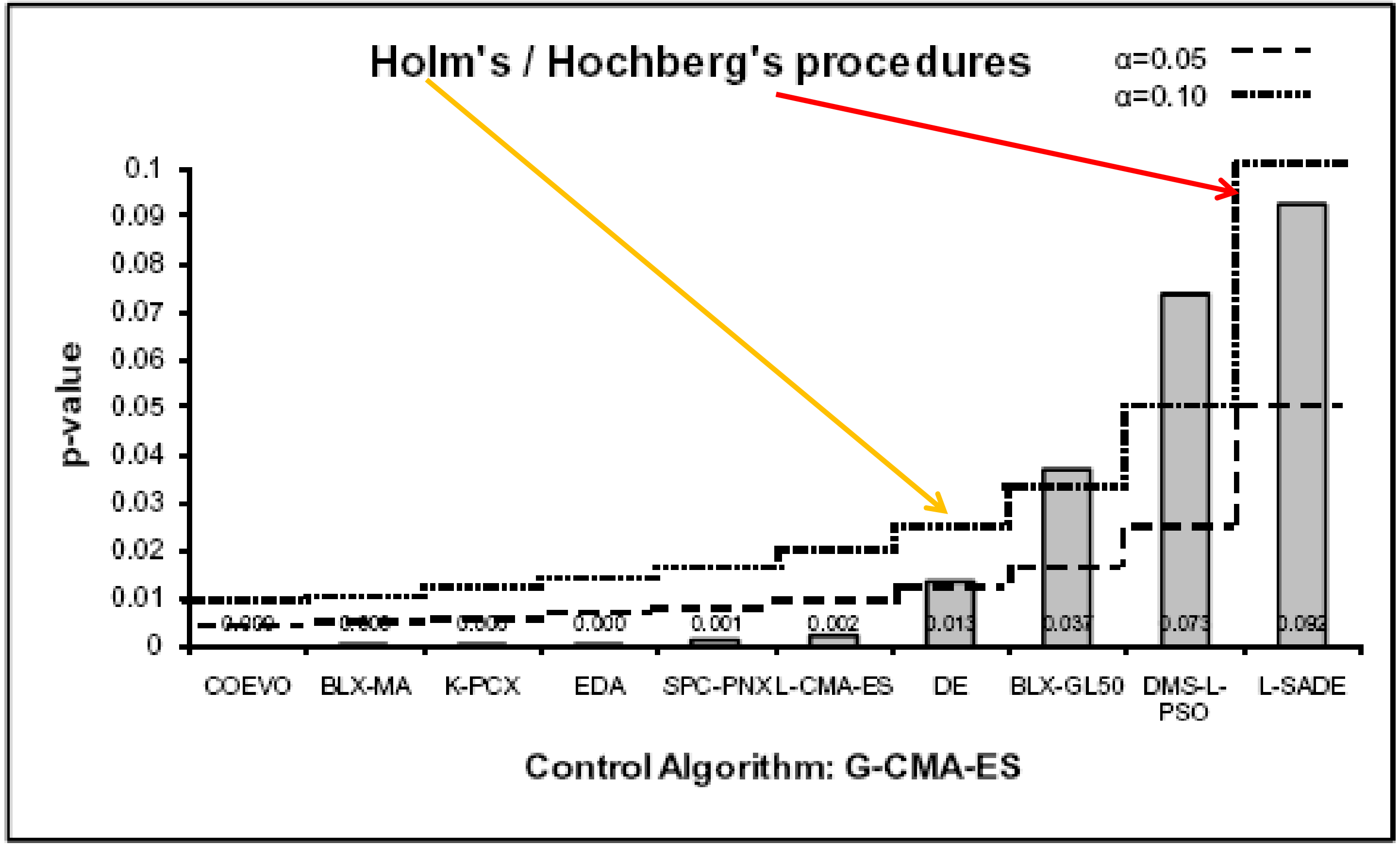


Fig. 11. Holm’s/Hochberg’s procedure for all functions (f1–f25).

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
- Basic non-parametric tests and case studies
- Advanced non-parametric tests and case studies
- Lessons Learned
- Books of Interest and References
- Software

OUTLINE (II)

- **Advanced non-parametric tests and case studies:**
  - For Multiple Comparisons involving control method
  - Post-hoc Procedures
  - Adjusted p-values
  - Detecting all pairwise differences in a multiple comparison
- Lessons Learned
  - Considerations on the use of nonparametric tests
  - Recommendations on the use of nonparametric tests
  - Frequent Questions
- Books of Interest and References
- Software

Advanced non-parametric tests and case studies

- **For Multiple Comparisons involving a control method**
- **Post-hoc Procedures**
- **Adjusted p-values**
- **Detecting all pairwise differences in a multiple comparison**

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

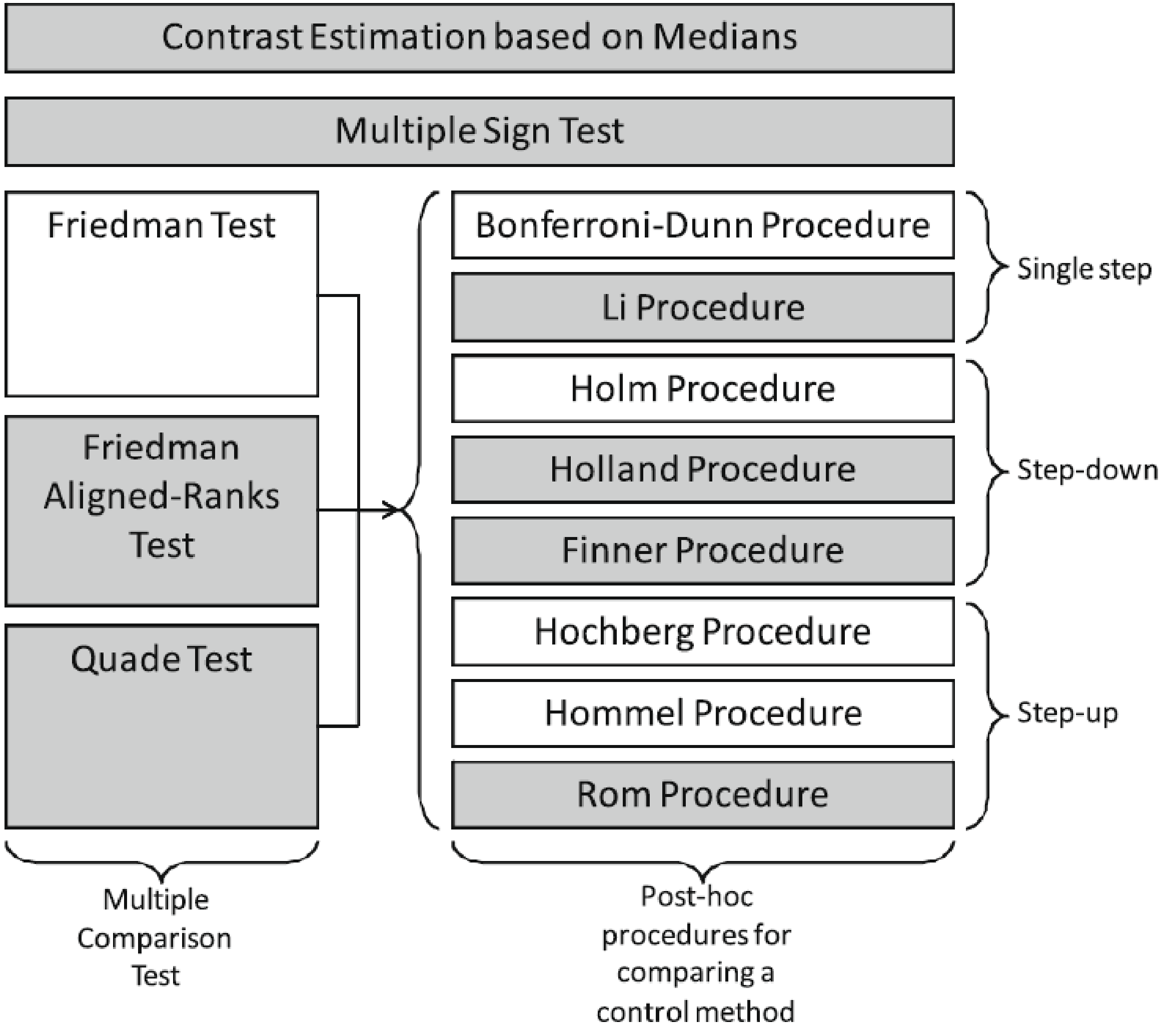
General Case Study used:

- 24 data sets from UCI and KEEL data-set
- Classifiers (from KEEL, standard parameters values):
  - PDFC
  - NNEP
  - IS-CHC + 1NN
  - FH-GBML
- 3 runs of 10fcv

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Multiple Comparison non-parametric procedures map.

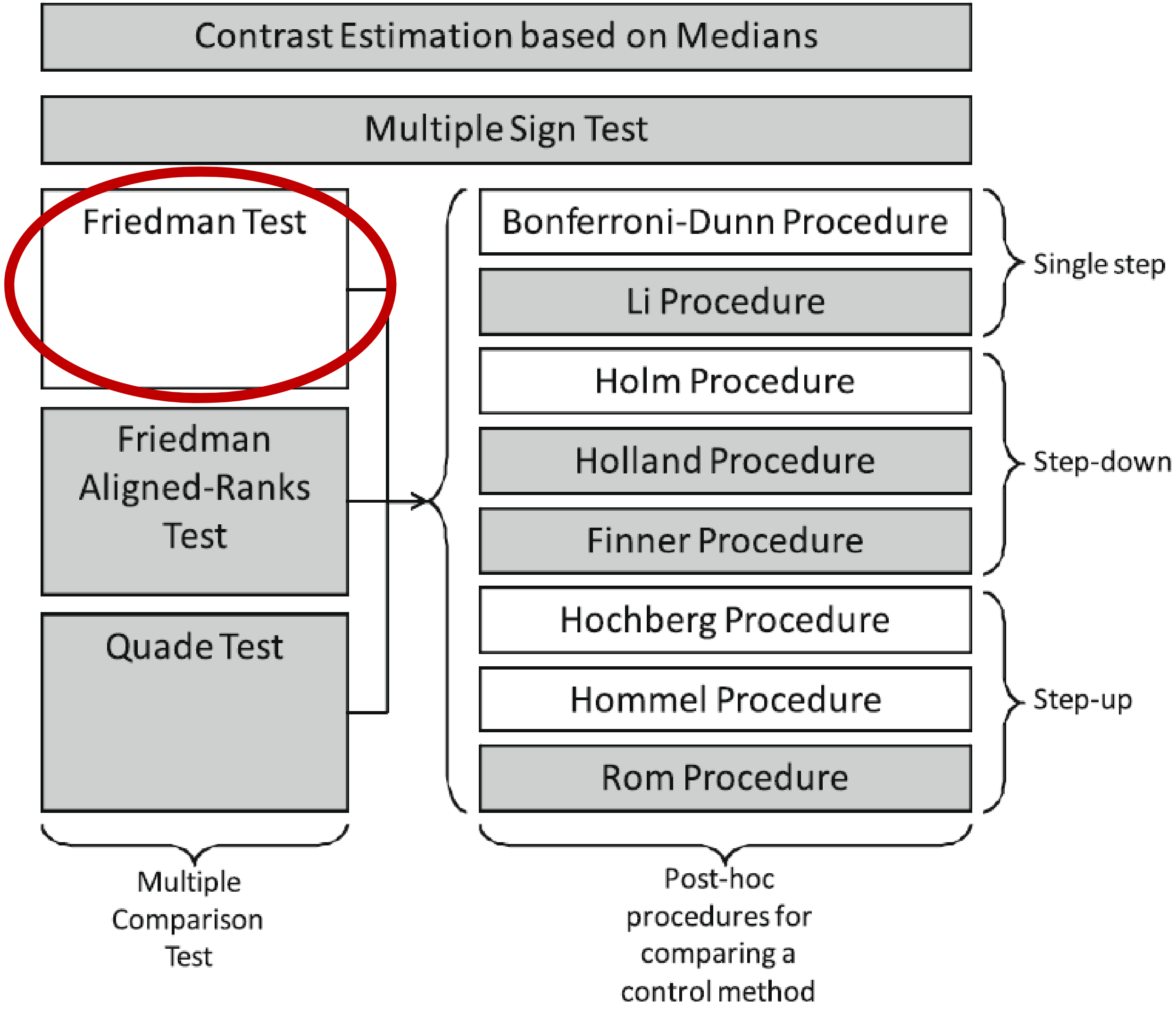


In white are depicted the basic non-parametric test, whereas in grey are depicted more advanced tests which will be presented next.

Source: 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.*

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method



Advanced Non-Parametric Tests and Case Studies

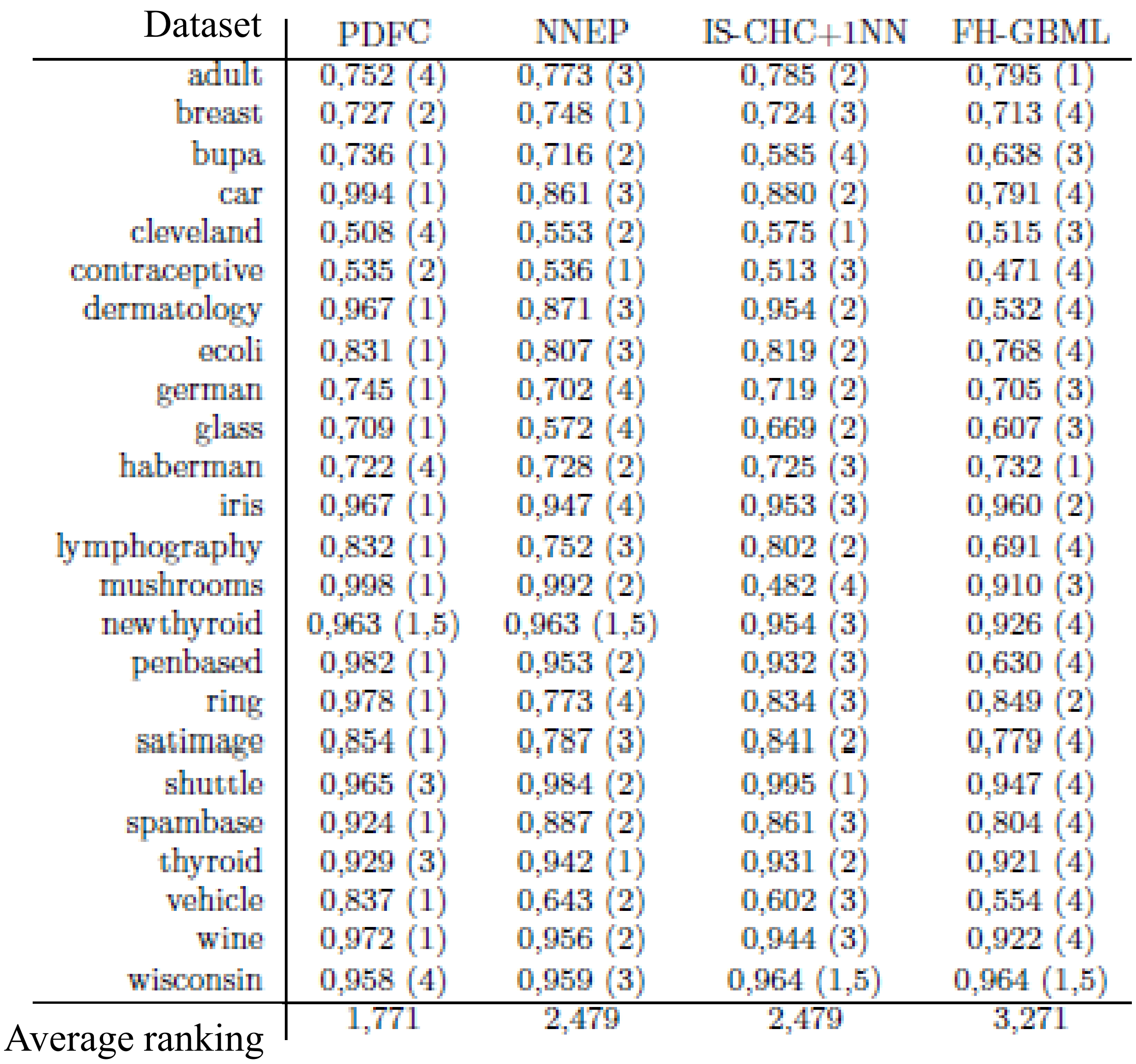
For Multiple Comparisons involving a Control Method

Friedman

and

Iman-Davenport

(only showed for comparison purposes in this case study)



Dataset	PDFC	NNEP	IS-CHC+1NN	FH-GBML
adult	0,752 (4)	0,773 (3)	0,785 (2)	0,795 (1)
breast	0,727 (2)	0,748 (1)	0,724 (3)	0,713 (4)
bupa	0,736 (1)	0,716 (2)	0,585 (4)	0,638 (3)
car	0,994 (1)	0,861 (3)	0,880 (2)	0,791 (4)
cleveland	0,508 (4)	0,553 (2)	0,575 (1)	0,515 (3)
contraceptive	0,535 (2)	0,536 (1)	0,513 (3)	0,471 (4)
dermatology	0,967 (1)	0,871 (3)	0,954 (2)	0,532 (4)
ecoli	0,831 (1)	0,807 (3)	0,819 (2)	0,768 (4)
german	0,745 (1)	0,702 (4)	0,719 (2)	0,705 (3)
glass	0,709 (1)	0,572 (4)	0,669 (2)	0,607 (3)
haberman	0,722 (4)	0,728 (2)	0,725 (3)	0,732 (1)
iris	0,967 (1)	0,947 (4)	0,953 (3)	0,960 (2)
lymphography	0,832 (1)	0,752 (3)	0,802 (2)	0,691 (4)
mushrooms	0,998 (1)	0,992 (2)	0,482 (4)	0,910 (3)
newthyroid	0,963 (1,5)	0,963 (1,5)	0,954 (3)	0,926 (4)
penbased	0,982 (1)	0,953 (2)	0,932 (3)	0,630 (4)
ring	0,978 (1)	0,773 (4)	0,834 (3)	0,849 (2)
satimage	0,854 (1)	0,787 (3)	0,841 (2)	0,779 (4)
shuttle	0,965 (3)	0,984 (2)	0,995 (1)	0,947 (4)
spambase	0,924 (1)	0,887 (2)	0,861 (3)	0,804 (4)
thyroid	0,929 (3)	0,942 (1)	0,931 (2)	0,921 (4)
vehicle	0,837 (1)	0,643 (2)	0,602 (3)	0,554 (4)
wine	0,972 (1)	0,956 (2)	0,944 (3)	0,922 (4)
wisconsin	0,958 (4)	0,959 (3)	0,964 (1,5)	0,964 (1,5)
Average ranking	1,771	2,479	2,479	3,271

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Friedman's measure: 16.255

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

Iman-Davenport's test:

F_F = 6.691, p-value for F(3,3*23) = 0.000497,

Therefore the null hypothesis is rejected.

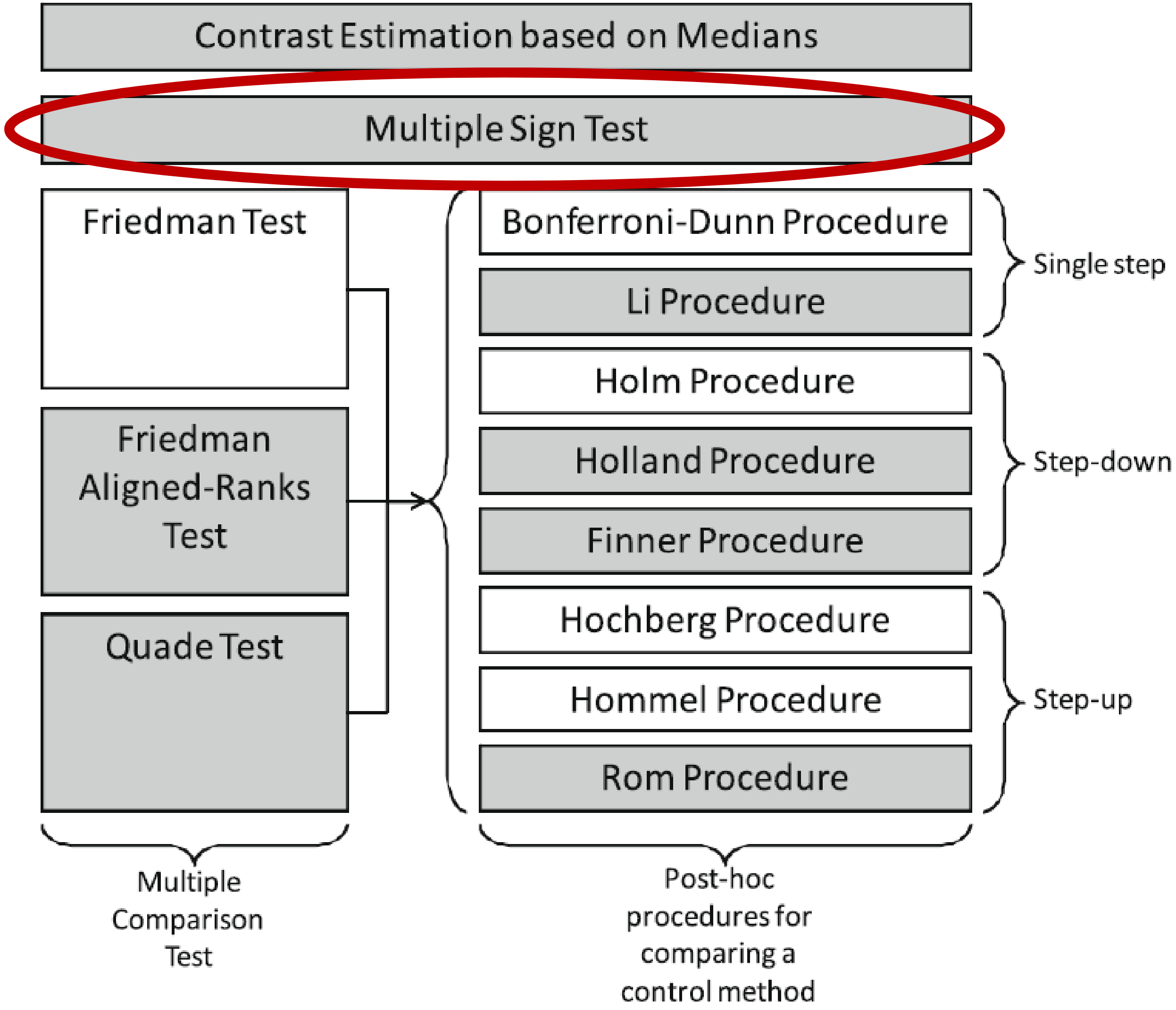
$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

$$\text{(Friedman)} \quad \chi_F^2 = \frac{12 \cdot 24}{4 \cdot 5}\left[(1.771^2 + 2.479^2 + 2.479^2 + 3.271^2) - \frac{4 \cdot 5^2}{4}\right] = 16.225$$

$$\text{(Iman–Davenport)} \quad F_F = \frac{23 \cdot 16.225}{24 \cdot 3 - 16.225} = 6.691$$

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method



Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Multiple sign test: The following procedure, allows us to compare all of the other algorithms with a control labeled algorithm. The technique, an extension of the familiar sign test, carries out the following steps:

1. Represent by x_{i1} and x_{ij} the performances of the control and the jth classifier in the ith data set.
2. Compute the signed differences $d_{ij} = x_{ij} \quad x_{i1}$. In other words, pair each performance with the control and, in each data set, subtract the control performance from the jth classifier.
3. Let r_j equal the number of differences, d_{ij}, that have the less frequently occurring sign (either positive or negative) within a pairing of an algorithm with the control.

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

4. Let M_1 be the median response of a sample of results of the control method and M_j be the median response of a sample of results of the jth algorithm. Apply one of the following decision rules:

   - For testing H_0: $M_j \geq M_1$ against H_1 : $M_j < M_1$, reject H_0 if the number of plus signs is less than or equal to the critical value of R_j appearing in Table A.1 in Appendix A (Ref. below) for k - 1 (number of algorithms excluding control), n and the chosen experimentwise error rate.

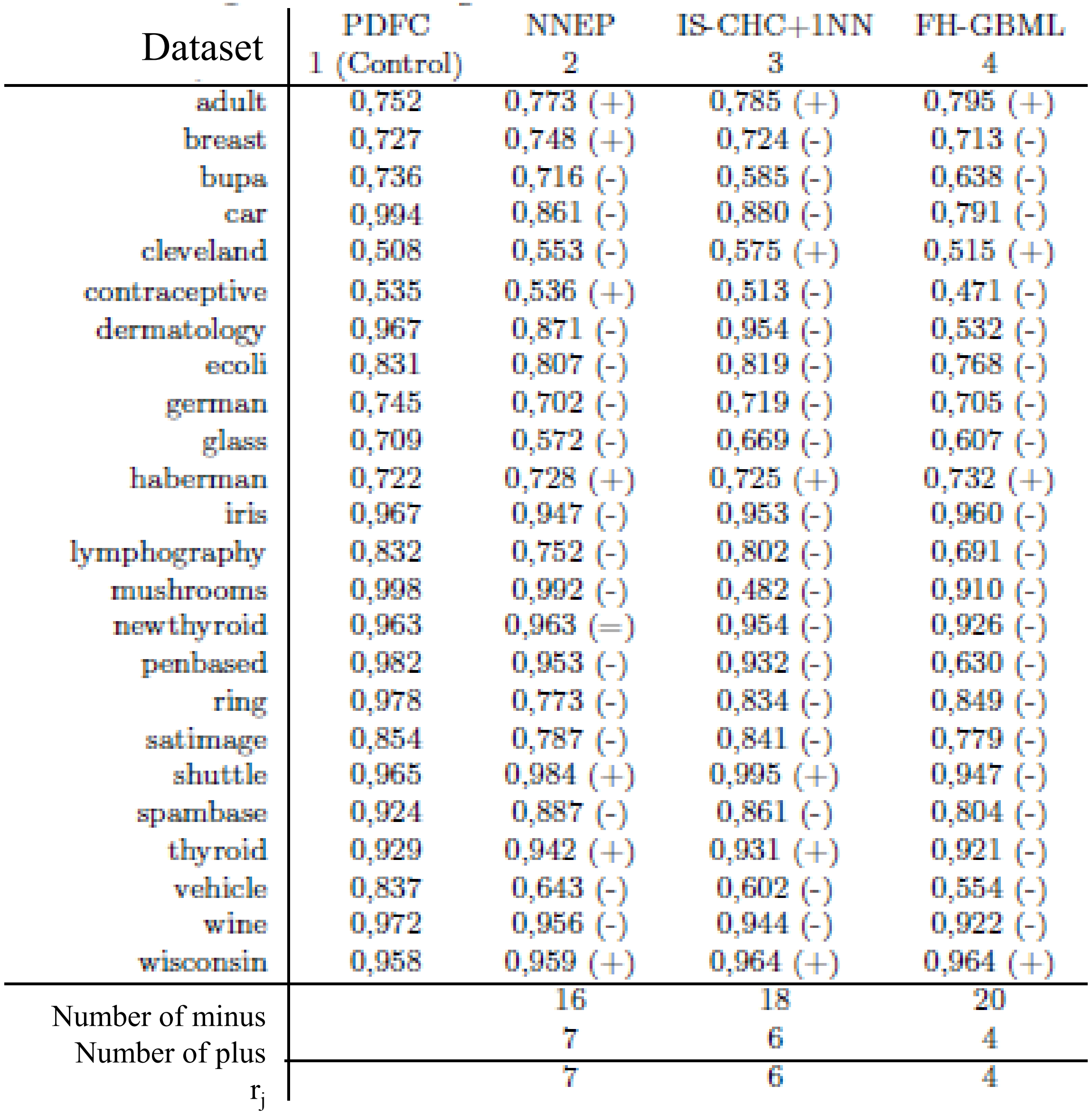
   - For testing H_0: $M_j \leq M_1$ against H_1 : M_j Z M_1, reject H_0 if the number of minus signs is less than or equal to the critical value of R_j appearing in Table A.1 in Appendix A for k - 1, n and the chosen experimentwise error rate.

Source: 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.*

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

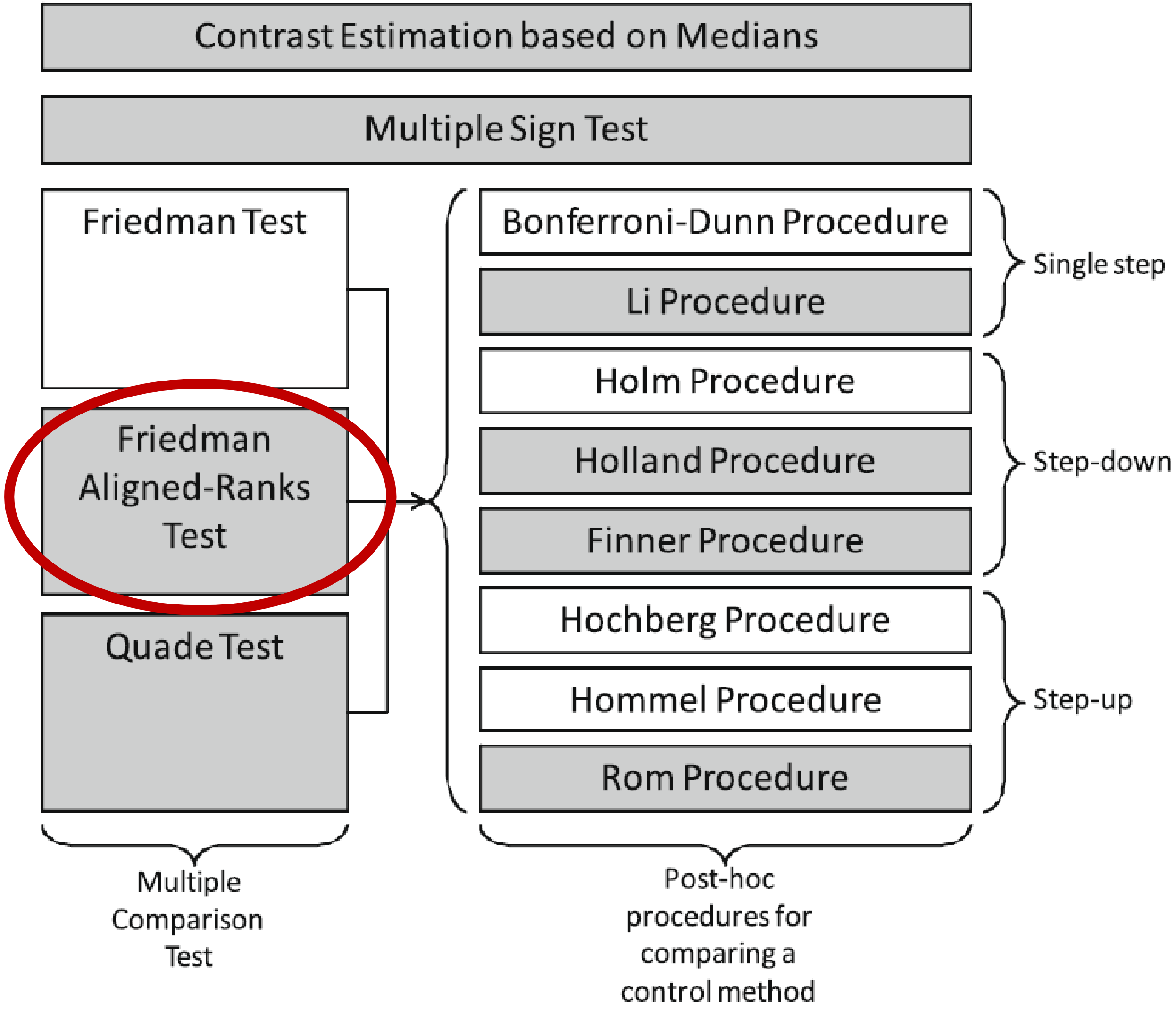
Multiple Sign Test



Dataset	PDFC	NNEP	IS-CHC+1NN	FH-GBML
	1 (Control)	2	3	4
adult	0,752	0,773 (+)	0,785 (+)	0,795 (+)
breast	0,727	0,748 (+)	0,724 (-)	0,713 (-)
bupa	0,736	0,716 (-)	0,585 (-)	0,638 (-)
car	0,994	0,861 (-)	0,880 (-)	0,791 (-)
cleveland	0,508	0,553 (-)	0,575 (+)	0,515 (+)
contraceptive	0,535	0,536 (+)	0,513 (-)	0,471 (-)
dermatology	0,967	0,871 (-)	0,954 (-)	0,532 (-)
ecoli	0,831	0,807 (-)	0,819 (-)	0,768 (-)
german	0,745	0,702 (-)	0,719 (-)	0,705 (-)
glass	0,709	0,572 (-)	0,669 (-)	0,607 (-)
haberman	0,722	0,728 (+)	0,725 (+)	0,732 (+)
iris	0,967	0,947 (-)	0,953 (-)	0,960 (-)
lymphography	0,832	0,752 (-)	0,802 (-)	0,691 (-)
mushrooms	0,998	0,992 (-)	0,482 (-)	0,910 (-)
newthyroid	0,963	0,963 (=)	0,954 (-)	0,926 (-)
penbased	0,982	0,953 (-)	0,932 (-)	0,630 (-)
ring	0,978	0,773 (-)	0,834 (-)	0,849 (-)
satimage	0,854	0,787 (-)	0,841 (-)	0,779 (-)
shuttle	0,965	0,984 (+)	0,995 (+)	0,947 (-)
spambase	0,924	0,887 (-)	0,861 (-)	0,804 (-)
thyroid	0,929	0,942 (+)	0,931 (+)	0,921 (-)
vehicle	0,837	0,643 (-)	0,602 (-)	0,554 (-)
wine	0,972	0,956 (-)	0,944 (-)	0,922 (-)
wisconsin	0,958	0,959 (+)	0,964 (+)	0,964 (+)
Number of minus		16	18	20
Number of plus		7	6	4
r_j		7	6	4

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method



Contrast Estimation based on Medians

Multiple Sign Test

Friedman Test

Friedman Aligned-Ranks Test

Quade Test

Multiple Comparison Test

Bonferroni-Dunn Procedure

Li Procedure

Single step

Holm Procedure

Holland Procedure

Finner Procedure

Step-down

Hochberg Procedure

Hommel Procedure

Rom Procedure

Step-up

Post-hoc procedures for comparing a control method

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

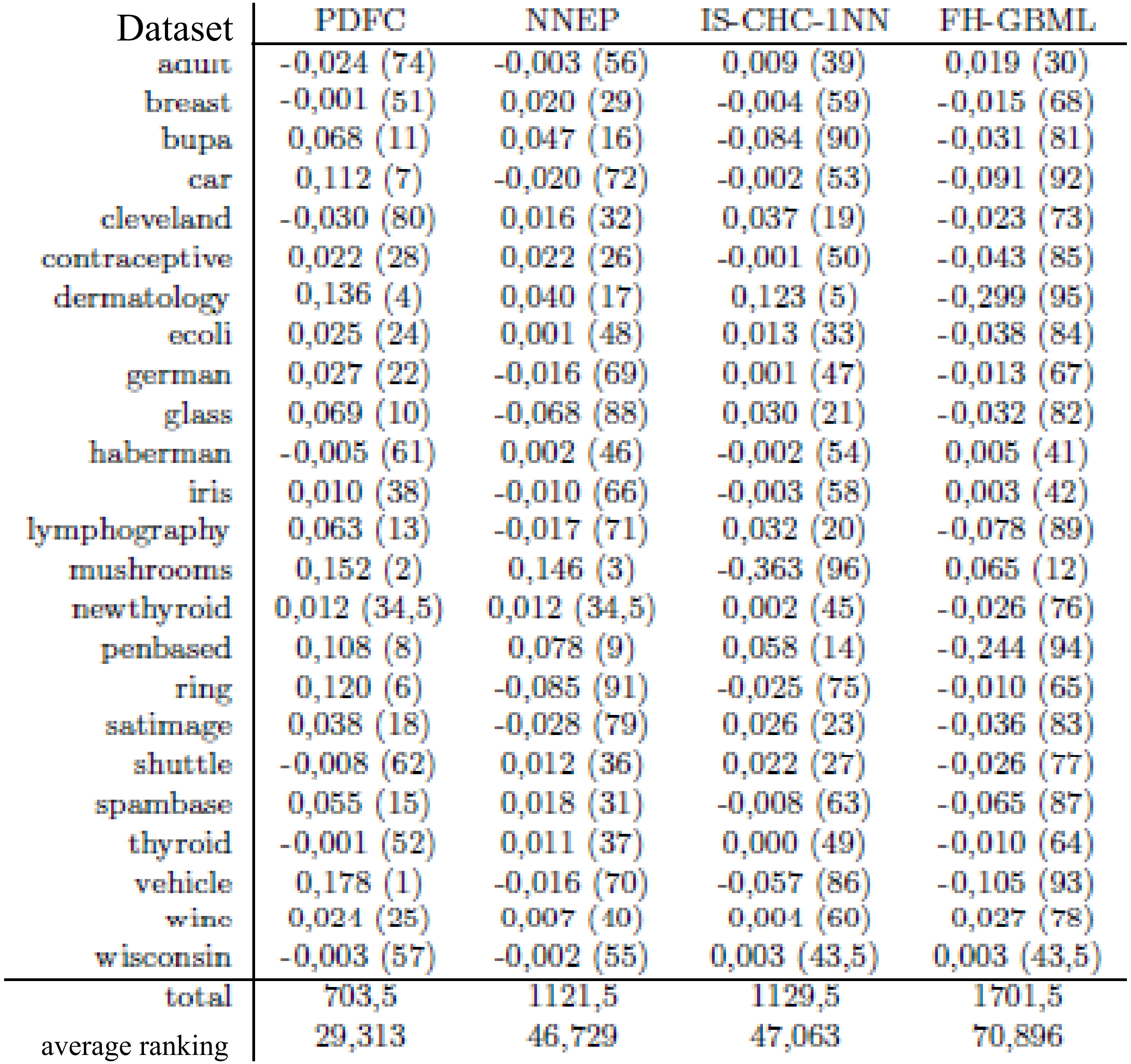
Aligned Ranks Friedman's test: a value of location is computed as the average performance achieved by all algorithms in each data set. Then, it calculates the difference between the performance obtained by an algorithm and the value of location. This step is repeated for algorithms and data sets. The resulting differences, called aligned observations, which keep their identities with respect to the data set and the combination of algorithms to which they belong, are then ranked from 1 to kn relative to each other. Then, the ranking scheme is the same as that employed by a multiple comparison procedure which employs independent samples; such as the Kruskal–Wallis test. The ranks assigned to the aligned observations are called aligned ranks.

$$T = \frac{(k-1)\left[\sum_{j=1}^{k} \widehat{R}_{.j}^{2} - (kn^{2}/4)(kn+1)^{2}\right]}{\{[kn(kn+1)(2kn+1)]/6\} - (1/k)\sum_{i=1}^{n} \widehat{R}_{i.}^{2}}$$

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Friedman Aligned Ranks



Dataset	PDFC	NNEP	IS-CHC-1NN	FH-GBML
adult	-0,024 (74)	-0,003 (56)	0,009 (39)	0,019 (30)
breast	-0,001 (51)	0,020 (29)	-0,004 (59)	-0,015 (68)
bupa	0,068 (11)	0,047 (16)	-0,084 (90)	-0,031 (81)
car	0,112 (7)	-0,020 (72)	-0,002 (53)	-0,091 (92)
cleveland	-0,030 (80)	0,016 (32)	0,037 (19)	-0,023 (73)
contraceptive	0,022 (28)	0,022 (26)	-0,001 (50)	-0,043 (85)
dermatology	0,136 (4)	0,040 (17)	0,123 (5)	-0,299 (95)
ecoli	0,025 (24)	0,001 (48)	0,013 (33)	-0,038 (84)
german	0,027 (22)	-0,016 (69)	0,001 (47)	-0,013 (67)
glass	0,069 (10)	-0,068 (88)	0,030 (21)	-0,032 (82)
haberman	-0,005 (61)	0,002 (46)	-0,002 (54)	0,005 (41)
iris	0,010 (38)	-0,010 (66)	-0,003 (58)	0,003 (42)
lymphography	0,063 (13)	-0,017 (71)	0,032 (20)	-0,078 (89)
mushrooms	0,152 (2)	0,146 (3)	-0,363 (96)	0,065 (12)
newthyroid	0,012 (34,5)	0,012 (34,5)	0,002 (45)	-0,026 (76)
penbased	0,108 (8)	0,078 (9)	0,058 (14)	-0,244 (94)
ring	0,120 (6)	-0,085 (91)	-0,025 (75)	-0,010 (65)
satimage	0,038 (18)	-0,028 (79)	0,026 (23)	-0,036 (83)
shuttle	-0,008 (62)	0,012 (36)	0,022 (27)	-0,026 (77)
spambase	0,055 (15)	0,018 (31)	-0,008 (63)	-0,065 (87)
thyroid	-0,001 (52)	0,011 (37)	0,000 (49)	-0,010 (64)
vehicle	0,178 (1)	-0,016 (70)	-0,057 (86)	-0,105 (93)
wine	0,024 (25)	0,007 (40)	0,004 (60)	0,027 (78)
wisconsin	-0,003 (57)	-0,002 (55)	0,003 (43,5)	0,003 (43,5)
total	703,5	1121,5	1129,5	1701,5
average ranking	29,313	46,729	47,063	70,896

Aligned Ranks Friedman's measure: 18.837

$$T = \frac{(k-1)\left[\sum_{j=1}^{k} \widehat{R}_{.j}^{2} - (kn^{2}/4)(kn+1)^{2}\right]}{\{[kn(kn+1)(2kn+1)]/6\} - (1/k)\sum_{i=1}^{n} \widehat{R}_{i.}^{2}}$$

The p-value of Chi^2 with 3 degrees of freedom is 0.000296. Hypothesis rejected

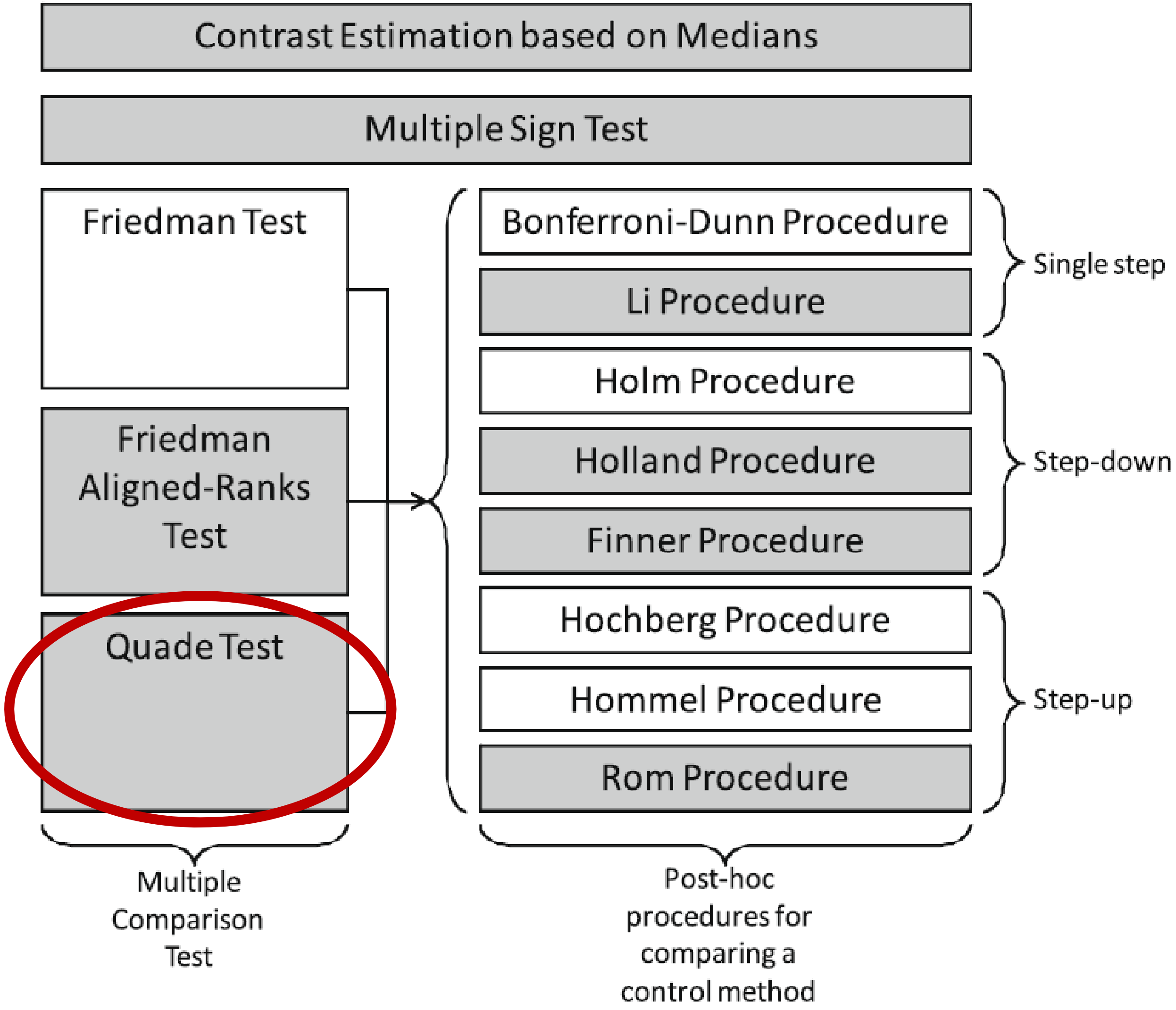
$$\sum_{j=1}^{k} \widehat{R}_{.j}^{2} = 703.5^{2} + 1121^{2} + 1129.5^{2} + 1701.5^{2} = 5,923,547$$

$$\sum_{i=1}^{n} \widehat{R}_{i.}^{2} = 199^{2} + 207^{2} + 198^{2} + \cdots + 199^{2} = 926,830$$

$$T = \frac{(4-1)[5,923,547 - (4 \cdot 24^{2}/4)(4 \cdot 24 + 1)^{2}]}{\{[4 \cdot 24(4 \cdot 24 + 1)(2 \cdot 4 \cdot 24 + 1)]/6\} - (1/4) \cdot 926,830} = 18.837$$

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method



Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Quade test: The Friedman test considers all data sets to be equal in terms of importance. An alternative to this could take into account the fact that some data sets are more difficult or the differences registered on the run of various algorithms over them are larger. The rankings computed on each data set could be scaled depending on the differences observed in the algorithms'performances.

The procedure starts finding the ranks in the same way as the Friedman test does. The next step requires the original values of performance of the classifiers. Ranks are assigned to the data sets themselves according to the size of the sample

range in each data set. The sample range within data set i is the difference between the largest and the smallest observations within that data set:

$$\text{Range in data set}: \; i = \max_j\{x_{ij}\} - \min_j\{x_{ij}\}$$

$$S_{ij} = Q_i\left[r_i^j - \frac{k+1}{2}\right]$$

$$A_2 = n(n+1)(2n+1)(k)(k+1)(k-1)/72$$

$$B = \frac{1}{n}\sum_{j=1}^{k} S_j^2$$

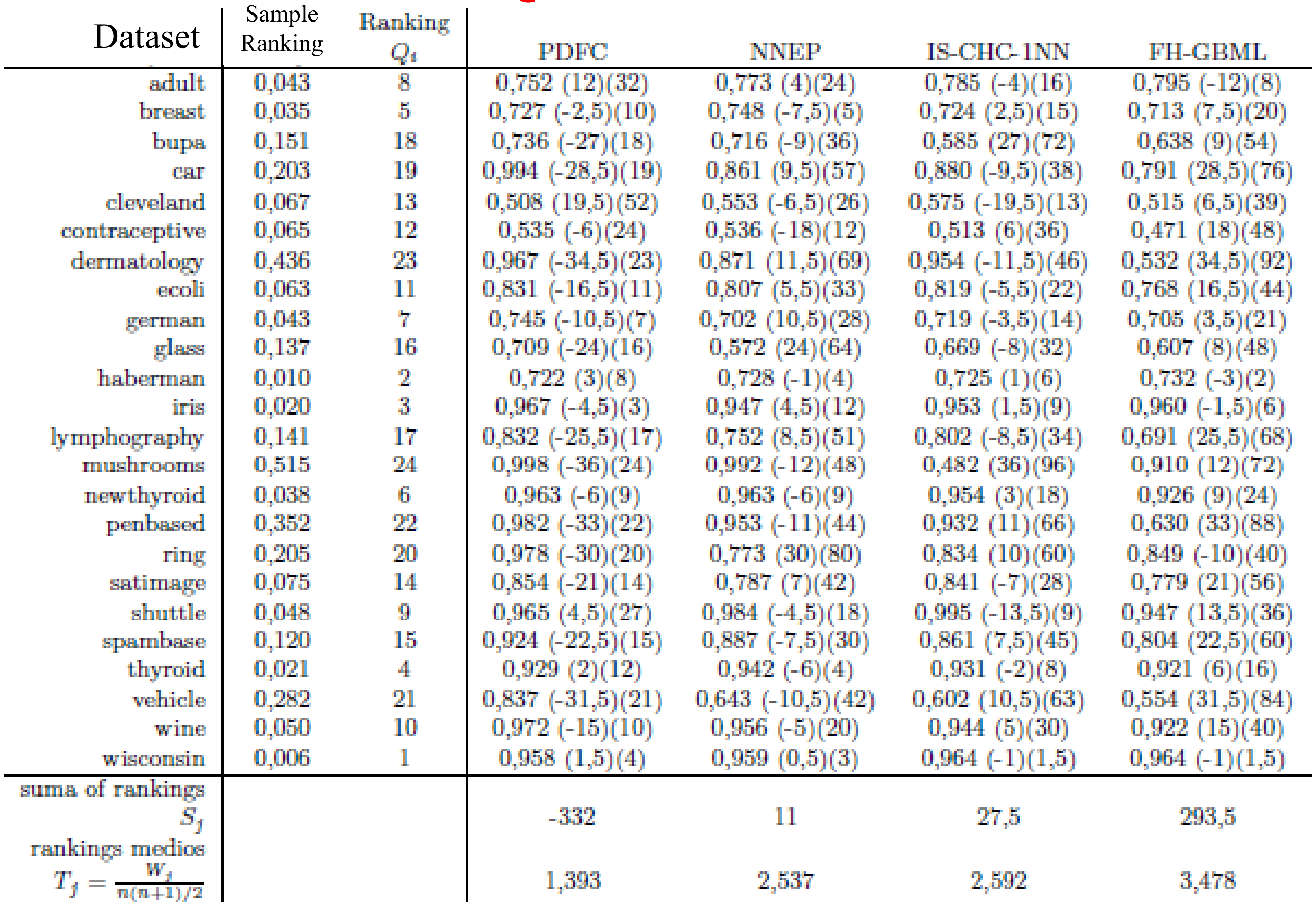
The test statistic is

$$T_3 = \frac{(n-1)B}{A_2 - B}$$

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Quade



Dataset	Sample Ranking	Ranking Q_i	PDFC	NNEP	IS-CHC-1NN	FH-GBML
adult	0,043	8	0,752 (12)(32)	0,773 (4)(24)	0,785 (-4)(16)	0,795 (-12)(8)
breast	0,035	5	0,727 (-2,5)(10)	0,748 (-7,5)(5)	0,724 (2,5)(15)	0,713 (7,5)(20)
bupa	0,151	18	0,736 (-27)(18)	0,716 (-9)(36)	0,585 (27)(72)	0,638 (9)(54)
car	0,203	19	0,994 (-28,5)(19)	0,861 (9,5)(57)	0,880 (-9,5)(38)	0,791 (28,5)(76)
cleveland	0,067	13	0,508 (19,5)(52)	0,553 (-6,5)(26)	0,575 (-19,5)(13)	0,515 (6,5)(39)
contraceptive	0,065	12	0,535 (-6)(24)	0,536 (-18)(12)	0,513 (6)(36)	0,471 (18)(48)
dermatology	0,436	23	0,967 (-34,5)(23)	0,871 (11,5)(69)	0,954 (-11,5)(46)	0,532 (34,5)(92)
ecoli	0,063	11	0,831 (-16,5)(11)	0,807 (5,5)(33)	0,819 (-5,5)(22)	0,768 (16,5)(44)
german	0,043	7	0,745 (-10,5)(7)	0,702 (10,5)(28)	0,719 (-3,5)(14)	0,705 (3,5)(21)
glass	0,137	16	0,709 (-24)(16)	0,572 (24)(64)	0,669 (-8)(32)	0,607 (8)(48)
haberman	0,010	2	0,722 (3)(8)	0,728 (-1)(4)	0,725 (1)(6)	0,732 (-3)(2)
iris	0,020	3	0,967 (-4,5)(3)	0,947 (4,5)(12)	0,953 (1,5)(9)	0,960 (-1,5)(6)
lymphography	0,141	17	0,832 (-25,5)(17)	0,752 (8,5)(51)	0,802 (-8,5)(34)	0,691 (25,5)(68)
mushrooms	0,515	24	0,998 (-36)(24)	0,992 (-12)(48)	0,482 (36)(96)	0,910 (12)(72)
newthyroid	0,038	6	0,963 (-6)(9)	0,963 (-6)(9)	0,954 (3)(18)	0,926 (9)(24)
penbased	0,352	22	0,982 (-33)(22)	0,953 (-11)(44)	0,932 (11)(66)	0,630 (33)(88)
ring	0,205	20	0,978 (-30)(20)	0,773 (30)(80)	0,834 (10)(60)	0,849 (-10)(40)
satimage	0,075	14	0,854 (-21)(14)	0,787 (7)(42)	0,841 (-7)(28)	0,779 (21)(56)
shuttle	0,048	9	0,965 (4,5)(27)	0,984 (-4,5)(18)	0,995 (-13,5)(9)	0,947 (13,5)(36)
spambase	0,120	15	0,924 (-22,5)(15)	0,887 (-7,5)(30)	0,861 (7,5)(45)	0,804 (22,5)(60)
thyroid	0,021	4	0,929 (2)(12)	0,942 (-6)(4)	0,931 (-2)(8)	0,921 (6)(16)
vehicle	0,282	21	0,837 (-31,5)(21)	0,643 (-10,5)(42)	0,602 (10,5)(63)	0,554 (31,5)(84)
wine	0,050	10	0,972 (-15)(10)	0,956 (-5)(20)	0,944 (5)(30)	0,922 (15)(40)
wisconsin	0,006	1	0,958 (1,5)(4)	0,959 (0,5)(3)	0,964 (-1)(1,5)	0,964 (-1)(1,5)
suma of rankings S_j			-332	11	27,5	293,5
rankings medios $T_j = \frac{W_j}{n(n+1)/2}$			1,393	2,537	2,592	3,478

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Quade measure: 21.967

With four algorithms and 24 data sets, T3 is distributed according to the F distribution with 4-1=3 and (4-1)*(24-1)=69 degrees of freedom. The p-value computed by using the F distribution is 0.000000000429, so the null hypothesis is rejected at a high level of significance.

$$A_2 = n(n+1)(2n+1)(k)(k+1)(k-1)/72$$

$$B = \frac{1}{n}\sum_{j=1}^{k} S_j{}^2$$

$$T_3 = \frac{(n-1)B}{A_2 - B}$$

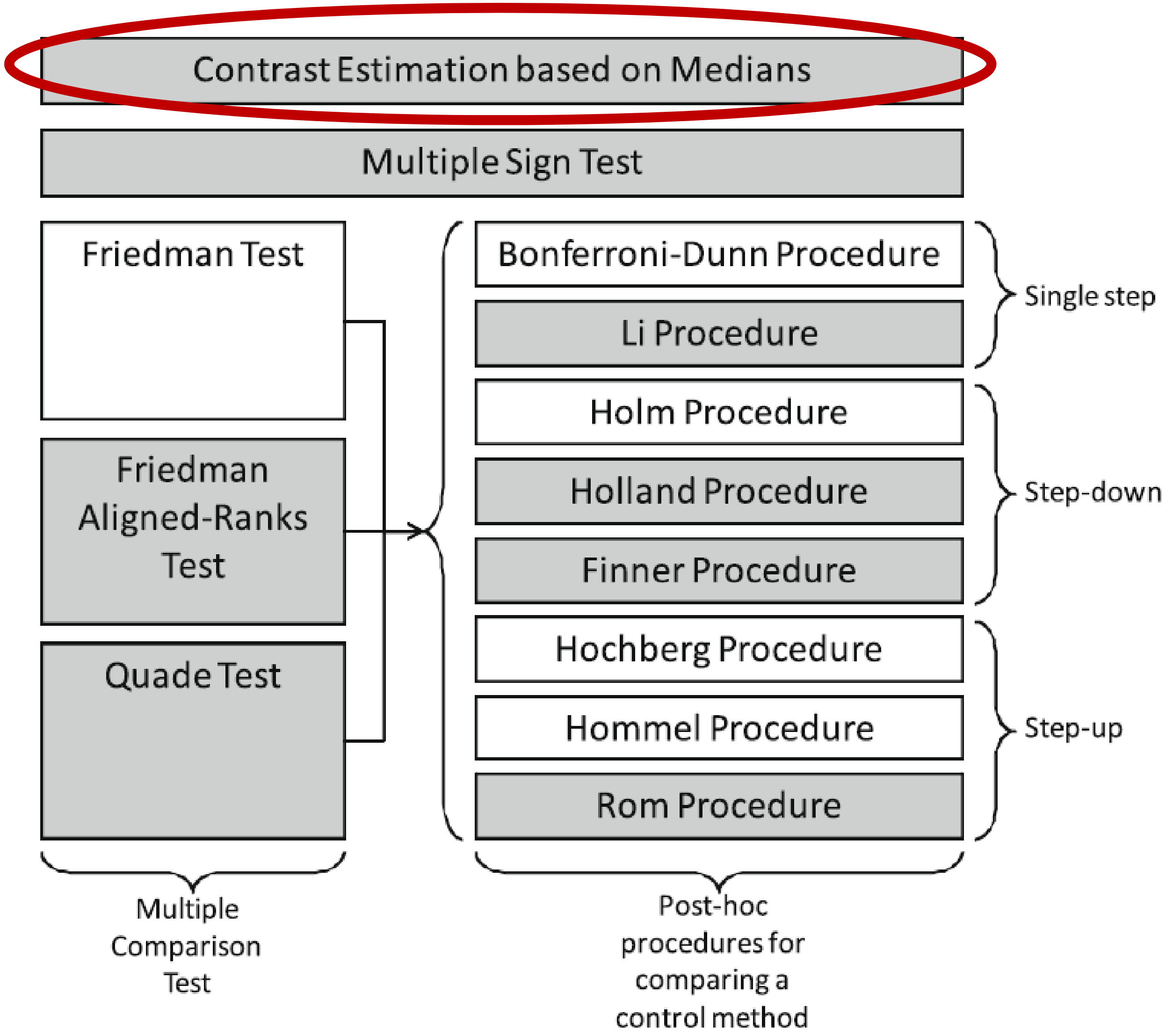
$$A_2 = 24(24+1)(2\cdot 24+1)4(4+1)(4-1)/72 = 24,500$$

$$B = \frac{1}{24}[(-332)^2 + 11^2 + 27.5^2 + 293.5^2] = 4068.479$$

$$T_3 = \frac{23\cdot 4068.479}{24,500 - 4068.479} = 21.967$$

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method



Contrast Estimation based on Medians

Multiple Sign Test

Friedman Test

Friedman Aligned-Ranks Test

Quade Test

Bonferroni-Dunn Procedure

Li Procedure

Holm Procedure

Holland Procedure

Finner Procedure

Hochberg Procedure

Hommel Procedure

Rom Procedure

Single step

Step-down

Step-up

Multiple Comparison Test

Post-hoc procedures for comparing a control method

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Contrast Estimation based on medians: Using the data resulting from the run of various classifiers over multiple data sets in an experiment, the researcher could be interested in the estimation of the difference between two classifiers' performance.

A procedure for this purpose assumes that the expected differences between performances of algorithms are the same across data sets. We assume that the performance is reflected by the magnitudes of the differences between the performances of the algorithms.

Consequently, we are interested in estimating the contrast between medians of samples of results considering all pairwise comparisons. It obtains a quantitative difference computed through medians between two algorithms over multiple data sets, but the value obtained will change when using other data sets in the experiment.

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Contrast Estimation Based on Means procedure:

1. Compute the difference between every pair of k algorithms in each of the n data set: $D_{i(uv)} = X_{iu} – X_{iv}$, only when u < v.

2. Compute the median of each set of differences Z_{uv}. It is the unadjusted estimator of $M_u – M_v$. Sice $Z_{vu} = Z_{uv}$, we have only to calculate the cases u < v. $Z_{uu} = 0$.

3. Compute the mean of each set of unadjusted medians having the same first subscript m_u:

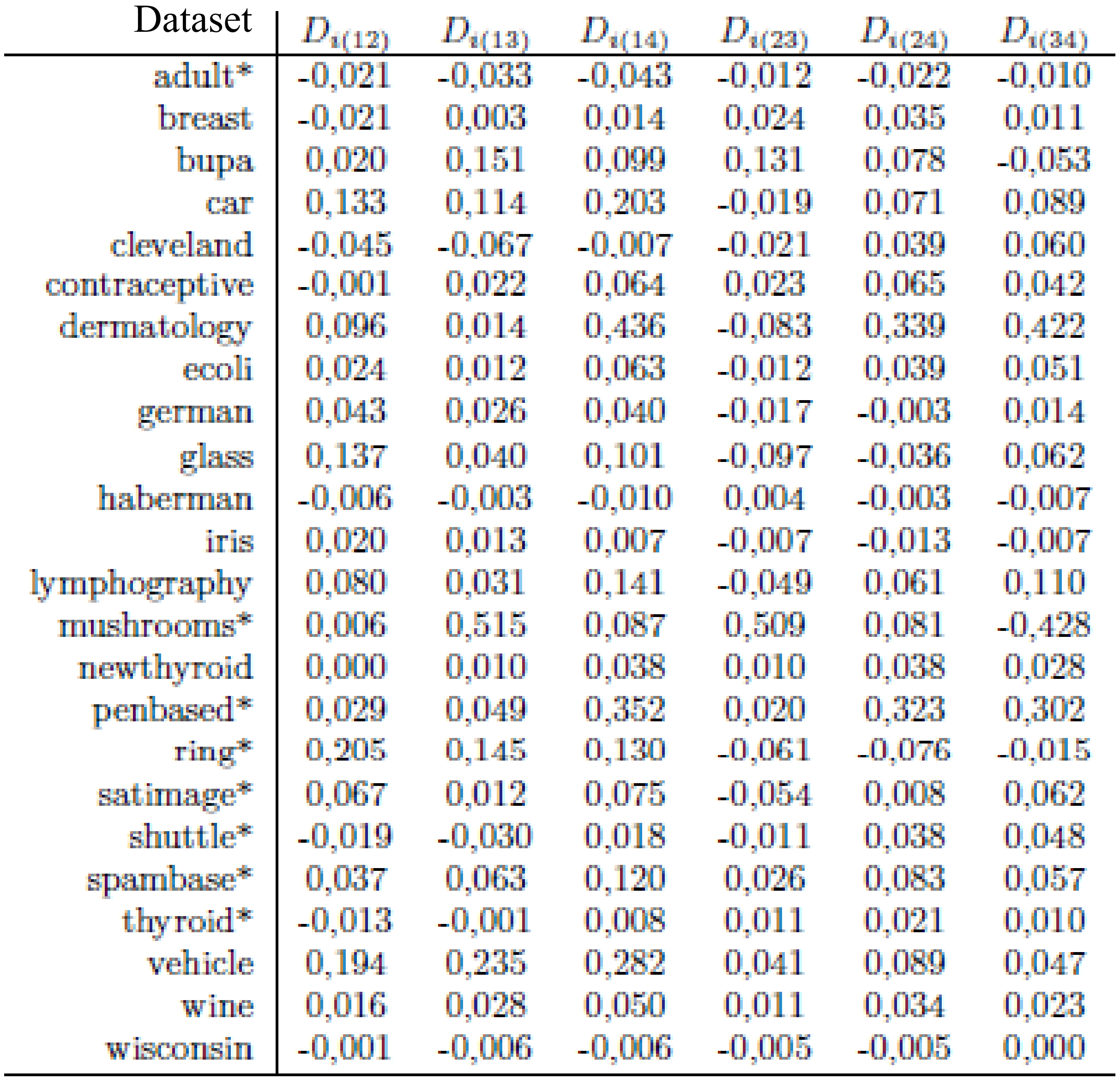
$$m_u = \frac{\sum_{j=1}^{k} Z_{uj}}{k}, u = 1, \ldots, k$$

4. The estimator of $M_u – M_v$ is $m_u - m_v$

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

Contrast Estimation based on Medians



Dataset	$D_{i(12)}$	$D_{i(13)}$	$D_{i(14)}$	$D_{i(23)}$	$D_{i(24)}$	$D_{i(34)}$
adult*	-0,021	-0,033	-0,043	-0,012	-0,022	-0,010
breast	-0,021	0,003	0,014	0,024	0,035	0,011
bupa	0,020	0,151	0,099	0,131	0,078	-0,053
car	0,133	0,114	0,203	-0,019	0,071	0,089
cleveland	-0,045	-0,067	-0,007	-0,021	0,039	0,060
contraceptive	-0,001	0,022	0,064	0,023	0,065	0,042
dermatology	0,096	0,014	0,436	-0,083	0,339	0,422
ecoli	0,024	0,012	0,063	-0,012	0,039	0,051
german	0,043	0,026	0,040	-0,017	-0,003	0,014
glass	0,137	0,040	0,101	-0,097	-0,036	0,062
haberman	-0,006	-0,003	-0,010	0,004	-0,003	-0,007
iris	0,020	0,013	0,007	-0,007	-0,013	-0,007
lymphography	0,080	0,031	0,141	-0,049	0,061	0,110
mushrooms*	0,006	0,515	0,087	0,509	0,081	-0,428
newthyroid	0,000	0,010	0,038	0,010	0,038	0,028
penbased*	0,029	0,049	0,352	0,020	0,323	0,302
ring*	0,205	0,145	0,130	-0,061	-0,076	-0,015
satimage*	0,067	0,012	0,075	-0,054	0,008	0,062
shuttle*	-0,019	-0,030	0,018	-0,011	0,038	0,048
spambase*	0,037	0,063	0,120	0,026	0,083	0,057
thyroid*	-0,013	-0,001	0,008	0,011	0,021	0,010
vehicle	0,194	0,235	0,282	0,041	0,089	0,047
wine	0,016	0,028	0,050	0,011	0,034	0,023
wisconsin	-0,001	-0,006	-0,006	-0,005	-0,005	0,000

Advanced Non-Parametric Tests and Case Studies

For Multiple Comparisons involving a Control Method

$$m_1 = \frac{0 + 0.02 + 0.018 + 0.064}{4} = 0.026$$

$$m_2 = \frac{-0.02 + 0 + (-0.006) + 0.038}{4} = 0.003$$

Our estimate is $m_1 - m_2$:

0.023

Contrast Estimation based on medians among all the algorithms of the case study presented

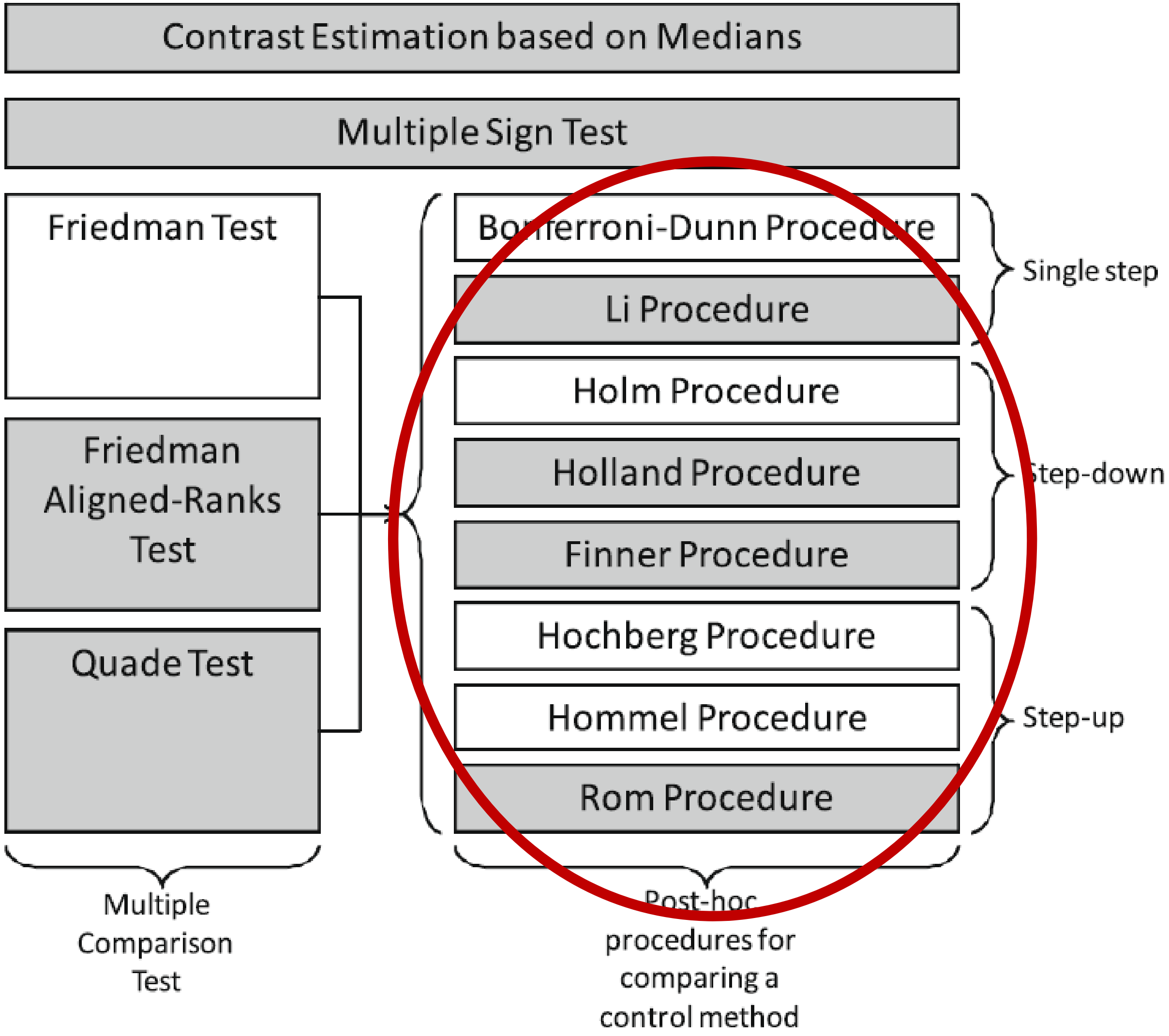
	PDFC	NNEP	IS-CHC+1NN	FH-GBML
PDFC	0.000	0.023	0.020	0.060
NNEP	-0.023	0.000	-0.003	0.037
IS-CHC+1NN	-0.020	0.003	0.000	0.040
FH-GBML	-0.060	-0.037	-0.040	0.000

Advanced non-parametric tests and case studies

- For Multiple Comparisons involving a control method
- Post-hoc Procedures
- Adjusted p-values
- Detecting all pairwise differences in a multiple comparison

Advanced Non-Parametric Tests and Case Studies

Post-hoc Procedures



Advanced Non-Parametric Tests and Case Studies

Post-hoc Procedures

Multiple Comparison tests are focused on the comparison between a control method, which is usually the proposed method, and a set of algorithms used in the empirical study. This set of comparisons is associated with a set or family of hypotheses, all of which are related to the control method. Any of the post hoc tests is suitable for application to nonparametric tests working over a family of hypotheses. The test statistic for comparing the ith algorithm and jth algorithm depends on the main nonparametric procedure used:

- Friedman $z = (R_i - R_j) \Big/ \sqrt{\frac{k(k+1)}{6n}}$
- Friedman Aligned Ranks $z = (\hat{R}_i - \hat{R}_j) \Big/ \sqrt{\frac{k(n+1)}{6}}$
- Quade $z = (T_i - T_j) \Big/ \sqrt{\frac{k(k+1)(2n+1)(k-1)}{18n(n+1)}}$ where $T_i = \frac{W_i}{n(n+1)/2}$, $T_j = \frac{W_j}{n(n+1)/2}$

Advanced Non-Parametric Tests and Case Studies

Post-hoc Procedures

REMEMBER: Three classical post-hoc procedures have been used in mutiple comparisons tests:

- **Bonferroni-Dunn:** controls the family-wise error rate by dividing a by the number of comparisons made ($k-1$).

- **Holm:** Step-down procedure that sequentially test the hypotheses ordered by their significance. We will denote the ordered p values by p_1, p_2, ..., so that $p_1 \leq p_2 \leq . . . \leq p_{k-1}$. It starts with the most significant p value. If p_1 is below $\alpha/(k-1)$, the corresponding hypothesis is rejected and we are allowed to compare p_2 with $\alpha /(k-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on.

- **Hochberg:** step-up procedure that works in the opposite direction, comparing the largest p value with α, the next largest with $\alpha /2$ and so forth until it encounters a hypothesis it can reject

Post-hoc Procedures

Hommel: is more complicated to compute and understand. First, we need to find the largest *j* for which $p_{n-j+k} > k\alpha/j$ for all $k = 1, . . . , j$. If no such *j* exists, we can reject all hypotheses, otherwise we reject all for which $p_i \leq \alpha/j$.

Holland: it also adjusts the value of a in a step-down manner, as Holm's method does. It rejects H_1 to H_{i-1} if *i* is the smallest integer so that $p_i > 1 - (1 - \alpha)^{k - i}$.

Advanced Non-Parametric Tests and Case Studies

Post-hoc Procedures

Finner: it also adjusts the value of a in a step-down manner, as Holm's or Holland's method do. It rejects H_1 to H_{i-1} if *i* is the smallest integer so that $p_i > 1 - (1 - \alpha)^{(k-1)/i}$.

Rom: Rom developed a modification to Hochberg's procedure to increase its power. It works in exactly the same way as the Hochberg procedure, except that the a values are computed through the expression

$$\alpha_{k-i} = \left[\sum_{j=1}^{i-1} \alpha^j - \sum_{j=1}^{i-2} \binom{i}{k} \alpha_{k-1-j}^{i-j} \right] \Big/ i$$

where $\alpha_{k-1} = \alpha$ and $\alpha_{k-2} = \alpha/2$.

Post-hoc Procedures

- **Li - 2 steps rejection procedure**:

- Step 1: Reject all H_i if $p_{k-1} \leq \alpha$. Otherwise, accept the hypothesis associated to p_{k-1} and got to step 2.

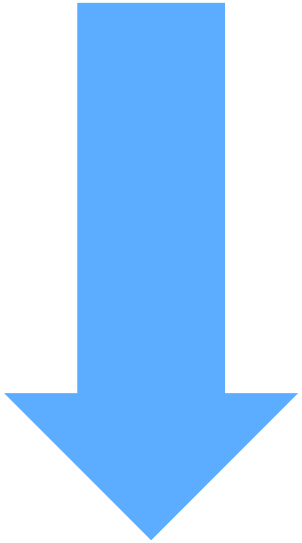
- Step 2: Reject any remaining H_i with $p_i \leq (1-p_{k-1})/(1-\alpha)\alpha$

Advanced Non-Parametric Tests and Case Studies

Post-hoc Procedures

A set of post-hoc procedures:

- one-step:
  - Bonferroni-Dunn
- step-down:
  - Holm
  - Holland
  - Finner
- step-up:
  - Hochberg
  - Hommel
  - Rom
- two-step:
  - Li



They are more powerful according this direction

Source: 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.*

Advanced non-parametric tests and case studies

- For Multiple Comparisons involving a control method
- Post-hoc Procedures
- Adjusted p-values
- Detecting all pairwise differences in a multiple comparison

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values

In statistical hypothesis testing, the p-value is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

The smallest level of significance that results in the rejection of the null hypothesis, the *p-value,* is a useful and interesting datum for many consumers of statistical analysis.

A p-value *provides* information about whether a statistical hypothesis test is significant or not, and it also indicates something about ”how significant” the result is: The smaller the *p-value, the stronger the evidence* against the null hypothesis. Most important, it does this without committing to a particular level of significance.

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values

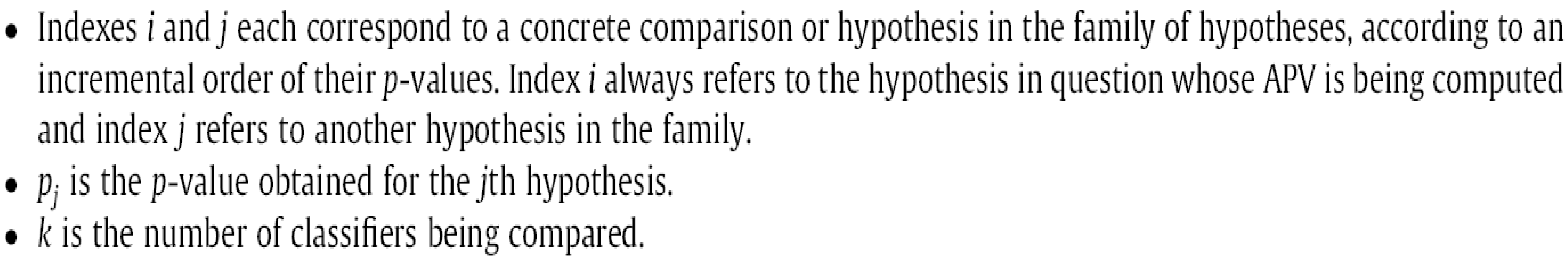
One way to solve this problem is to report adjusted p-values (APVs) which take into account that multiple tests are conducted.

An APV can be compared directly with any chosen significance level α.

We recommend the use of APVs due to the fact that they provide more information in a statistical analysis.

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values



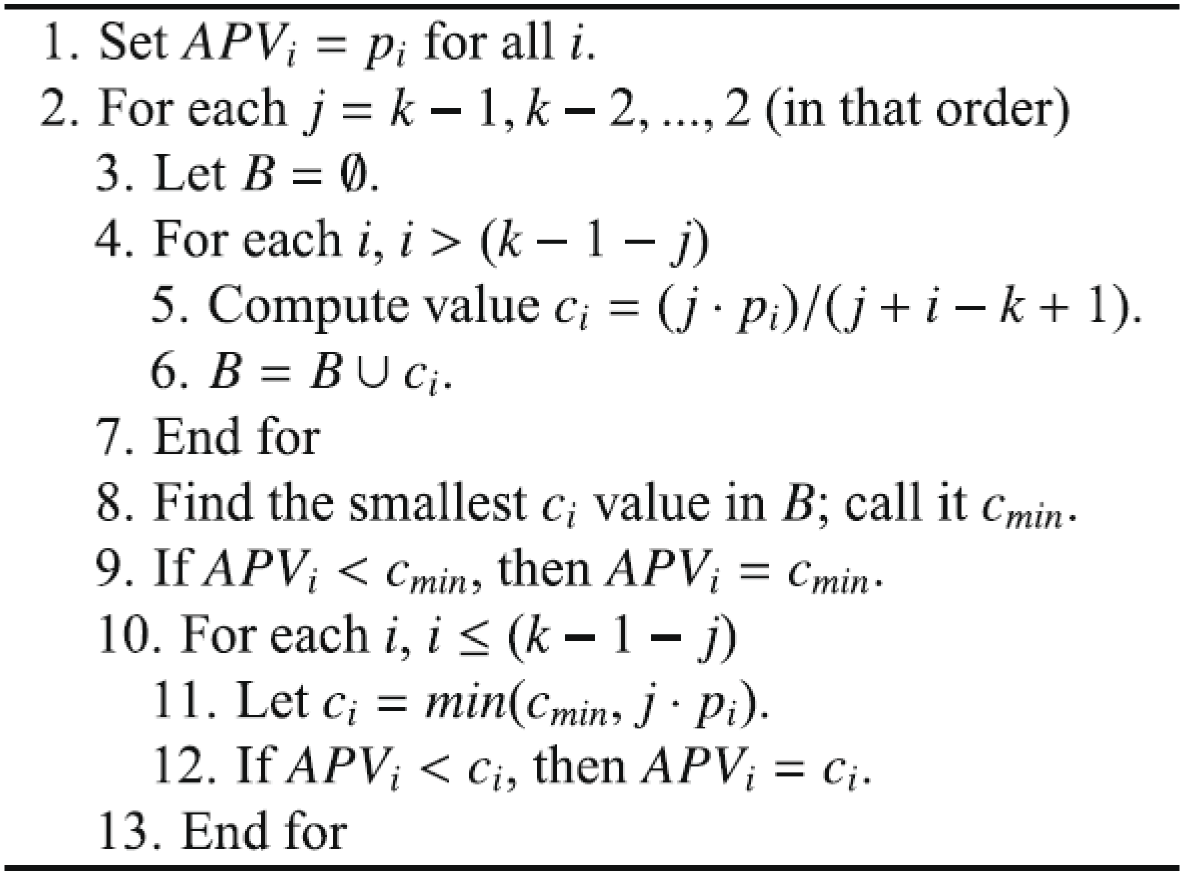
- Indexes *i* and *j* each correspond to a concrete comparison or hypothesis in the family of hypotheses, according to an incremental order of their *p*-values. Index *i* always refers to the hypothesis in question whose APV is being computed and index *j* refers to another hypothesis in the family.
- p_j is the *p*-value obtained for the *j*th hypothesis.
- *k* is the number of classifiers being compared.

APVs for each post-hoc procedure:

- one-step:
  - Bonferroni-Dunn APV_i: $\min\{v; 1\}$, where $v = (k-1)p_i$.
- step-down:
  - Holm APV_i: $\min\{v; 1\}$, where $v = \max\{(k-j)p_j : 1 \leqslant j \leqslant i\}$.
  - Holland APV_i: $\min\{v; 1\}$, where $v = \max\{1-(1-p_j)^{k-j} : 1 \leqslant j \leqslant i\}$
  - Finner APV_i: $\min\{v; 1\}$, where $v = \max\{1-(1-p_j)^{(k-1)/j} : 1 \leqslant j \leqslant i\}$
- step-up:
  - Hochberg APV_i: $\max\{(k-j)p_j : (k-1) \geqslant j \geqslant i\}$
  - Hommel (very difficult to compute, next slide)
  - Rom $APV_i : \max\{(r_{k-j})p_j : (k-1) \geqslant j \geqslant i\}$
- two-step:
  - Li $APV_i : p_i/(p_i + 1 - p_{k-1})$

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values



```
1. Set $APV_i = p_i$ for all $i$.
2. For each $j = k - 1, k - 2, ..., 2$ (in that order)
    3. Let $B = \emptyset$.
    4. For each $i$, $i > (k - 1 - j)$
        5. Compute value $c_i = (j \cdot p_i)/(j + i - k + 1)$.
        6. $B = B \cup c_i$.
    7. End for
    8. Find the smallest $c_i$ value in $B$; call it $c_{min}$.
    9. If $APV_i < c_{min}$, then $APV_i = c_{min}$.
    10. For each $i$, $i \leq (k - 1 - j)$
        11. Let $c_i = min(c_{min}, j \cdot p_i)$.
        12. If $APV_i < c_i$, then $APV_i = c_i$.
    13. End for
```

Fig. 2. Algorithm for calculating APVs based on Hommel's procedure.

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values

APVs in CEC'2005 Case Study

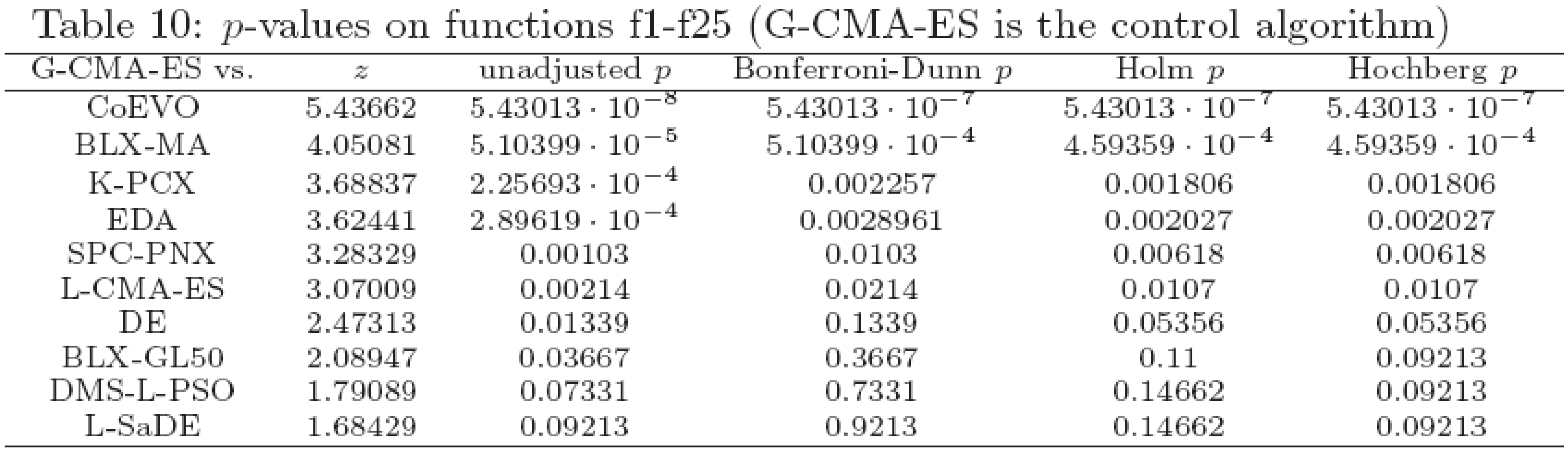


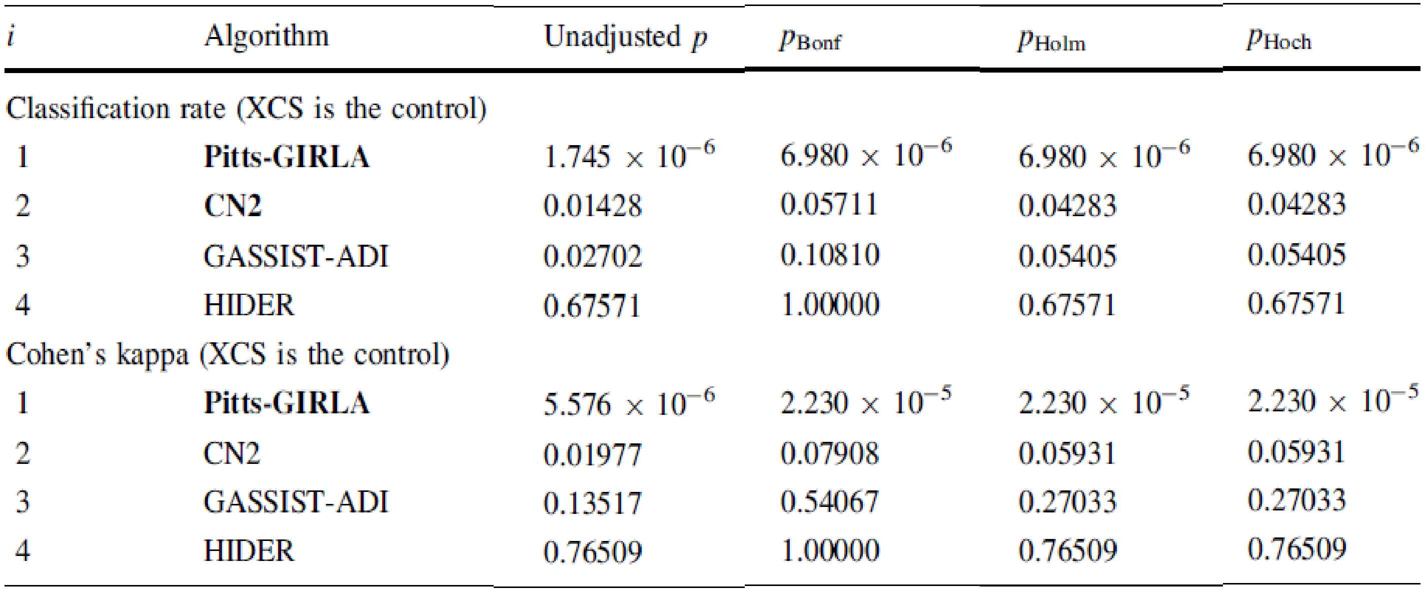
Table 10: p-values on functions f1-f25 (G-CMA-ES is the control algorithm)

G-CMA-ES vs.	z	unadjusted p	Bonferroni-Dunn p	Holm p	Hochberg p
CoEVO	5.43662	$5.43013 \cdot 10^{-8}$	$5.43013 \cdot 10^{-7}$	$5.43013 \cdot 10^{-7}$	$5.43013 \cdot 10^{-7}$
BLX-MA	4.05081	$5.10399 \cdot 10^{-5}$	$5.10399 \cdot 10^{-4}$	$4.59359 \cdot 10^{-4}$	$4.59359 \cdot 10^{-4}$
K-PCX	3.68837	$2.25693 \cdot 10^{-4}$	0.002257	0.001806	0.001806
EDA	3.62441	$2.89619 \cdot 10^{-4}$	0.0028961	0.002027	0.002027
SPC-PNX	3.28329	0.00103	0.0103	0.00618	0.00618
L-CMA-ES	3.07009	0.00214	0.0214	0.0107	0.0107
DE	2.47313	0.01339	0.1339	0.05356	0.05356
BLX-GL50	2.08947	0.03667	0.3667	0.11	0.09213
DMS-L-PSO	1.79089	0.07331	0.7331	0.14662	0.09213
L-SaDE	1.68429	0.09213	0.9213	0.14662	0.09213

- **In practice, Hochberg's method is more powerful than Holm's one (but this difference is rather small), in this the results are in favour of Hochberg's method.**

APVs in GBMLs Case Study

Adjusted *p*-values for the comparison of the control algorithm in each measure with the remaining algorithms



i	Algorithm	Unadjusted p	p_{Bonf}	p_{Holm}	p_{Hoch}
Classification rate (XCS is the control)					
1	**Pitts-GIRLA**	1.745×10^{-6}	6.980×10^{-6}	6.980×10^{-6}	6.980×10^{-6}
2	**CN2**	0.01428	0.05711	0.04283	0.04283
3	GASSIST-ADI	0.02702	0.10810	0.05405	0.05405
4	HIDER	0.67571	1.00000	0.67571	0.67571
Cohen's kappa (XCS is the control)					
1	**Pitts-GIRLA**	5.576×10^{-6}	2.230×10^{-5}	2.230×10^{-5}	2.230×10^{-5}
2	CN2	0.01977	0.07908	0.05931	0.05931
3	GASSIST-ADI	0.13517	0.54067	0.27033	0.27033
4	HIDER	0.76509	1.00000	0.76509	0.76509

- If the adjusted *p* for each method is lower than the desired level of confidence α (0.05 in our case), the algorithms are worse from bottom to top (stress in bold for 0.05)
- In practice, Hochberg's method is more powerful than Holm's one (but this difference is rather small), in this our study the results are the same.

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values

APVs in ANNs Case Study

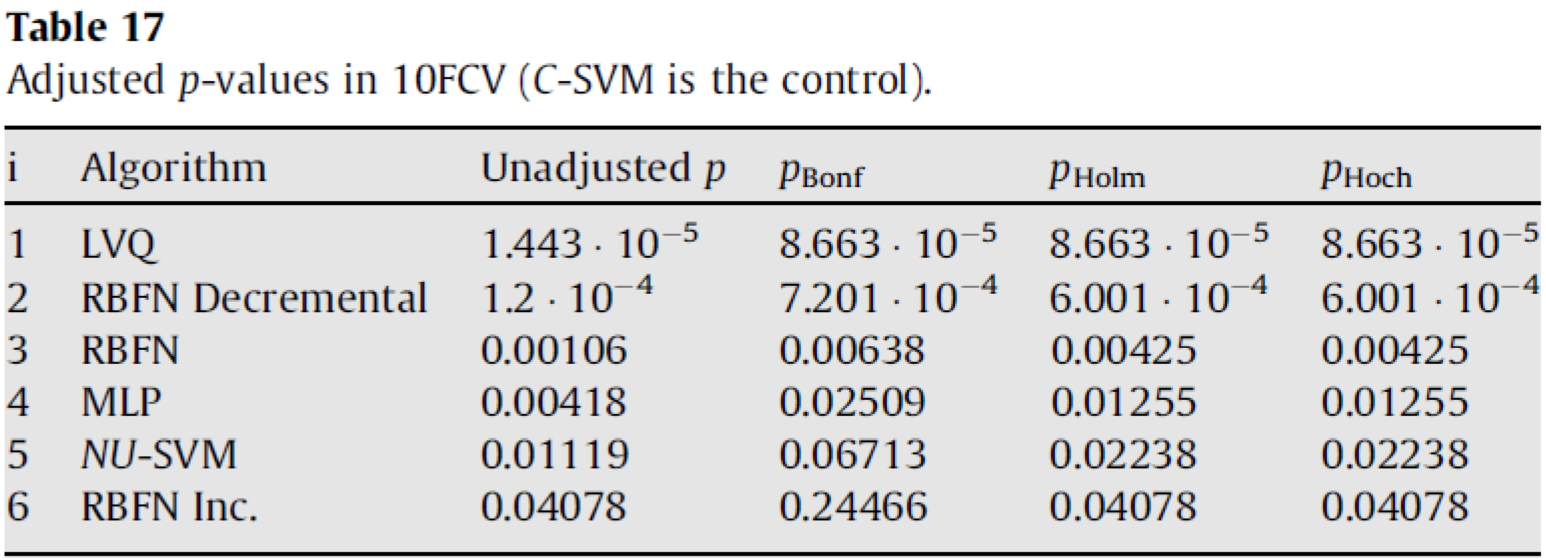


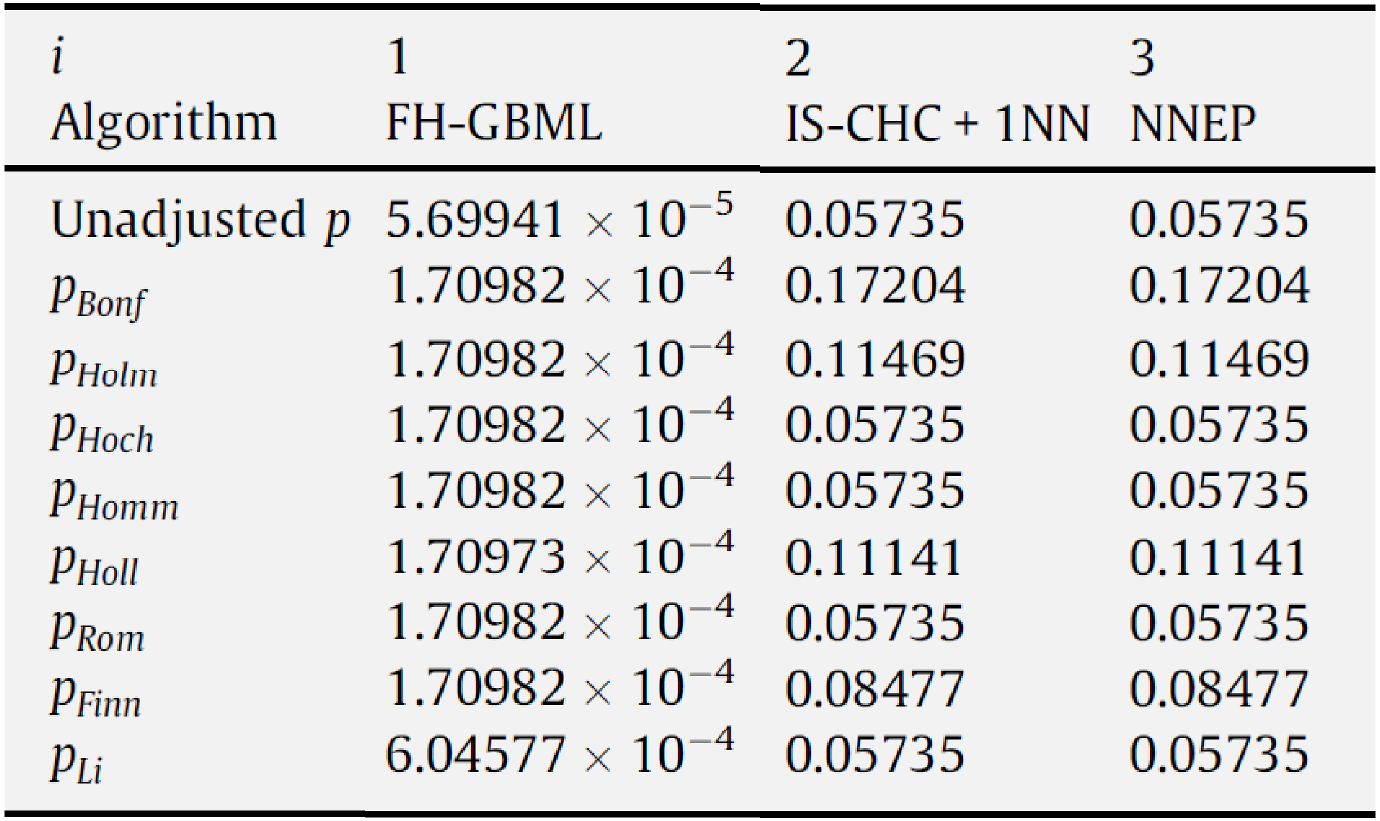
Table 17
Adjusted *p*-values in 10FCV (*C*-SVM is the control).

i	Algorithm	Unadjusted *p*	p_{Bonf}	p_{Holm}	p_{Hoch}
1	LVQ	$1.443 \cdot 10^{-5}$	$8.663 \cdot 10^{-5}$	$8.663 \cdot 10^{-5}$	$8.663 \cdot 10^{-5}$
2	RBFN Decremental	$1.2 \cdot 10^{-4}$	$7.201 \cdot 10^{-4}$	$6.001 \cdot 10^{-4}$	$6.001 \cdot 10^{-4}$
3	RBFN	0.00106	0.00638	0.00425	0.00425
4	MLP	0.00418	0.02509	0.01255	0.01255
5	*NU*-SVM	0.01119	0.06713	0.02238	0.02238
6	RBFN Inc.	0.04078	0.24466	0.04078	0.04078

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values

APVs for all post-hoc procedures in Friedman PDFC is the control

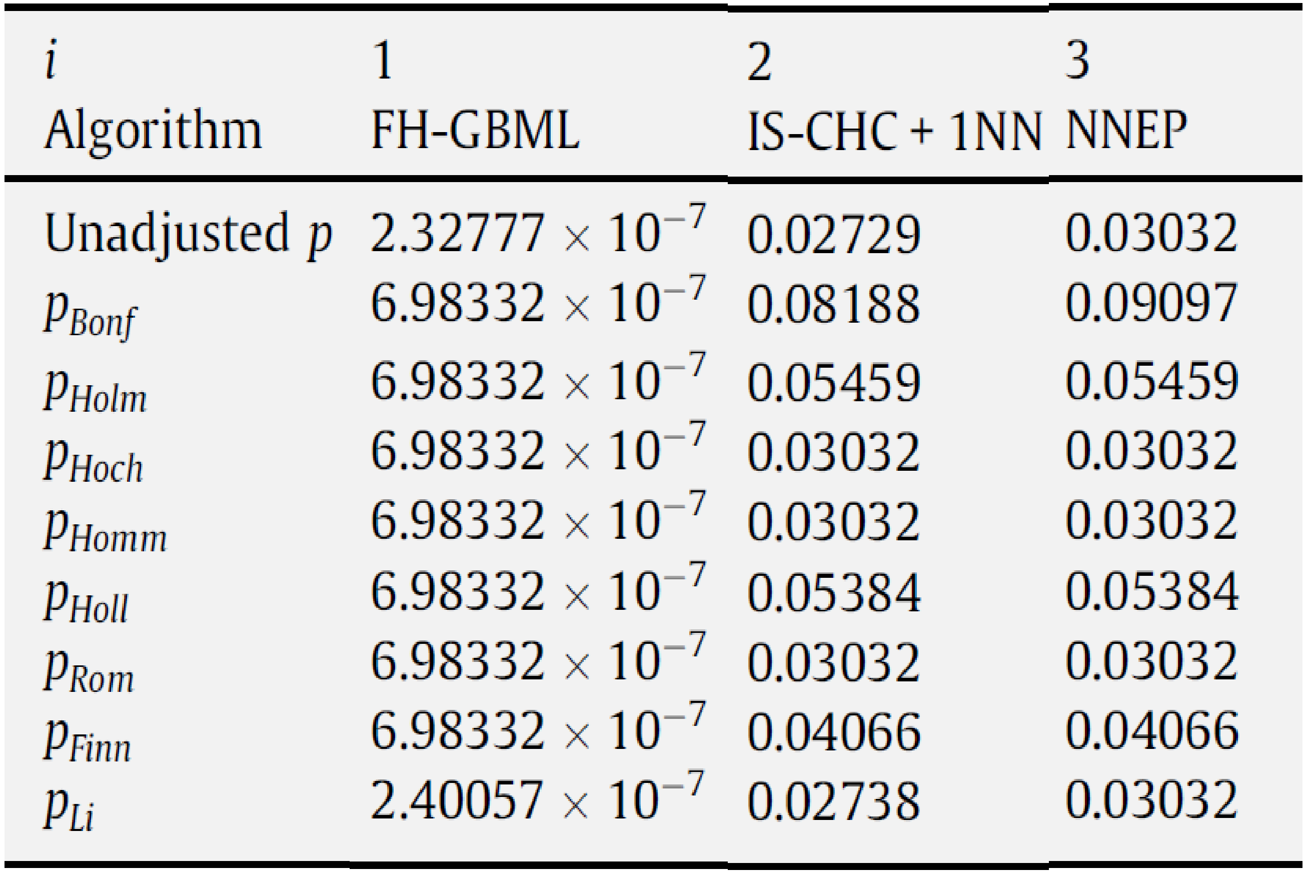


i	1	2	3
Algorithm	FH-GBML	IS-CHC + 1NN	NNEP
Unadjusted p	5.69941×10^{-5}	0.05735	0.05735
p_{Bonf}	1.70982×10^{-4}	0.17204	0.17204
p_{Holm}	1.70982×10^{-4}	0.11469	0.11469
p_{Hoch}	1.70982×10^{-4}	0.05735	0.05735
p_{Homm}	1.70982×10^{-4}	0.05735	0.05735
p_{Holl}	1.70973×10^{-4}	0.11141	0.11141
p_{Rom}	1.70982×10^{-4}	0.05735	0.05735
p_{Finn}	1.70982×10^{-4}	0.08477	0.08477
p_{Li}	6.04577×10^{-4}	0.05735	0.05735

Advanced Non-Parametric Tests and Case Studies

Adjusted P-Values

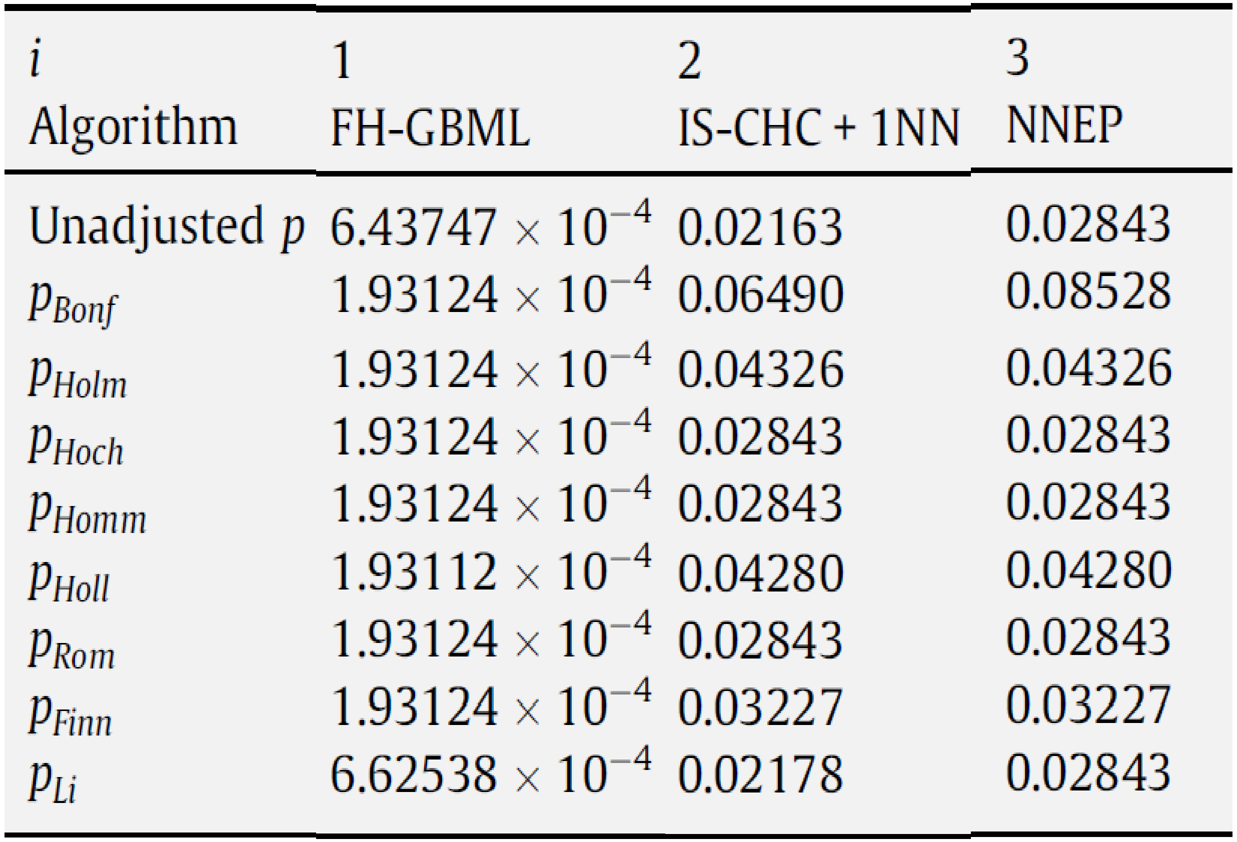
APVs for all post-hoc procedures in Friedman Aligned Ranks PDFC is the control



i	1	2	3
Algorithm	FH-GBML	IS-CHC + 1NN	NNEP
Unadjusted p	2.32777×10^{-7}	0.02729	0.03032
p_{Bonf}	6.98332×10^{-7}	0.08188	0.09097
p_{Holm}	6.98332×10^{-7}	0.05459	0.05459
p_{Hoch}	6.98332×10^{-7}	0.03032	0.03032
p_{Homm}	6.98332×10^{-7}	0.03032	0.03032
p_{Holl}	6.98332×10^{-7}	0.05384	0.05384
p_{Rom}	6.98332×10^{-7}	0.03032	0.03032
p_{Finn}	6.98332×10^{-7}	0.04066	0.04066
p_{Li}	2.40057×10^{-7}	0.02738	0.03032

APVs for all post-hoc procedures in Quade

PDFC is the control



i	1	2	3
Algorithm	FH-GBML	IS-CHC + 1NN	NNEP
Unadjusted p	6.43747×10^{-4}	0.02163	0.02843
p_{Bonf}	1.93124×10^{-4}	0.06490	0.08528
p_{Holm}	1.93124×10^{-4}	0.04326	0.04326
p_{Hoch}	1.93124×10^{-4}	0.02843	0.02843
p_{Homm}	1.93124×10^{-4}	0.02843	0.02843
p_{Holl}	1.93112×10^{-4}	0.04280	0.04280
p_{Rom}	1.93124×10^{-4}	0.02843	0.02843
p_{Finn}	1.93124×10^{-4}	0.03227	0.03227
p_{Li}	6.62538×10^{-4}	0.02178	0.02843

Advanced non-parametric tests and case studies

- For Multiple Comparisons involving a control method
- Post-hoc Procedures
- Adjusted p-values
- Detecting all pairwise differences in a multiple comparison

Detecting all pairwise differences in a multiple comparison:

Until now, we have studied the techniques for multiple comparison using a control method. But, under some circumstances, it would be interesting to conduct a test over all possible comparisons involved in the experimental study

It is the usual case in review papers. In these cases, the repetition of comparisons choosing different control classifiers may lose the control of the family-wise error.

The post-hoc procedures need to control the FWER under more restrictive corrections because the family of hypotheses is formed now for k(k-1)/2 comparisons instead of (k-1).

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison

Remember (Friedman):

A set of pairwise comparisons can be associated with a set or family of hypotheses. Any of the post-hoc tests which can be applied to non-parametric tests work over a family of hypotheses.

The test statistics for comparing the i-th and j-th classifier is

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

The *z* value is used to find the corresponding probability (p-value) from the table of normal distribution, which is then compared with an appropriate level of significance a (Table A1 in Sheskin, 2003)

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison

REMEMBER: Two classical post-hoc procedures have been used in mutiple comparisons tests and also valid in n x n comparisons:

- **Bonferroni-Dunn (Nemenyi in n x n comparisons):** controls the family-wise error rate by dividing a by the number of comparisons made $m = k(k-1)/2$.

- **Holm:** Step-down procedure that sequentially test the hypotheses ordered by their significance. We will denote the ordered p values by $p_1, p_2, \ldots$, so that $p_1 \leq p_2 \leq . . . \leq p_{k-1}$. It starts with the most significant p value. If p_1 is below $\alpha/(m-1)$, the corresponding hypothesis is rejected and we are allowed to compare p_2 with $\alpha /(m-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on.

- Hochberg, Hommel, Rom, Finner are also valid….

Logically Related Hypotheses:

The hypotheses being tested belonging to a family of all pairwise comparisons are logically interrelated so that not all combinations of true and false hypotheses are possible.

As a simple example of such a situation suppose that we want to test the three hypotheses of pairwise equality associated with the pairwise comparisons of three classifiers C_i; i = 1,2,3. It is easily seen from the relations among the hypotheses that if any one of them is false, at least one other must be false. For example, if C_1 is different than C_2, then it is not possible that C_1 has the same performance than C_3 and C_2 has the same performance than C_3. C_3 must be different than C_1 or C_2 or both.

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison

Shaffer's procedure: following Holm's step down method, at stage *j*, instead of rejecting H_i if $p_i \leq \alpha$ / (m-i+1), reject H_i if $p_i \leq \alpha$ / t_i, where t_i is the maximum number of hypotheses which can be true given that any (*i* - *1*) hypotheses are false.

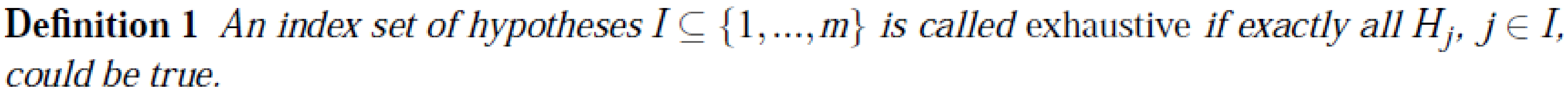
It is a static procedure, that is, t_1, …, t_m are fully determined for the given hypotheses H_1, …, H_m, independent of the observed p-values. The possible numbers of true hypotheses, and thus the values of t_i can be obtained from the recursive formula

$$S(k) = \bigcup_{j=1}^{k} \left\{ \binom{j}{2} + x : x \in S(k-j) \right\},$$

where S(k) is the set of possible numbers of true hypotheses with k classifiers being compared, k ≥ 2, and S(0) = S(1) = {0}.

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison



Definition 1 *An index set of hypotheses* $I \subseteq \{1,\ldots,m\}$ *is called* exhaustive *if exactly all* H_j, $j \in I$, *could be true.*

Bergmann-Hommel's procedure: Reject all Hj with j not in A, where the acceptance set

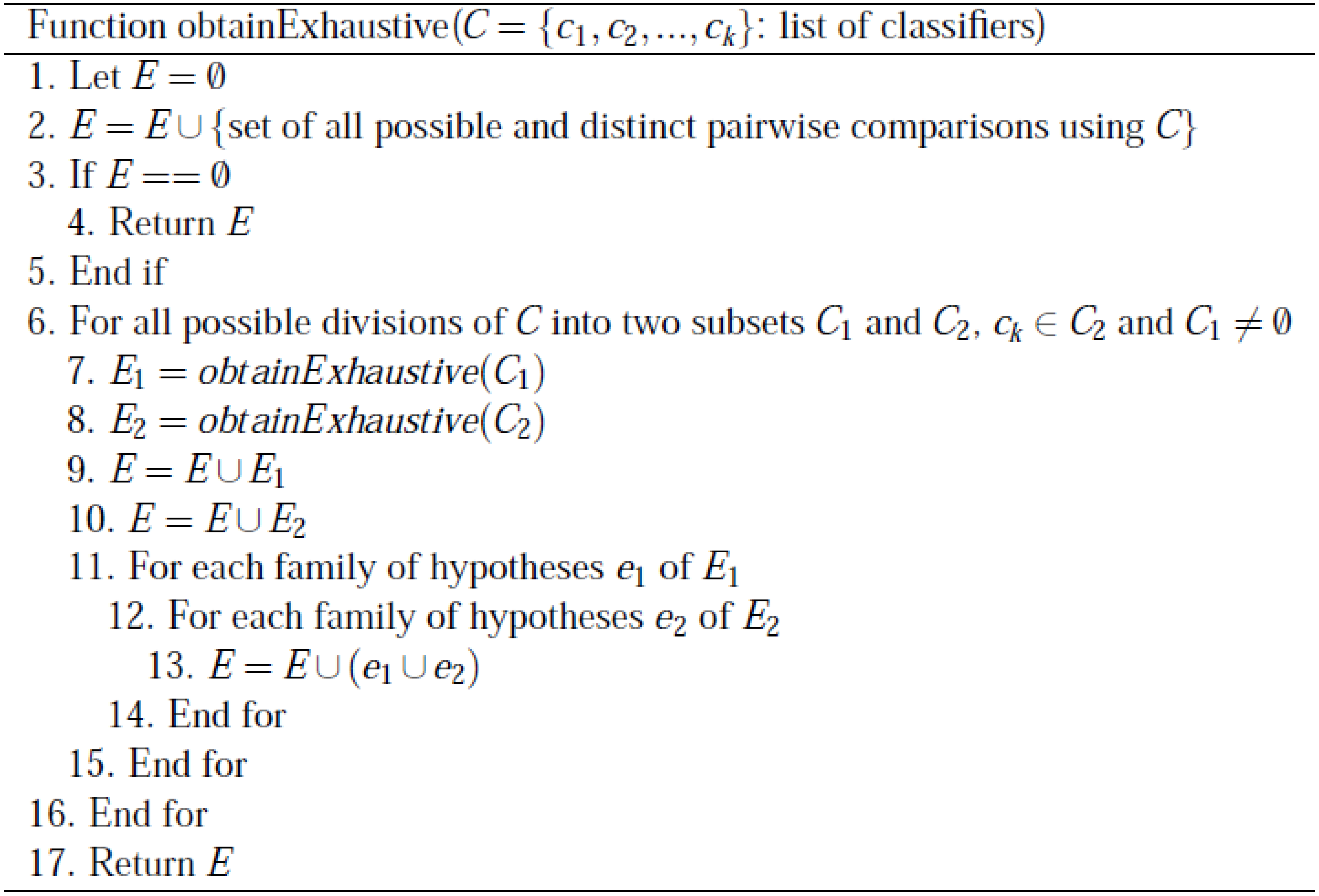
$$A = \bigcup \{I : I\ exhaustive,\ min\{P_i : i \in I\} > \alpha/|I|\}$$

is the index set of null hypotheses which are retained.

For this procedure, one has to check for each subset I of {1,…,m} if I is exhaustive, which leads to intensive computation. Due to this fact, we will obtain a set, named E, which will contain all the possible exhaustive sets of hypotheses for a certain comparison. Once the E set is obtained, the hypotheses that do not belong to the A set are rejected.

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison



```
Function obtainExhaustive($C = \{c_1, c_2, ..., c_k\}$: list of classifiers)
Let $E = \emptyset$
$E = E \cup \{$set of all possible and distinct pairwise comparisons using $C\}$
If $E == \emptyset$
    Return $E$
End if
For all possible divisions of $C$ into two subsets $C_1$ and $C_2$, $c_k \in C_2$ and $C_1 \neq \emptyset$
    $E_1 = obtainExhaustive(C_1)$
    $E_2 = obtainExhaustive(C_2)$
    $E = E \cup E_1$
    $E = E \cup E_2$
    For each family of hypotheses $e_1$ of $E_1$
        For each family of hypotheses $e_2$ of $E_2$
            $E = E \cup (e_1 \cup e_2)$
        End for
    End for
End for
Return $E$
```

Figure 1: Algorithm for obtaining all exhaustive sets

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison

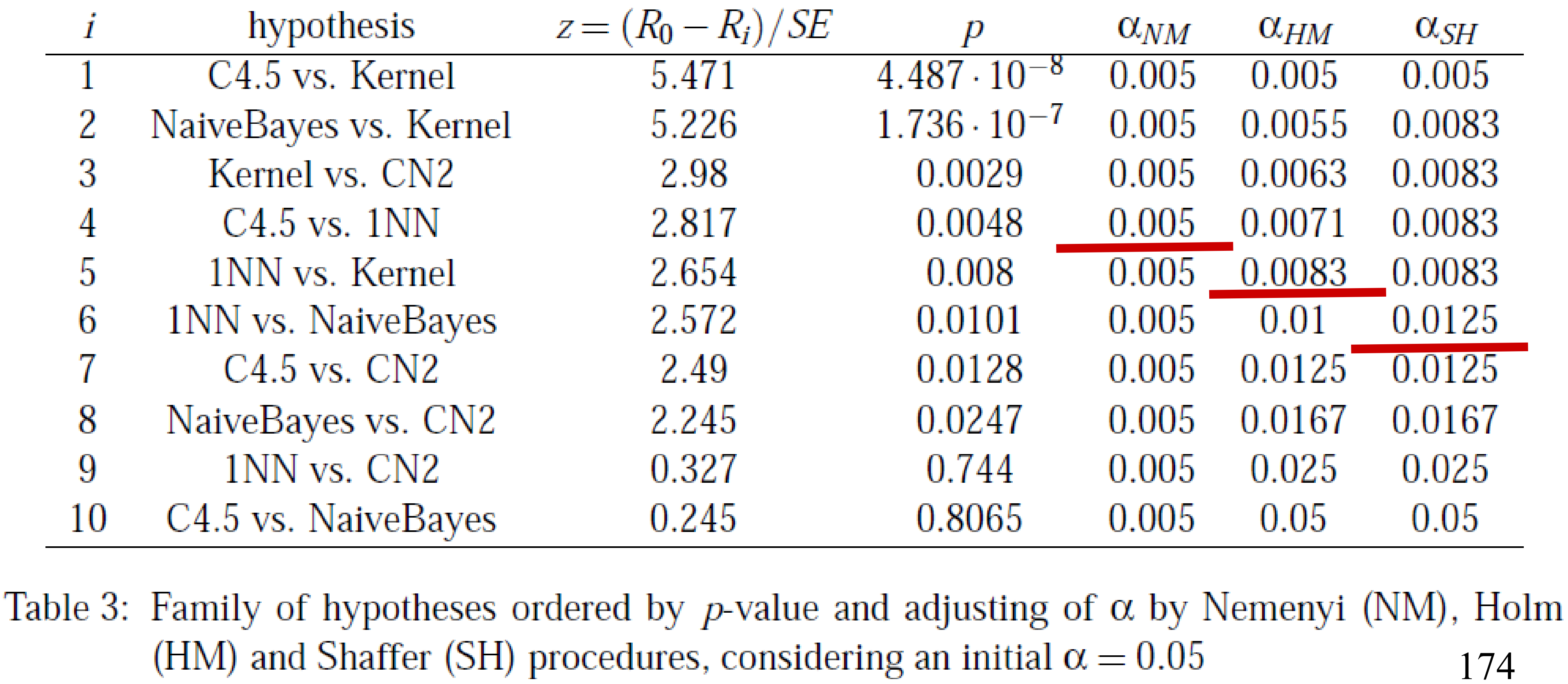
Case Study used:

- 30 data sets from UCI and KEEL data-set
- Classifiers (from KEEL, standard parameters values):
  - C4.5
  - 1NN
  - Naïve Bayes
  - Kernel
  - CN2
- Rankings computed by Friedman test

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison

	C4.5	1NN	NaiveB	Kernel	CN2
Average Rank	**2.100**	**3.250**	**2.200**	**4.333**	**3.117**

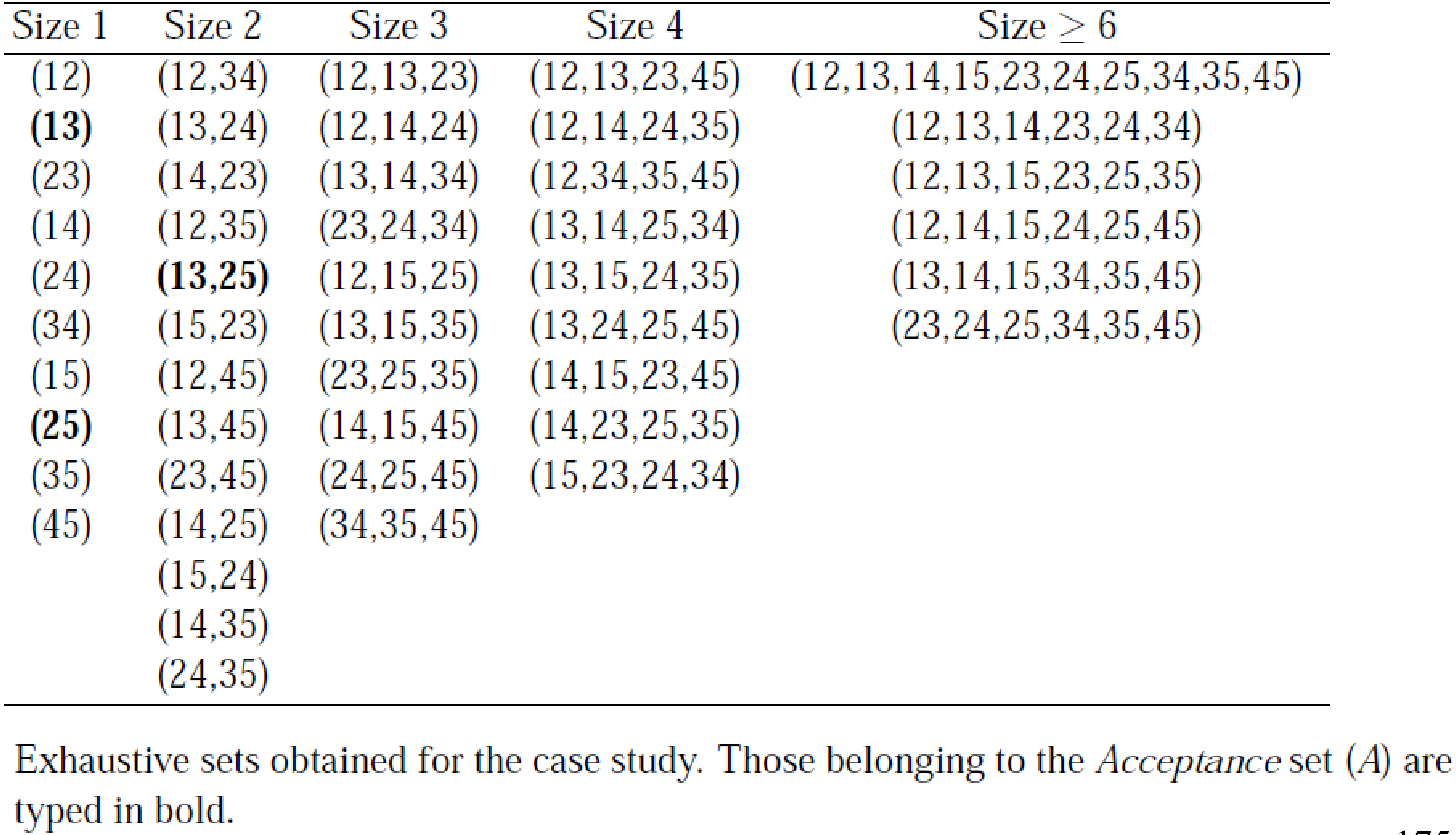


i	hypothesis	$z = (R_0 - R_i)/SE$	p	α_{NM}	α_{HM}	α_{SH}
1	C4.5 vs. Kernel	5.471	$4.487 \cdot 10^{-8}$	0.005	0.005	0.005
2	NaiveBayes vs. Kernel	5.226	$1.736 \cdot 10^{-7}$	0.005	0.0055	0.0083
3	Kernel vs. CN2	2.98	0.0029	0.005	0.0063	0.0083
4	C4.5 vs. 1NN	2.817	0.0048	0.005	0.0071	0.0083
5	1NN vs. Kernel	2.654	0.008	0.005	0.0083	0.0083
6	1NN vs. NaiveBayes	2.572	0.0101	0.005	0.01	0.0125
7	C4.5 vs. CN2	2.49	0.0128	0.005	0.0125	0.0125
8	NaiveBayes vs. CN2	2.245	0.0247	0.005	0.0167	0.0167
9	1NN vs. CN2	0.327	0.744	0.005	0.025	0.025
10	C4.5 vs. NaiveBayes	0.245	0.8065	0.005	0.05	0.05

Table 3: Family of hypotheses ordered by p-value and adjusting of α by Nemenyi (NM), Holm (HM) and Shaffer (SH) procedures, considering an initial $\alpha = 0.05$

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison



Size 1	Size 2	Size 3	Size 4	Size $\geq$ 6
(12)	(12,34)	(12,13,23)	(12,13,23,45)	(12,13,14,15,23,24,25,34,35,45)
(13)	(13,24)	(12,14,24)	(12,14,24,35)	(12,13,14,23,24,34)
(23)	(14,23)	(13,14,34)	(12,34,35,45)	(12,13,15,23,25,35)
(14)	(12,35)	(23,24,34)	(13,14,25,34)	(12,14,15,24,25,45)
(24)	**(13,25)**	(12,15,25)	(13,15,24,35)	(13,14,15,34,35,45)
(34)	(15,23)	(13,15,35)	(13,24,25,45)	(23,24,25,34,35,45)
(15)	(12,45)	(23,25,35)	(14,15,23,45)	
(25)	(13,45)	(14,15,45)	(14,23,25,35)	
(35)	(23,45)	(24,25,45)	(15,23,24,34)	
(45)	(14,25)	(34,35,45)		
	(15,24)			
	(14,35)			
	(24,35)			

Exhaustive sets obtained for the case study. Those belonging to the *Acceptance* set (A) are typed in bold.

Case Study used:

- ❑ Nemenyi's test rejects the hypotheses [1–4] since the corresponding p-values are smaller than the adjusted α's.
- ❑ Holm's procedure rejects the hypotheses [1–5].
- ❑ Shaffer's static procedure rejects the hypotheses [1–6].
- ❑ Bergmann-Hommel's dynamic procedure first obtains the exhaustive index set of hypotheses. It obtains 51 index sets. We can see them in the previous slide. From the index sets, it computes the A set. It rejects all hypotheses H_j with j not in A, so it rejects the hypotheses [1–8].

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison

- m is the number of possible comparisons in an all pairwise comparisons design; that is, $m = \frac{k\cdot(k-1)}{2}$.
- t_j is the maximum number of hypotheses which can be true given that any $(j-1)$ hypotheses are false

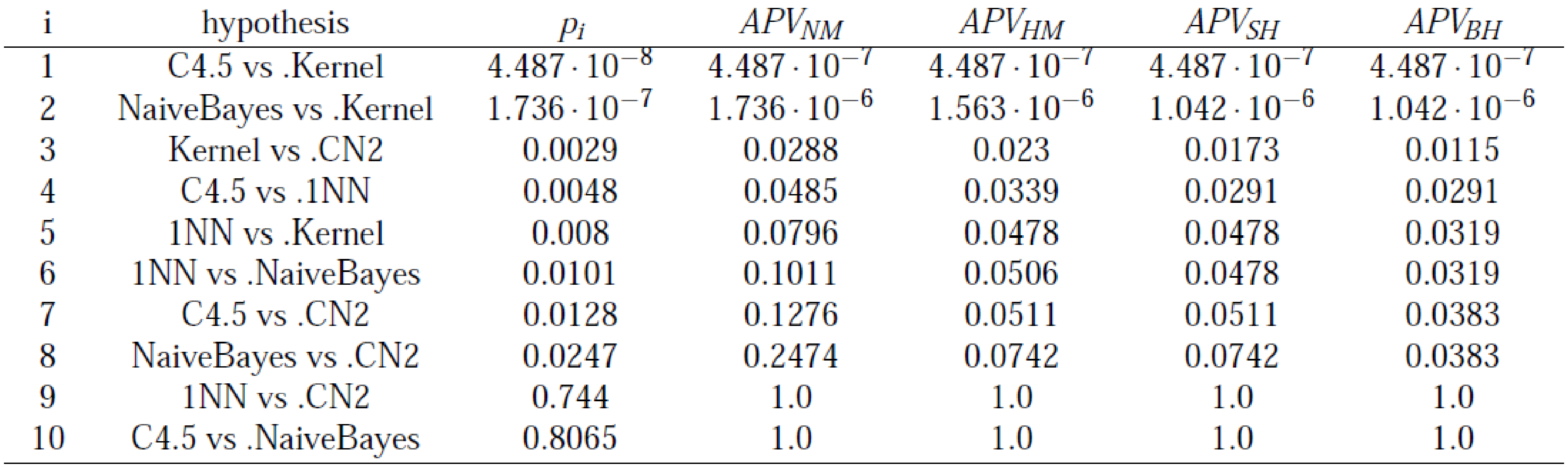
APVs for each post-hoc procedure:

- one-step:
  - Nemenyi APV_i: $min\{v;1\}$, where $v = m \cdot p_i$.
- step-down:
  - Holm APV_i: $min\{v;1\}$, where $v = max\{(m-j+1)p_j : 1 \leq j \leq i\}$.
  - Shaffer APV_i: $min\{v;1\}$, where $v = max\{t_j p_j : 1 \leq j \leq i\}$.
  - Bergmann-Hommel

APV_i: $min\{v;1\}$, where $v = max\{|I| \cdot min\{p_j, j \in I\} : I\, exhaustive, i \in I\}$.

Advanced Non-Parametric Tests and Case Studies

Detecting all Pairwise differences in a Multiple Comparison



i	hypothesis	p_i	APV_{NM}	APV_{HM}	APV_{SH}	APV_{BH}
1	C4.5 vs .Kernel	$4.487 \cdot 10^{-8}$	$4.487 \cdot 10^{-7}$	$4.487 \cdot 10^{-7}$	$4.487 \cdot 10^{-7}$	$4.487 \cdot 10^{-7}$
2	NaiveBayes vs .Kernel	$1.736 \cdot 10^{-7}$	$1.736 \cdot 10^{-6}$	$1.563 \cdot 10^{-6}$	$1.042 \cdot 10^{-6}$	$1.042 \cdot 10^{-6}$
3	Kernel vs .CN2	0.0029	0.0288	0.023	0.0173	0.0115
4	C4.5 vs .1NN	0.0048	0.0485	0.0339	0.0291	0.0291
5	1NN vs .Kernel	0.008	0.0796	0.0478	0.0478	0.0319
6	1NN vs .NaiveBayes	0.0101	0.1011	0.0506	0.0478	0.0319
7	C4.5 vs .CN2	0.0128	0.1276	0.0511	0.0511	0.0383
8	NaiveBayes vs .CN2	0.0247	0.2474	0.0742	0.0742	0.0383
9	1NN vs .CN2	0.744	1.0	1.0	1.0	1.0
10	C4.5 vs .NaiveBayes	0.8065	1.0	1.0	1.0	1.0

APVs obtained in the example by Nemenyi (NM), Holm (HM), Shaffer's static (SH) Bergmann-Hommel's dynamic (BH)

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
- Basic non-parametric tests and case studies
- Advanced non-parametric tests and case studies
- Lessons Learned
- Books of Interest and References
- Software

OUTLINE (II)

- Advanced non-parametric tests and case studies:
  - For Multiple Comparisons involving control method
  - Post-hoc Procedures
  - Adjusted p-values
  - Detecting all pairwise differences in a multiple comparison
- **Lessons Learned**
  - Considerations on the use of nonparametric tests
  - Recommendations on the use of nonparametric tests
  - Frequent Questions
- Books of Interest and References
- Software

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Lessons Learned

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

Lessons Learned

Considerations on the Use of Non-Parametric Tests

On the use of non-parametric tests:

The need of using non-parametric tests given that the necessary conditions for using parametric tests are not verified.

Lessons Learned

Considerations on the Use of Non-Parametric Tests

Wilcoxon’s test

❑ Wilcoxon’s test computes a ranking based on differences between functions independently, whereas Friedman and derivative procedures compute the ranking between algorithms.

❑ Wilcoxon’s test is highly influenced by the number of case of study (functions, data sets …). The N value determines the critical values to search in the statistical table.

It is highly influenced by outliers when N is below or equal to 11.

Lessons Learned

Considerations on the Use of Non-Parametric Tests

Multiple comparison (1)

□ A multiple comparison must be carried out first by using a statistical method for testing the differences among the related samples means. Then to use a post-hoc statistical procedures.

□ Holm’s procedure is a very good test.

Hochberg’s method can rejects more hypothesis than Holm’s one.

Lessons Learned

Considerations on the Use of Non-Parametric Tests

Multiple comparison (2)

- ❑ An appropriate number of algorithms in contrast with an appropriate number of case problems are needed to be used in order to employ each type of test. The number of algorithms used in multiple comparisons procedures must be lower than the number of case problems

- ❑ Both, the Friedman Aligned Rank test and the Quade test, can be used under the same circumstances as the Friedman test. The differences in power between them are unknown, but we encourage the use of these tests when the number of algorithms to be compared is low.

Lessons Learned

Considerations 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

Considerations on the Use of Non-Parametric Tests

Advantages of Nonparametric Tests

- Can treat data which are inherently in ranks as well as data whose seemingly numerical scores have the strength in ranks
- Easier to learn and apply than parametric tests

  (only one run for all cases of test)

If sample sizes as small as N=6 are used, there is no alternative to using a nonparametric test

Lessons Learned

Considerations on the Use of Non-Parametric Tests

Advantages of Nonparametric Tests

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

Considerations on the Use of Non-Parametric Tests

Design of Experiments in Data Mining/Computational Intelligence

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

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

Lessons Learned

Recommendations on the Use of Non-Parametric Tests

Wilcoxon's test

❑ The influence of the number of case problems used is more noticeable in multiple comparisons procedures than in Wilcoxon's test.

❑ It is highly influenced by outliers when N is below or equal to 11.

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)

❑ Holm, Hochberg, Finner and Li are the more recommended post-hoc test to be used due to their trade-off between simplicity and power.

❑ The power of the Li test is highly influenced by the first p-value of the family and when it is lower than 0.5, the test will perform very well.

❑ 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

Multiple comparison with a control (3)

- ❑ An alternative to directly performing a comparison between a control algorithm and a set of algorithms is the Multiple Sign-test. We recommend its use when the differences reported by the control algorithm with respect to the rest of methods are very clear.

- ❑ The Contrast Estimation in nonparametric statistics is used for computing the real differences between two algorithms, considering the median measure the most important.

Lessons Learned

Recommendations on the Use of Non-Parametric Tests

All pairwise comparisons in multiple comparison

□ We do not recommend the use of Nemenyi’s test, because it is a very conservative procedure and many of the obvious differences may not be detected.

□ Conducting the Shaffer static procedure means a not very significant increase of the difficulty with respect to the Holm procedure.

□ Bergmann-Hommel’s procedure is the best performing one, but it is also the most difficult to understand and computationally expensive. We recommend its usage when the situation requires so.

Lessons Learned

- Considerations on the use of non-parametric tests
- 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 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.

Lessons Learned

Frequent Questions

❑ When the Friedman Aligned and Quade tests are recommendable instead of classical Friedman?

❑ The difference of power among the three methods are small and very dependent of the sample of results to be analyzed. Theoretical studies demonstrate that the Aligned Friedman and the Quade tests have better performance when we compare not more than 4 algorithms. The Quade test also assumes some risk because it considers that the more relevant problems are also those which present higher differences in performance among the methods, and it is not always true.

Lessons Learned

Frequent Questions

❑ Which post-hoc procedures should be used?

❑ We consider that the Holm test must appear in a comparison, wheres Bonferroni does not. Hochberg and Li tests could act as a complement when their use allow us to reject more hypotheses than Holm's. All rejected hyphotesis by any procedure is correctly rejected because all procedures perform a strong control of the FWER.

❑However, some tests, such as Li, are influenciated by the unadjusted p-values of the initial hypotheses and when the are lesser than 0.5, is the only case in which the test achieves its best performance of power.

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
- Basic non-parametric tests and case studies
- Advanced non-parametric tests and case studies:
- Lessons Learned
- Books of Interest and References
- Software

Books of interest and References

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*.

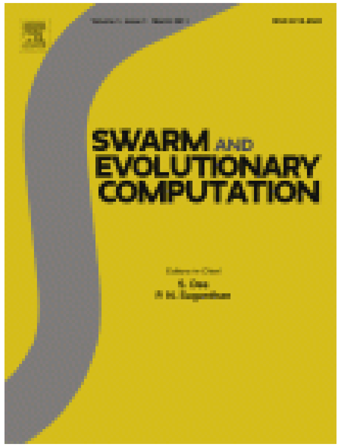
P3: S. García, A. Fernández, J. Luengo, F. Herrera, **A Study of Statistical Techniques and Performance Measures for Genetics-Based Machine Learning: Accuracy and Interpretability**. *Soft Computing 13:10 (2009) 959-977, doi:10.1007/s00500-008-0392-y*.

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

A tutorial:

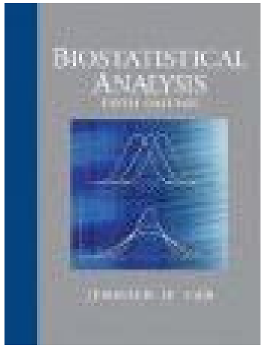


J. Derrac, S. García, D. Molina, F. Herrera.

A Practical Tutorial on the Use of Nonparametric Statistical Tests as a Methodology for Comparing Evolutionary and Swarm Intelligence Algorithms.

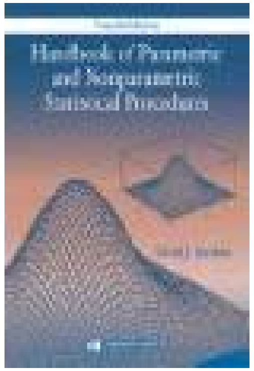
Swarm and Evolutionary Computation 1:1 (2011) 3-18.

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.

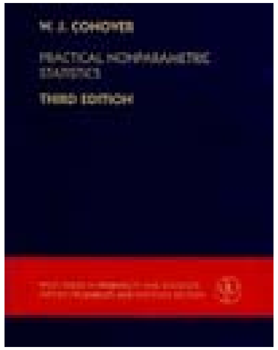
Books of interest and References

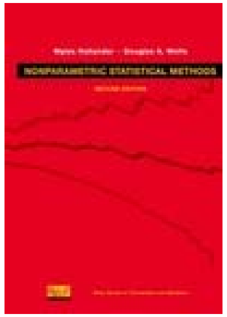
W.W. Daniel. Applied Nonparametric Statistics.

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

Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–803 (1988)

Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian J. Statist. 6, 65–70 (1979)

Iman, R.L., Davenport, J.M.: Approximations of the critical region of the Friedman statistic. Commun. Stat. 18, 571–595 (1980)

Shaffer, J.P.: Multiple hypothesis testing. Annu. Rev. Psychol. 46, 561–584 (1995)

Wright, S.P.: Adjusted *p-values for simultaneous inference. Biometrics 48, 1005–1013 (1992)*

K. Doksum, Robust procedures for some linear models with one observation per cell, Annals of Mathematical Statistics 38 (1967) 878–883.

O.J. Dunn, Multiple comparisons among means, Journal of the American Statistical Association 56 (1961) 52–64.

H. Finner, On a monotonicity problem in step-down multiple test procedures, Journal of the American Statistical Association 88 (1993) 920–923.

M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (1937) 674–701.

Books of interest and References

M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics 11 (1940) 86–92.

J.L. Hodges, E.L. Lehmann, Ranks methods for combination of independent experiments in analysis of variance, Annals of Mathematical Statistics 33 (1962) 482–497.

B.S. Holland, M.D. Copenhaver, An improved sequentially rejective Bonferroni test procedure, Biometrics 43 (1987) 417–423.

G. Hommel, A stagewise rejective multiple test procedure based on a modified Bonferroni test, Biometrika 75 (1988) 383–386.

J. Li, A two-step rejection procedure for testing multiple hypotheses, Journal of Statistical Planning and Inference 138 (2008) 1521–1527.

D. Quade, Using weighted rankings in the analysis of complete blocks with additive block effects, Journal of the American Statistical Association 74 (1979) 680–683.

A.L. Rhyne, R.G.D. Steel, Tables for a treatments versus control multiple comparisons sign test, Technometrics 7 (1965) 293–306.

D.M. Rom, A sequentially rejective test procedure based on a modified Bonferroni inequality, Biometrika 77 (1990) 663–665.

R.G.D. Steel, A multiple comparison sign test: treatments versus control, Journal of American Statistical Association 54 (1959) 767–775.

Books of interest and References

G. Bergmann and G. Hommel. Improvements of general multiple test procedures for redundant systems of hypotheses. In P. Bauer, G. Hommel, and E. Sonnemann, editors, *Multiple Hypotheses Testing, pages 100–115. Springer, Berlin, 1988.*

Y. Hochberg and D. Rom. Extensions of multiple testing procedures based on Simes' test. *Journal of Statistical Planning and Inference, 48:141–152, 1995.*

G. Hommel and G. Bernhard. A rapid algorithm and a computer program for multiple test procedures using procedures using logical structures of hypotheses. *Computer Methods and Programs in Biomedicine, 43:213–216, 1994.*

G. Hommel and G. Bernhard. Bonferroni procedures for logically related hypotheses. *Journal of Statistical Planning and Inference, 82:119–128, 1999.*

P. B. Nemenyi. *Distribution-free Multiple Comparisons. PhD thesis, Princeton University, 1963.*

J.P. Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81(395):826–831, 1986.*

R.J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73:* 751–754, 1986.

Statistical Analysis of Experiments in Data Mining and Computational Intelligence

OUTLINE

- **Introduction to Inferential Statistics**
- **Conditions for the safe use of parametric tests**
- **Basic non-parametric tests and case studies**
- **Advanced non-parametric tests and case studies:**
- **Lessons Learned**
- **Books of Interest and References**
- **Software**

Software

Software for conducting multiple comparisons tests with a control

http://sci2s.ugr.es/sicidm/controlTest.zip

Read data of results of k algorithms over N case problems in CSV format. The data can correspond to accuracy, AUC or any other performance measure.

Compute the rankings through the Friedman Aligned Ranks and Quade procedures of k algorithms over N case problems.

Compute the Friedman and Iman-Davenport, Friedman Aligned-Ranks and Quade Statistics corresponding to the input data.

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*

Software

Software for conducting multiple comparisons tests with a control

http://sci2s.ugr.es/sicidm/controlTest.zip

Show the tables with the set of hypotheses, unadjusted p-values for each comparison and adjusted level of significance for Bonferroni-Dunn, Holm and Hochberg, Hommel, Holland, Rom, Finner and Li procedures: **1 x n comparison**.

Show the table with adjusted p-values for the procedures 1 x n mentioned in the previous item.

Give a report detailing the rejected hypotheses considering the levels of significance α = 0.05 and α = 0.10.

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*

Software

Software for conducting all pairwise comparisons

http://sci2s.ugr.es/sicidm/multipleTest.zip

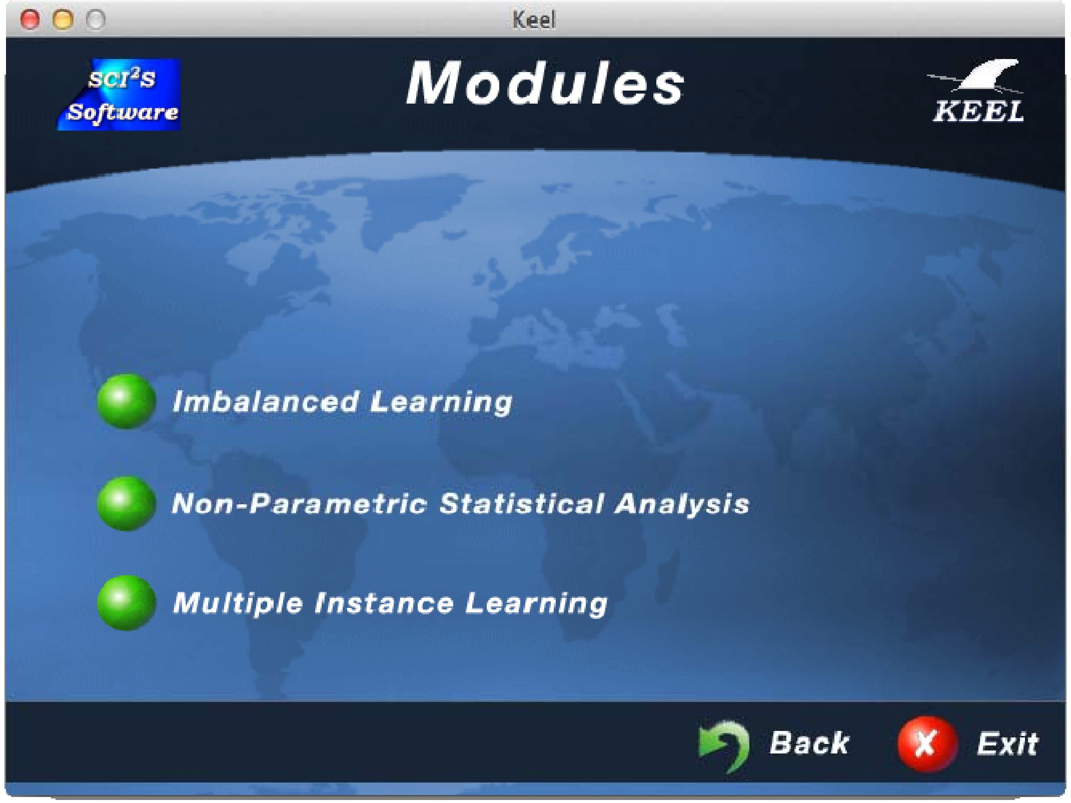
We offer a software developed in JAVA which calculates all the multiple comparisons procedures described in this talk and the JMLR paper.

It allows as input files in CSV format and obtains as output a LaTeX file with tabulated information about Friedman, Iman-Davenpor. Bonferroni-Dunn, Holm, Hochberg, Shaffer and Bergamnn-Hommel tests. It also computes and shows the adjusted p-values.

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*

Software

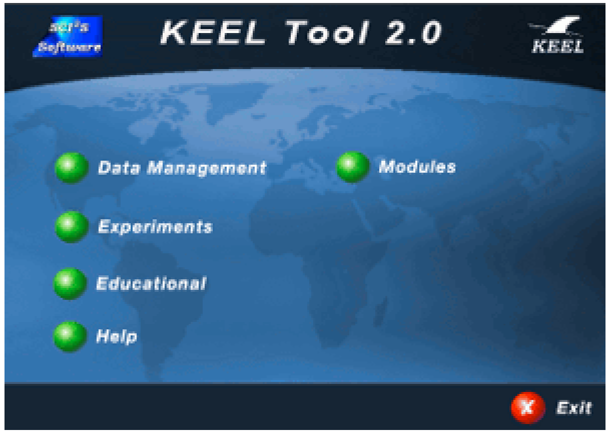
http://www.keel.es

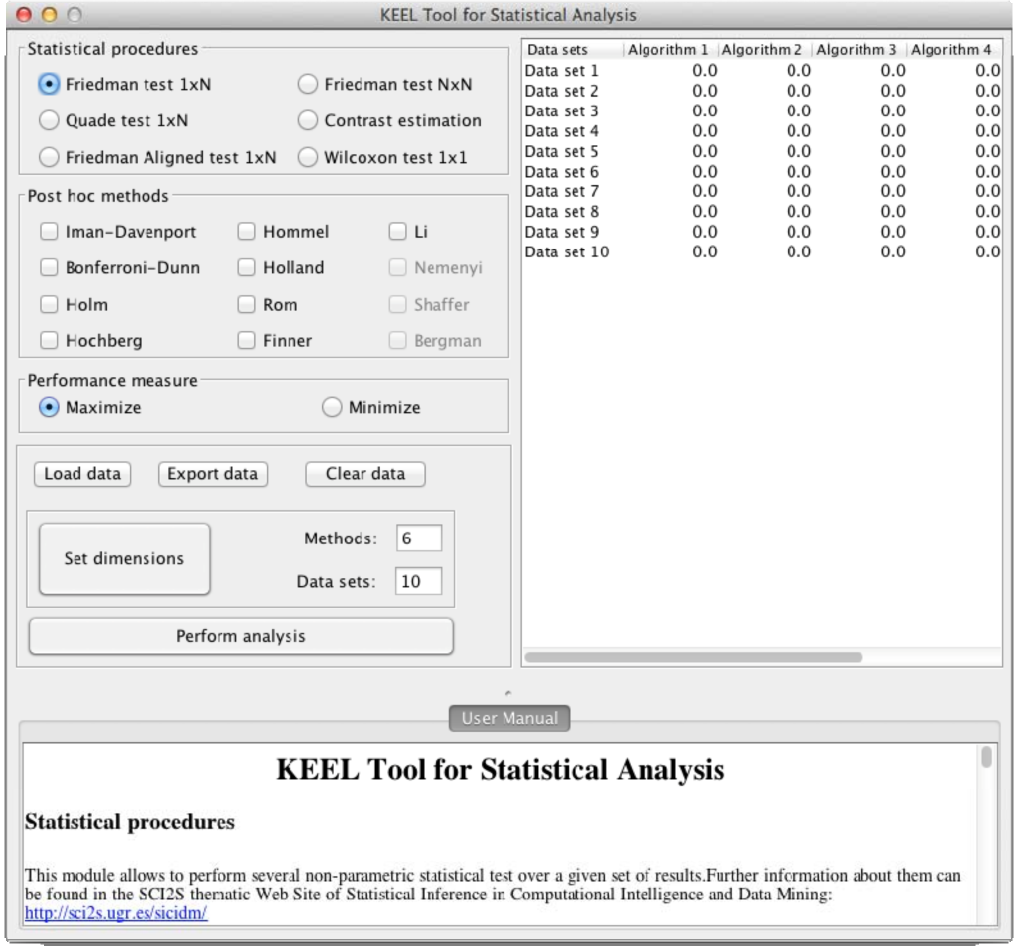

Keel
SCI²S Software
Modules
KEEL
Imbalanced Learning
Non-Parametric Statistical Analysis
Multiple Instance Learning
Back
Exit

J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera. KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 13:3 (2009) 307-318

Software

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
Data set 1	0.0	0.0	0.0	0.0
Data set 2	0.0	0.0	0.0	0.0
Data set 3	0.0	0.0	0.0	0.0
Data set 4	0.0	0.0	0.0	0.0
Data set 5	0.0	0.0	0.0	0.0
Data set 6	0.0	0.0	0.0	0.0
Data set 7	0.0	0.0	0.0	0.0
Data set 8	0.0	0.0	0.0	0.0
Data set 9	0.0	0.0	0.0	0.0
Data set 10	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/

J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera. KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 13:3 (2009) 307-318

Statistical Analysis of Experiments

OUTLINE

- Introduction to Inferential Statistics
- Conditions for the safe use of parametric tests
- Basic non-parametric tests and case studies
- Advanced non-parametric tests and case studies:
- Lessons Learned
- Books of Interest and References
- Software

Statistical Analysis of Experiments