CSE/ECE 848 Introduction to Evolutionary Computation

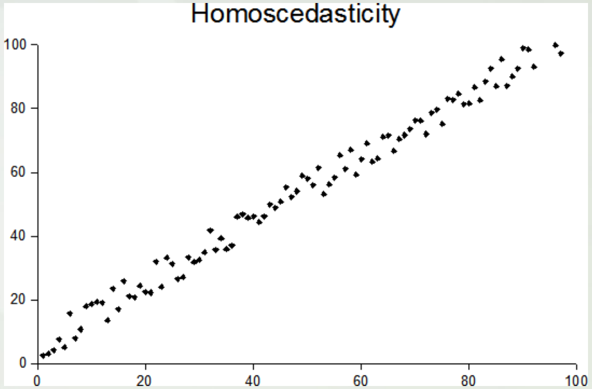
Module 3 - Lecture 14 - Part 3

Comparison of EC Methods: Statistical Tests & Methods

Wolfgang Banzhaf, CSE
John R. Koza Chair in Genetic Programming

Outcomes reported - Now what?

- Suppose we have now reported medians or other performance measures for number of algorithms - Now what can we conclude?
- Statistical analysis and tests are the tools for deciding whether there are clear (statistically significant) differences between the different algorithms
- Parametric vs non-parametric statistics tests have different underlying assumptions
  - Parametric: Independence, normality, homoscedasticity (“homogeneity of variance”)
  - Non-parametric: These assumptions are not required
- Descriptive vs inferential statistics
  - Descriptive St.: Describes data
  - Inferential St.: Allows to make predictions
    - Hypothesis testing
    - Confidence Intervals



Statistical Tests

- Without hypothesis - no test!
- Null hypothesis - H_0 There are no effects/differences
- Alternative hypothesis - H_1 There are significant differences between the different algorithms under comparison
- Significance level α can be defined, or one calculated the p-value, the probability of obtaining results at least as extreme as the observed results, assuming H_0 is correct
- Whereas α allows a Boolean decision, the p-value gives a measure of significance of the results - the smaller the p-value, the stronger the evidence against H_0

The data[1]

n=25; k=9

Table 1
Average error obtained in the 25 benchmark functions.

Function	PSO	IPOP-CMA-ES	CHC	SSGA	SS-BLX	SS-Arit	DE-Bin	DE-Exp	SaDE
F1	$1.234 \cdot 10^{-4}$	0.000	2.464	$8.420 \cdot 10^{-9}$	$3.402 \cdot 10$	1.064	$7.716 \cdot 10^{-9}$	$8.260 \cdot 10^{-9}$	$8.416 \cdot 10^{-9}$
F2	$2.595 \cdot 10^{-2}$	0.000	$1.180 \cdot 10^{2}$	$8.719 \cdot 10^{-5}$	1.730	5.282	$8.342 \cdot 10^{-9}$	$8.181 \cdot 10^{-9}$	$8.208 \cdot 10^{-9}$
F3	$5.174 \cdot 10^{4}$	0.000	$2.699 \cdot 10^{5}$	$7.948 \cdot 10^{4}$	$1.844 \cdot 10^{5}$	$2.535 \cdot 10^{5}$	$4.233 \cdot 10$	$9.935 \cdot 10$	$6.560 \cdot 10^{3}$
F4	2.488	$2.932 \cdot 10^{3}$	$9.190 \cdot 10$	$2.585 \cdot 10^{-3}$	6.228	5.755	$7.686 \cdot 10^{-9}$	$8.350 \cdot 10^{-9}$	$8.087 \cdot 10^{-9}$
F5	$4.095 \cdot 10^{2}$	$8.104 \cdot 10^{-10}$	$2.641 \cdot 10^{2}$	$1.343 \cdot 10^{2}$	2.185	$1.443 \cdot 10$	$8.608 \cdot 10^{-9}$	$8.514 \cdot 10^{-9}$	$8.640 \cdot 10^{-9}$
F6	$7.310 \cdot 10^{2}$	0.000	$1.416 \cdot 10^{6}$	6.171	$1.145 \cdot 10^{2}$	$4.945 \cdot 10^{2}$	$7.956 \cdot 10^{-9}$	$8.391 \cdot 10^{-9}$	$1.612 \cdot 10^{-2}$
F7	$2.678 \cdot 10$	$1.267 \cdot 10^{3}$	$1.269 \cdot 10^{3}$	$1.271 \cdot 10^{3}$	$1.966 \cdot 10^{3}$	$1.908 \cdot 10^{3}$	$1.266 \cdot 10^{3}$	$1.265 \cdot 10^{3}$	$1.263 \cdot 10^{3}$
F8	$2.043 \cdot 10$	$2.001 \cdot 10$	$2.034 \cdot 10$	$2.037 \cdot 10$	$2.035 \cdot 10$	$2.036 \cdot 10$	$2.033 \cdot 10$	$2.038 \cdot 10$	$2.032 \cdot 10$
F9	$1.438 \cdot 10$	$2.841 \cdot 10$	5.886	$7.286 \cdot 10^{-9}$	4.195	5.960	4.546	$8.151 \cdot 10^{-9}$	$8.330 \cdot 10^{-9}$
F10	$1.404 \cdot 10$	$2.327 \cdot 10$	7.123	$1.712 \cdot 10$	$1.239 \cdot 10$	$2.179 \cdot 10$	$1.228 \cdot 10$	$1.118 \cdot 10$	$1.548 \cdot 10$
F11	5.590	1.343	1.599	3.255	2.929	2.858	2.434	2.067	6.796
F12	$6.362 \cdot 10^{2}$	$2.127 \cdot 10^{2}$	$7.062 \cdot 10^{2}$	$2.794 \cdot 10^{2}$	$1.506 \cdot 10^{2}$	$2.411 \cdot 10^{2}$	$1.061 \cdot 10^{2}$	$6.309 \cdot 10$	$5.634 \cdot 10$
F13	1.503	1.134	$8.297 \cdot 10$	$6.713 \cdot 10$	$3.245 \cdot 10$	$5.479 \cdot 10$	1.573	$6.403 \cdot 10$	$7.070 \cdot 10$
F14	3.304	3.775	2.073	2.264	2.796	2.970	3.073	3.158	3.415
F15	$3.398 \cdot 10^{2}$	$1.934 \cdot 10^{2}$	$2.751 \cdot 10^{2}$	$2.920 \cdot 10^{2}$	$1.136 \cdot 10^{2}$	$1.288 \cdot 10^{2}$	$3.722 \cdot 10^{2}$	$2.940 \cdot 10^{2}$	$8.423 \cdot 10$
F16	$1.333 \cdot 10^{2}$	$1.170 \cdot 10^{2}$	$9.729 \cdot 10$	$1.053 \cdot 10^{2}$	$1.041 \cdot 10^{2}$	$1.134 \cdot 10^{2}$	$1.117 \cdot 10^{2}$	$1.125 \cdot 10^{2}$	$1.227 \cdot 10^{2}$
F17	$1.497 \cdot 10^{2}$	$3.389 \cdot 10^{2}$	$1.045 \cdot 10^{2}$	$1.185 \cdot 10^{2}$	$1.183 \cdot 10^{2}$	$1.279 \cdot 10^{2}$	$1.421 \cdot 10^{2}$	$1.312 \cdot 10^{2}$	$1.387 \cdot 10^{2}$
F18	$8.512 \cdot 10^{2}$	$5.570 \cdot 10^{2}$	$8.799 \cdot 10^{2}$	$8.063 \cdot 10^{2}$	$7.668 \cdot 10^{2}$	$6.578 \cdot 10^{2}$	$5.097 \cdot 10^{2}$	$4.482 \cdot 10^{2}$	$5.320 \cdot 10^{2}$
F19	$8.497 \cdot 10^{2}$	$5.292 \cdot 10^{2}$	$8.798 \cdot 10^{2}$	$8.899 \cdot 10^{2}$	$7.555 \cdot 10^{2}$	$7.010 \cdot 10^{2}$	$5.012 \cdot 10^{2}$	$4.341 \cdot 10^{2}$	$5.195 \cdot 10^{2}$
F20	$8.509 \cdot 10^{2}$	$5.264 \cdot 10^{2}$	$8.960 \cdot 10^{2}$	$8.893 \cdot 10^{2}$	$7.463 \cdot 10^{2}$	$6.411 \cdot 10^{2}$	$4.928 \cdot 10^{2}$	$4.188 \cdot 10^{2}$	$4.767 \cdot 10^{2}$
F21	$9.138 \cdot 10^{2}$	$4.420 \cdot 10^{2}$	$8.158 \cdot 10^{2}$	$8.522 \cdot 10^{2}$	$4.851 \cdot 10^{2}$	$5.005 \cdot 10^{2}$	$5.240 \cdot 10^{2}$	$5.420 \cdot 10^{2}$	$5.140 \cdot 10^{2}$
F22	$8.071 \cdot 10^{2}$	$7.647 \cdot 10^{2}$	$7.742 \cdot 10^{2}$	$7.519 \cdot 10^{2}$	$6.828 \cdot 10^{2}$	$6.941 \cdot 10^{2}$	$7.715 \cdot 10^{2}$	$7.720 \cdot 10^{2}$	$7.655 \cdot 10^{2}$
F23	$1.028 \cdot 10^{3}$	$8.539 \cdot 10^{2}$	$1.075 \cdot 10^{3}$	$1.004 \cdot 10^{3}$	$5.740 \cdot 10^{2}$	$5.828 \cdot 10^{2}$	$6.337 \cdot 10^{2}$	$5.824 \cdot 10^{2}$	$6.509 \cdot 10^{2}$
F24	$4.120 \cdot 10^{2}$	$6.101 \cdot 10^{2}$	$2.959 \cdot 10^{2}$	$2.360 \cdot 10^{2}$	$2.513 \cdot 10^{2}$	$2.011 \cdot 10^{2}$	$2.060 \cdot 10^{2}$	$2.020 \cdot 10^{2}$	$2.000 \cdot 10^{2}$
F25	$5.099 \cdot 10^{2}$	$1.818 \cdot 10^{3}$	$1.764 \cdot 10^{3}$	$1.747 \cdot 10^{3}$	$1.794 \cdot 10^{3}$	$1.804 \cdot 10^{3}$	$1.744 \cdot 10^{3}$	$1.742 \cdot 10^{3}$	$1.738 \cdot 10^{3}$

[1] Adopted from Derrac et al, “A practical tutorial on the use of non-parametric statistical tests for comparing evolutionary and swarm intelligence algorithms”, Swarm and Evolutionary Computation, 1 (2011) 3-18

Types of Comparisons

- Pairwise comparisons
  - Sign test
  - Wilcoxon test
- Multiple comparison (1 vs N)
  - Multiple sign test
  - Friedman test
  - …
- Multiple comparison (N vs N)
  - Friedman test

Sign Test

- Simple and popular: Count the number of times an algorithm is the winner.
- Under the Null hypothesis, both algorithms should win n/2 times
- Number is distributed in a binomial distribution which allows to apply the z-test: If number of wins is larger than $n/2 + 1.96 \cdot \sqrt{n}/2$ the algorithm is significantly better with p-value p< 0.05

Table 4
Critical values for the two-tailed sign test at $\alpha = 0.05$ and $\alpha = 0.1$. An algorithm is significantly better than another if it performs better on at least the cases presented in each row.

#Cases	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$\alpha = 0.05$	5	6	7	7	8	9	9	10	10	11	12	12	13	13	14	15	15	16	17	18	18
$\alpha = 0.1$	5	6	6	7	7	8	9	9	10	10	11	12	12	13	13	14	14	15	16	16	17

Table 5
Example of Sign test for pairwise comparisons. SaDE shows a significant improvement over PSO, CHC, and SSGA, with a level of significance $\alpha = 0.05$, and over SS-Arit, with a level of significance $\alpha = 0.1$.

SaDE	PSO	IPOP-CMA-ES	CHC	SSGA	SS-BLX	SS-Arit	DE-Bin	DE-Exp
Wins (+)	20	15	20	18	16	17	13	9
Loses (−)	5	10	5	7	9	8	12	16
Detected differences	$\alpha = 0.05$	–	$\alpha = 0.05$	$\alpha = 0.05$	–	$\alpha = 0.1$	–	–

Wilcoxon Test

- Signed rank test for answering, whether two samples represent two different populations
- Let d_i be the difference in performance score between two algorithms on problem i out of n
- Differences are ranked according to their absolute values. In case of ties, use the average rank
- We calculate positive and negative ranking scores …

$$R^+ = \sum_{d_i>0} \mathrm{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \mathrm{rank}(d_i)$$

$$R^- = \sum_{d_i<0} \mathrm{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \mathrm{rank}(d_i)$$

- and associated p-values

Wilcoxon Test II

- with the following result:

Table 6
Wilcoxon signed ranks test results. SaDE shows an improvement over PSO, CHC, and SSGA, with a level of significance $\alpha = 0.01$, over IPOP-CMA-ES and SS-Arit, with $\alpha = 0.05$, and over SS-BLX, with $\alpha = 0.1$.

Comparison	R^+	R^-	*p*-value	Comparison	R^+	R^-	*p*-value
SaDE versus PSO	261	64	0.00673	SaDE versus SS-BLX	232	93	0.06262
SaDE versus IPOP-CMA-ES	239	86	0.03934	SaDE versus SS-Arit	243	82	0.02958
SaDE versus CHC	287	38	0.00038	SaDE versus DE-Bin	176	149	>0.2
SaDE versus SSGA	260	65	0.00737	SaDE versus DE-Exp	119	206	>0.2

- which means: SaDE is significantly better than
  - PSO, CHC, SSGA with level of significance $\alpha = 0.01$
  - IPOP-CMA-ES, SS-Arit with $\alpha = 0.05$
  - SS-BLX with $\alpha = 0.1$

Friedman Test

- Multiple comparison test, asking the following question: In a set of k>=2 samples, do at least 2 of the samples represent populations with different median values?
- Null hypothesis: Equality of medians
- Calculation:
  - For each problem i rank values from 1 (best) to k (worst) r_i^j
  - For each algorithm j, average the ranks obtained for all problems $R^j = 1/n \, \Sigma_i \, r_i^j$
  - Friedman statistic

$$F_f = \frac{12n}{k(k+1)}\left[\Sigma_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

Friedman Test II

- On a toy example:

Table 7
Error rates achieved (Example 4).

Error	A	B	C	D
P1	2.711	3.147	2.515	2.612
P2	7.832	9.828	7.832	7.921
P3	0.012	0.532	0.122	0.005
P4	3.431	4.111	3.401	3.401

Table 8
Friedman ranks (Example 4).

Friedman	A	B	C	D
P1	3	4	1	2
P2	1.5	4	1.5	3
P3	2	4	3	1
P4	3	4	1.5	1.5
Average	2.375	4	1.250	1.875

Friedman Test III

- and for our algorithms:
- DE-Exp comes out best

- p-values can be calculated from the statistics
- and suggest strongly, that there are significant differences among the algorithms considered

Algorithms	Friedman
PSO	7
IPOP-CMA-ES	4.84
CHC	6.28
SSGA	5.5
SS-BLX	4.64
SS-Arit	5.4
DE-Bin	4
DE-Exp	3.5
SaDE	3.84
Statistic	35.99733
p-value	0.000018

- There are other tests, like the Quade test, but we are not going to discuss them here.

Post-hoc Procedures

- Disadvantage of Friedman (et al.) tests: They only detect that there IS a difference, but they cannot pinpoint which of the many algorithms compared differ significantly.
- To that end, a family of comparisons can be defined
  - Using k-1 hypotheses for comparison with a control method (k=1)
  - Using k*(k-1)/2 hypotheses for comparison all against all
- Then we order according to p-value (surest), from lowest to highest to get a picture

p-value calculation

- The p-value for a member of the family can be obtained by converting the rankings R_i and R_j of algorithms i and j into a z- score:

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

- The z-score can be translated into an (un-adjusted) p-value

- This p-value results from a one-to-one comparison, and needs to be corrected to say something for multiple tests
  - Bonferroni adjustment: Multiply each p-value by k-1

$$\text{Bonferroni } APV_i : \min\{v, 1\}, \text{ where } v = (k-1)p_i.$$

  - Other procedures: Holm & Hochberg

Example with DE-Exp as Control

- Result: No statistical difference between the last three algorithms and the control

Friedman	Unadjusted	Bonferroni
PSO	0.000006	0.000050
CHC	0.000332	0.002656
SSGA	0.009823	0.078586
SS-Arit	0.014171	0.113371
IPOP-CMA-ES	0.083642	0.669139
SS-BLX	0.141093	1.0
DE-Bin	0.518605	1.0
SaDE	0.660706	1.0