UNIVERSITÉ DU LUXEMBOURG

Assessing Statistical Confidence to Your Experimental Comparisons

Bernabé Dorronsoro

Based on:
Takagi, Statistical Tests for Computational Intelligence Research and Human Subjective Tests
García, Statistical Analysis of Experiments in Data Mining and Computational Intelligence

Outline

- Motivations
- Pairwise comparisons
  - Parametric
  - Non-parametric
- Multiple comparisons
  - Unpaired
    - Parametric
    - Non-parametric
  - Paired
    - Parametric
    - Non-parametric
- Summary

Motivations

- Deciding when an algorithm is better than other one is not trivial
- Just comparing averages is not scientifically rigorous

	Algorithm 1	Algorithm 2	
Average Results	22.30	20.4	$f^*(x) = 0.0$
Data	21.7	20.6	
	26.5	22.8	
	19.8	17.7	
	22.4	21.5	
	21.1	19.4	

77.66% probability that they are different

You cannot show the superiority of your method without statistical tests

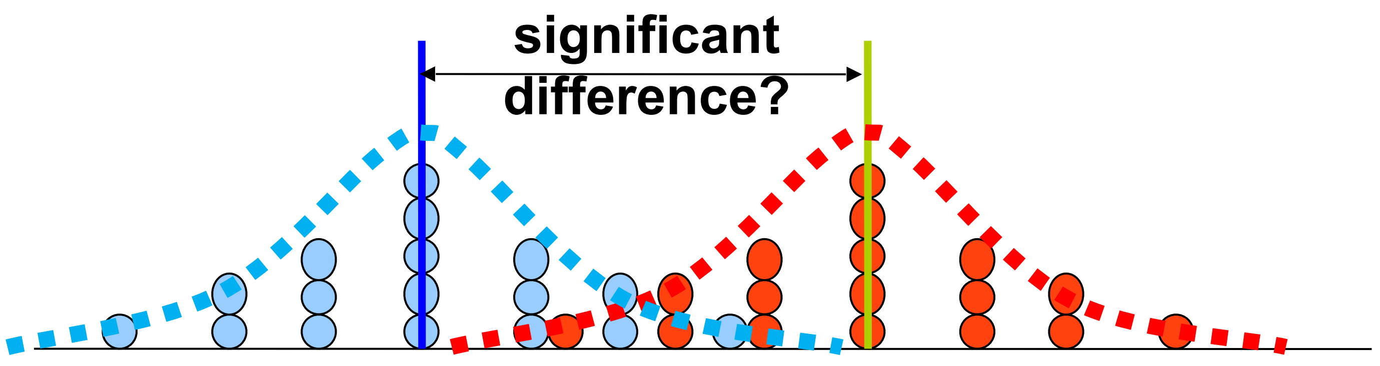
How to show significance?

Statistical tests will tell you if these data distributions are similar or not, with the desired degree of confidence (95%, 99%, …)

Pairwise Comparisons

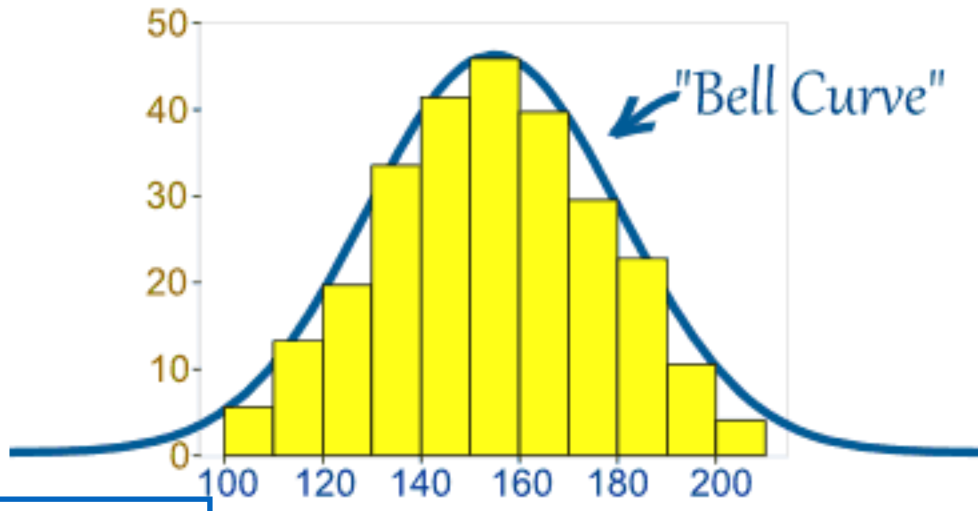
Pairwise Statistical Tests

- Stochastic processes provide different results in every experiment
  - We need to perform many tests
  - Need of statistical tests: median or mean do not reflect the best algorithm



- Statistical test
  - Assume the null hypothesis:
    - The two distributions are the same
  - Analyze the distribution of the data provided by the algorithms
    - They always provide a p-value
  - p-value: The smallest level of significance that results in the rejection of the null hypothesis
    - If p-value is less than 0.05 => the hypothesis is rejected with 95% confidence
    - If p-value is less than 0.01 => the hypothesis is rejected with 99% confidence
- Number of samples
  - Depends
    - Data distribution
    - Test to apply
  - Minimum of 30
  - Recommended 100

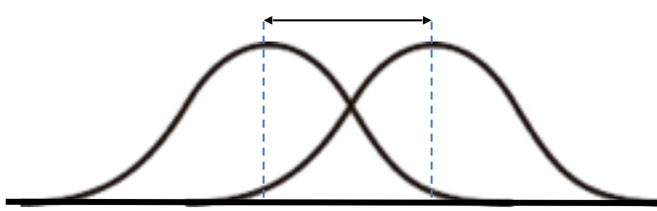
Pairwise Parametric Test



- Student's test (*t*-test)
  - Requirements:
    - Normally distributed data (Shapiro-Wilks)

Matlab (download swtest.m)
[stat, pval, H] = swtest(data)

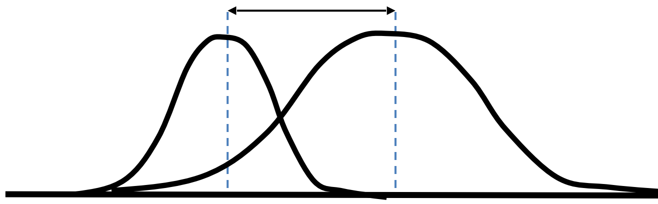
R
shapiro.test(data)



- Equality of variances (F-test)

Matlab
[H, pval] = vartest(data1, data2)

R
var.test(data1, data2)



- If p-value≥0.05 in **all cases**, then we can apply *t*-test

Matlab
[H, pval] = ttest(data1, data2)

R
t.test(data1, data2)

p-value: Confidence interval

≥ 0.05 Data follow the same distribution with 95% confidence

≤ 0.05 Data follow different distributions with 95% confidence

Pairwise Non-Parametric Test

- Wilcoxon signed-rank test
  - Does not require data to be normally distributed
    - Alternative to *t*-test for not normally distributed populations

Matlab

[H, pval] = signrank(data1, data2)

R

wilcox.test(data1, data2)

p-value: Confidence interval

$\geq$ 0.05 Data follow the same distribution with 95% confidence

$\leq$ 0.05 Data follow different distributions with 95% confidence

Pairwise Tests

UNPAIRED

Repeated measurements of a single sample

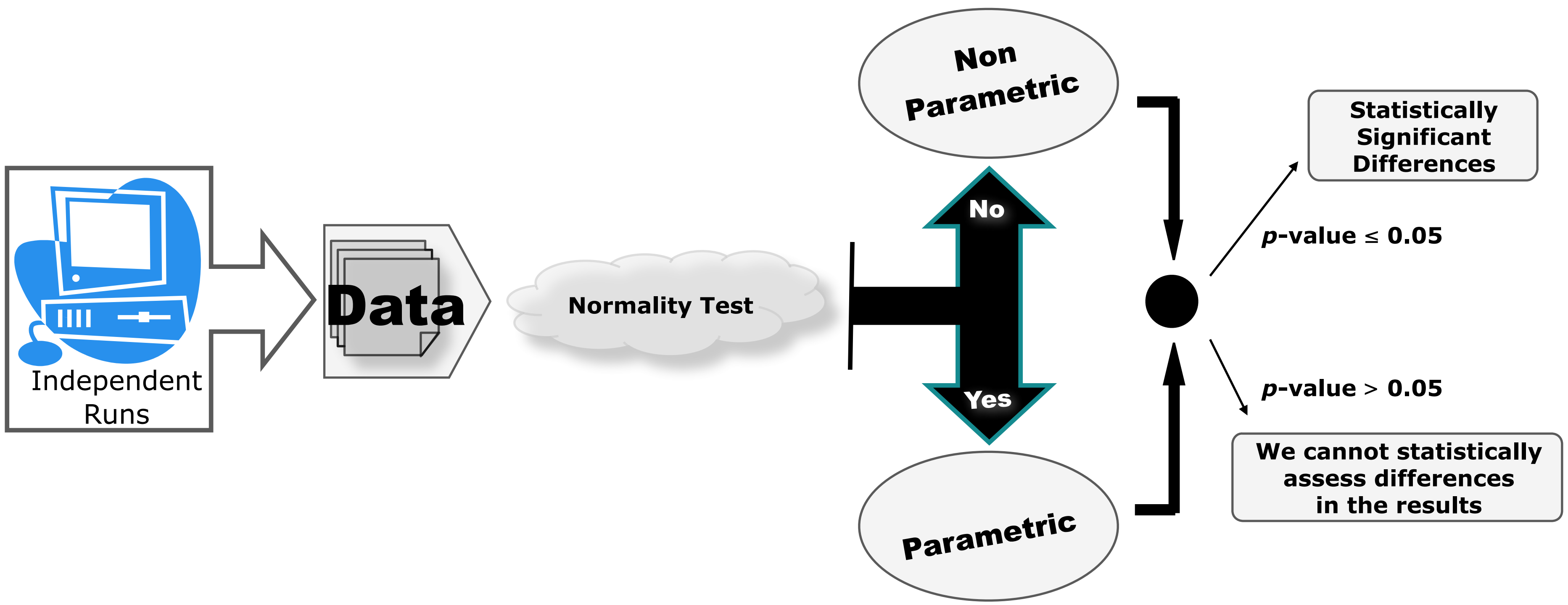
Alg. 1	Alg. 2
25.3	76.0
55.5	81.9
45.0	76.2
58.0	63.5
51.6	83.2
65.5	96.0
42.7	44.4
34.6	66.3
54.5	77.4
79.3	79.9

PAIRED

Matched or related samples

	Alg. 1	Alg. 2
aud	25.3	76.0
aus	55.5	81.9
bal	45.0	76.2
bpa	58.0	63.5
bps	51.6	83.2
bre	65.5	96.0
cmc	42.7	44.4
gls	34.6	66.3
h-c	54.5	77.4
hep	79.3	79.9

Process for Pairwise Statistical Tests



Unpaired Multiple Comparisons

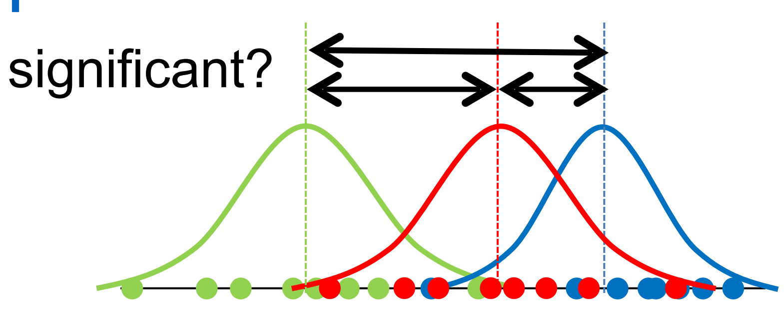
Unpaired Multiple Comparisons

Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5
25.3	76.0	79.0	83.3	76.0
55.5	81.9	85.2	81.9	43.7
45.0	76.2	78.5	65.8	96.1
58.0	63.5	65.8	79.0	96.7
51.6	83.2	80.1	95.3	76.2
65.5	96.0	95.4	49.8	88.4
42.7	44.4	52.1	69.0	86.3
34.6	66.3	65.8	77.9	72.2
54.5	77.4	73.6	80.0	95.4
79.3	79.9	78.9	95.3	87.6

- How to compare multiple algorithms on the same problem?
- Pairwise comparisons on every pair
  - p-values in a pairwise comparison is independent from another one
  - Extract a conclusion involving more than one pairwise comparison

Accumulated error coming from the combination of the pairwise comparisons

For 95% confidence and 10 algorithms, the probability of making one or more errors is 0.37



- We need statistical tests for multiple comparisons

Unpaired Mult. Comparisons Parametric Test

- ANOVA test
  - Requires data to be normally distributed
    - Normally distributed data (Shapiro-Wilks)
    - Equality of variances (F-test)

Matlab

[pval, anovatab, stats] = anova1([Alg1,Alg2,Alg3])

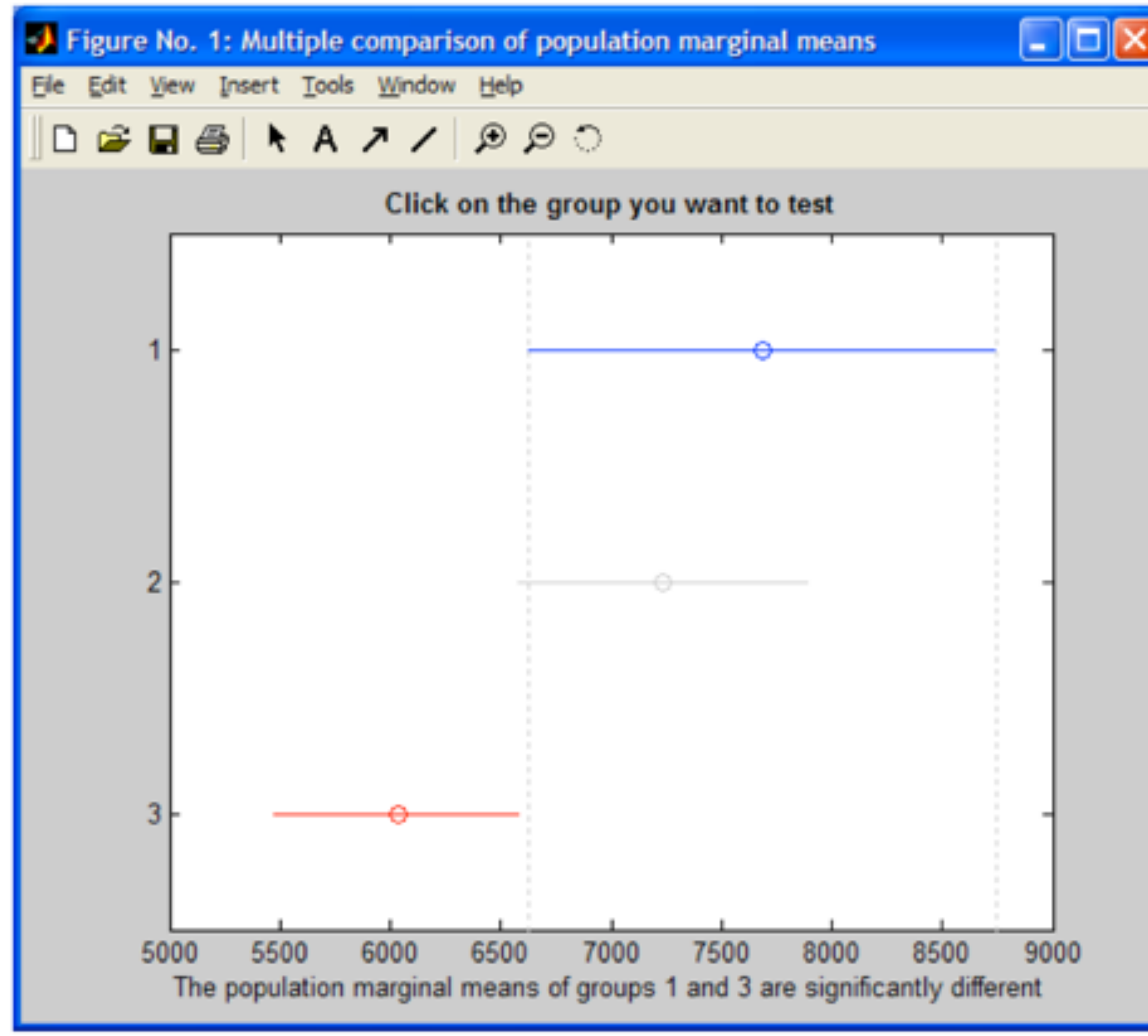
3 x nb samples

One column

Matlab

multcompare(stats)

Tukey method



p-value: Confidence interval

≤ 0.05 Statistically significant differences with 95% confidence

BETWEEN ANY TWO ALGORITHMS!

Unpaired Mult. Comparisons Parametric Test

- ANOVA test
  - Requires
    - Normally distributed data (Shapiro-Wilks)
    - Equality of variances (F-test)

R

res = aov(Solution~Algorithm, test1)

test1 <- read.table(“results.dat”,col.names=c(“Solution”,”Algorithm”))

```
Solution Algorithm
0.491758 alg1
0.15651 alg1
...
2.40222e-07 alg3
```

Results file format

```
0.491758 alg1
0.15651 alg1
0.000977229 alg1
...
0.319085 alg1
0.11174 alg1
2.57778e-07 alg2
2.57778e-07 alg2
2.57778e-07 alg2
...
2.57778e-07 alg2
2.57778e-07 alg2
0.00037624 alg3
2.40222e-07 alg3
0.105578 alg3
...
2.40222e-07 alg3
2.40222e-07 alg3
```

R

summary(res)

```
            Df Sum Sq Mean Sq F value    Pr(>F)
Algorithm    2 2.8244  1.4122  20.301 5.812e-08 ***
Residuals   87 6.0520  0.0696
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

p-value: Confidence interval

$\leq$ 0.05 Statistically significant differences with 95% confidence

BETWEEN ANY TWO ALGORITHMS!

Unpaired Mult. Comparisons Parametric Test

- ANOVA test
  - Apply the Tukey method to check the differences between the algorithms

R

resTukey <- TukeyHSD(res,"Algorithm")

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Solution ~ Algorithm, data = test1)

$Algorithm
                            diff         lwr        upr
algorithm2-algorithm1 -0.4175242 -0.57990596 -0.2551424
algorithm3-algorithm1 -0.3111089 -0.47349067 -0.1487271
algorithm3-algorithm2  0.1064153 -0.05596652  0.2687971
```

Significant

Not Significant

Difference between two algorithms is significant if

Interval [lwr,upr] does not contain value 0

Unpaired Mult. Comp. Non-Parametric Test

- Kruskal-Wallis test
  - Does not require data to be normally distributed
    - Alternative to ANOVA for not normally distributed populations

Matlab

[pval, kwtab, stats] = kruskalwallis([Alg1,Alg2,Alg3])

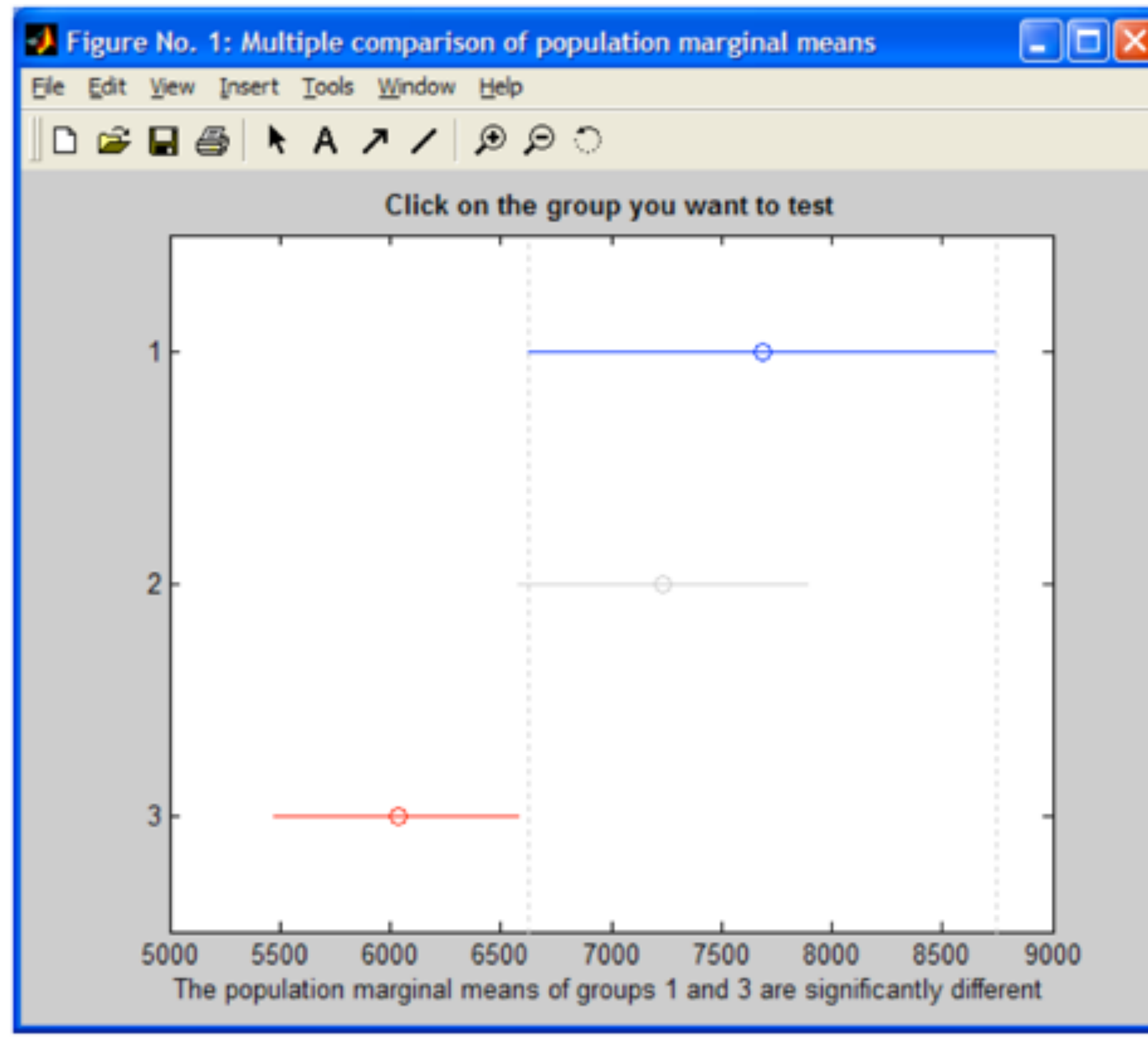
3 x nb samples

One column

Matlab

multcompare(stats)

Tukey method



p-value: Confidence interval

≤ 0.05 Statistically significant differences with 95% confidence

BETWEEN ANY TWO ALGORITHMS!

Unpaired Mult. Comp. Non-Parametric Test

- Kruskal-Wallis test
  - Does not require data to be normally distributed
    - Alternative to ANOVA for not normally distributed populations

R

res = kruskal.test(Solution~Algorithm, test1)

test1 <- read.table("results.dat",col.names=c("Solution","Algorithm"))

```
Solution Algorithm
0.491758 alg1
0.15651 alg1
...
2.40222e-07 alg3
```

Results file format

```
0.491758 alg1
0.15651 alg1
0.000977229 alg1
...
0.319085 alg1
0.11174 alg1
2.57778e-07 alg2
2.57778e-07 alg2
2.57778e-07 alg2
...
2.57778e-07 alg2
2.57778e-07 alg2
0.00037624 alg3
2.40222e-07 alg3
0.105578 alg3
...
2.40222e-07 alg3
2.40222e-07 alg3
```

p-value: Confidence interval

$\leq$ 0.05 Statistically significant differences with 95% confidence

BETWEEN ANY TWO ALGORITHMS!

Paired Multiple Comparisons

Paired Multiple Comparisons

- Example for classification algorithms

  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**

Paired Multiple Comparisons

- Alg. 4
  - Best for 8 problems
  - Avg. 78.0
- Alg. 2
  - Best for 4 problems
  - Avg. 80.0
- Which one is better?
- Is any of them better than the others?

	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**

Paired Multiple Comparisons

- We need statistical tests for paired multiple comparisons

	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**

Friedman Rank

- Friedman Rank
  - Checks if there is statistical significance in the behavior of the algorithms
  - Ranks them, from better (lower rank) to worse (higher rank)

MULTIPLETEST package

java Friedman data.csv > tests.tex

Algorithm	Rank
Alg.7	3.0666666666666678
Alg.3	3.1166666666666667
Alg.5	3.4833333333333334
Alg.6	3.4833333333333334
Alg.4	3.9166666666666668
Alg.2	4.0333333333333332
Alg.1	6.9000000000000001

p-value = 3.6831E-11

Pairwise Paired Multiple Comparisons

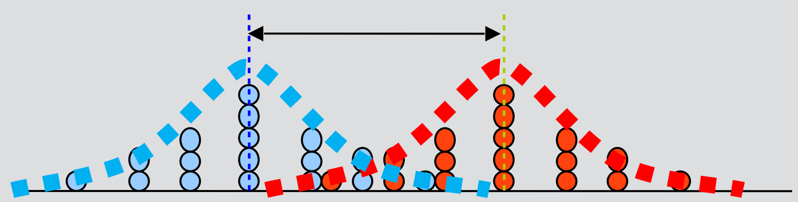
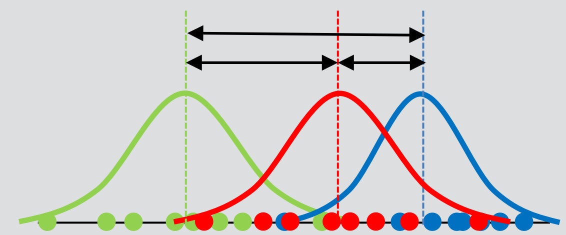
MULTIPLETEST package

java Friedman data.csv > tests.tex

Table 7: Adjusted p-values

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	Alg.1 vs .Alg.7	6.3057851282898035E-12	1.3242148769408586E-10	1.3242148769408586E-10	1.3242148769408586E-10	1.3242148769408586E-10
2	Alg.1 vs .Alg.3	1.1776893734796516E-11	2.4731476843072686E-10	2.355378746959303E-10	1.7665340602194773E-10	1.7665340602194773E-10
3	Alg.1 vs .Alg.6	9.037280264360894E-10	1.8978288555157876E-8	1.71708325022857E-8	1.3555920396541341E-8	9.941008290796983E-9
4	Alg.1 vs .Alg.5	9.03728026436099E-10	1.8978288555158078E-8	1.71708325022857E-8	1.3555920396541485E-8	9.941008290797088E-9
5	Alg.1 vs .Alg.4	8.861368315025284E-8	1.8608873461553095E-6	1.5064326135542983E-6	1.3292052472537925E-6	9.747505146527812E-7
6	Alg.1 vs .Alg.2	2.754953845326433E-7	5.7854030751855095E-6	4.407926152522293E-6	4.1324307679896495E-6	3.0304492298590765E-6
7	Alg.2 vs .Alg.7	0.08308118696647453	1.744704926295965	1.246217804497118	1.246217804497118	1.246217804497118
8	Alg.2 vs .Alg.3	0.1002920667156204	2.1061334010280284	1.4040889340186855	1.246217804497118	1.246217804497118
9	Alg.4 vs .Alg.7	0.12752957690954794	2.678121115100507	1.6578844998241233	1.4028253460050273	1.2752957690954794
10	Alg.3 vs .Alg.4	0.15149399240421982	3.1813738404886163	1.817927908850638	1.666433916446418	1.2752957690954794
11	Alg.2 vs .Alg.6	0.32410190296180597	6.806139962197925	3.5651209325798656	3.5651209325798656	2.2687133207326418
12	Alg.2 vs .Alg.5	0.3241019029618067	6.806139962197941	3.5651209325798656	3.5651209325798656	2.2687133207326418
13	Alg.4 vs .Alg.6	0.43721859962502607	9.181590592125547	3.9349673966252348	3.9349673966252348	2.2687133207326418
14	Alg.4 vs .Alg.5	0.437218599625027	9.181590592125568	3.9349673966252348	3.9349673966252348	2.2687133207326418
15	Alg.5 vs .Alg.7	0.45505276752074586	9.556108117935663	3.9349673966252348	3.9349673966252348	3.185369372645221
16	Alg.6 vs .Alg.7	0.45505276752074675	9.556108117935683	3.9349673966252348	3.9349673966252348	3.185369372645221
17	Alg.3 vs .Alg.5	0.5109393498748492	10.729726347371834	3.9349673966252348	3.9349673966252348	3.185369372645221
18	Alg.3 vs .Alg.6	0.5109393498748502	10.729726347371855	3.9349673966252348	3.9349673966252348	3.185369372645221
19	Alg.2 vs .Alg.4	0.8343194288581424	17.52070800602099	3.9349673966252348	3.9349673966252348	3.185369372645221
20	Alg.3 vs .Alg.7	0.9285715917804449	19.500003427389345	3.9349673966252348	3.9349673966252348	3.185369372645221
21	Alg.5 vs .Alg.6	0.9999999999999987	20.99999999999997	3.9349673966252348	3.9349673966252348	3.185369372645221

Summary

		PAIRWISE COMPARISON	MULTIPLE COMPARISON
Data Distribution			
Parametric Test (normality & homoscedasticity)	unpaired (independent)	unpaired t-test	ANOVA
	paired (related)	paired t-test	two-way ANOVA
Non-parametric Test	unpaired (independent)	unpaired Wilcoxon signed-ranks test	Kruskal-Wallis test
	paired (related)	paired Wilcoxon signed-ranks test	Friedman test Holm Shaffer Bergmann-Hommel

References

- http://sci2s.ugr.es/sicidm/
  - Software: CONTROLTEST & MULTIPLETEST
- Cites for your papers

  - S. García, D. Molina, M. Lozano, and F. Herrera, *A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'05 special session on real parameter optimization*, Journal of Heuristics, 15:617–644, 2009.
  - D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, CRC Press, Boca Raton, FL, 2003.
  - J. H. Zar, *Biostatistical Analysis*, Prentice Hall, Upper Saddle River, NJ, 1999.