

Neyman–Pearson Theory of Testing and Mayo’s Extensions Applied to Evolutionary Computing

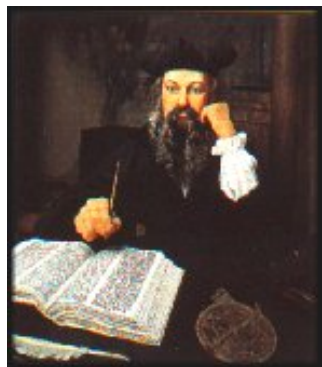
Thomas Bartz-Beielstein

Algorithm Engineering
Universität Dortmund

June 2006

- 1 Goals
- 2 History
- 3 Sequential Parameter Optimization
- 4 Observed Significance Plots

Scientific Goals?



- Why is astronomy considered scientific—and astrology not?
- And what about experimental research in computer science (evolutionary computation)?

Figure: Nostradamus

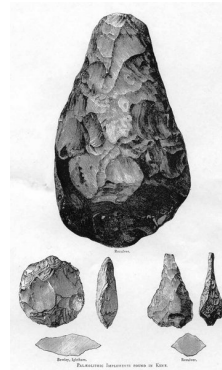
Goals in Evolutionary Computation

- (RG-1) *Investigation*. Specifying optimization problems, analyzing programs. Detect important parameters
- (RG-2) *Comparison*. Comparing the performance of programs
- (RG-3) *Conjecture*. Good: demonstrate performance. Better: explain and understand performance
- (RG-4) *Quality*. Robustness (includes insensitivity to exogenous factors, minimization of the variability) [Mon01]

What About Theory?

- Given: Hard real world optimization problems, e.g., chemical engineering, airfoil optimization, bioinformatics
- Many theoretical results are too abstract, do not match with reality
- Real programs, not algorithms
- Develop problem specific programs, experimentation is necessary
- Experimentation requires statistics
- But: There are three kinds of lies: lies, damned lies, and statistics (Mark Twain or Benjamin Disraeli), why should we care?
- Because it is the only tool we can rely on (at the moment, i.e., 2006)

A Totally Subjective History of Experimentation in Evolutionary Computation



- Palaeolithic
- Yesterday
- Today
- Tomorrow

Stone Age: Experimentation Based on Mean Values

- First phase (foundation and development, before 1980)
- Comparison based on mean values, no statistics
- Development of standard benchmark sets (sphere function etc.)
- Today: Everybody knows that mean values are not sufficient

Stone Age Example: Comparison Based on Mean Values

Example (Particle swarm optimization: swarm size)

- Experimental setup:
 - 4 test functions: Sphere, Rosenbrock, Rastrigin, Griewangk
 - Program's parameter: default setting
- Results: Table form, e.g.,

Table: Mean fitness values for the Rosenbrock function

Population	Dimension	Generation	Fitness
20	10	1000	96,1725
20	20	1500	214,6764

- Conclusion: "Under all the testing cases, the program always converges very quickly"

Yesterday: Mean Values and Simple Statistics



- Second phase (move to mainstream, 1980-2000)
- Statistical methods introduced, mean values, standard deviations, tutorials
- t test, p value, ...
- Comparisons mainly on standard benchmark sets
- Questionable assumptions

Today: Based on Correct Statistics



- Third phase (Correct statistics, since 2000)
 - Statistical tools for EC
 - Conferences, tutorials, workshops, e.g., Workshop On Empirical Methods for the Analysis of Algorithms (EMAA) (<http://www.imada.sdu.dk/~marco/EMAA>)

Today: Based on Correct Statistics

Today: Based on Correct Statistics

Example (Good practice)

Table 3: Results of the algorithms with population of 20

Test functions	SGA mean best (std. dev.)	FDGA			t-value between SGA to the best FDGA	Best algorithm
		OGA mean best (std. dev.)	MGA mean best (std. dev.)	EA mean best (std. dev.)		
f_1	8.060e+000 1.553e+000	8.568e+000 1.667e+000	8.654e+000 1.508e+000	8.272e+000 1.572e+000	-6.76 *	SGA
f_2	7.801e-001 4.583e-000	4.247e+000 1.321e+000	3.544e+000 2.087e+000	3.509e+000 1.497e+000	-4.00 *	SGA
f_3	6.405e+000 1.800e+000	9.272e+000 1.807e+000	8.660e+000 1.861e+000	8.637e+000 1.986e+000	-5.65 *	SGA
f_4	1.350e+002 3.349e+002	9.220e+002 2.807e+002	8.207e+002 2.599e+002	8.227e+002 2.485e+002	-11.89 *	SGA
f_5	2.747e-002 3.083e-002	6.823e-002 5.477e-002	8.205e-002 5.202e-002	6.247e-002 5.599e-002	-3.87 *	SGA
f_6	2.079e-003 9.184e-004	2.705e-003 3.528e-006	2.591e-003 3.320e-006	2.583e-003 2.737e-006	15.81 *	FDGA
f_7	2.079e-003 9.184e-004	4.333e-011 7.549e-012	4.019e-011 8.049e-012	4.006e-011 8.329e-012	1.91 *	FDGA
f_8	7.121e+001 7.121e+001	5.015e+001 4.112e+001	5.177e+001 3.757e+001	4.064e+001 4.106e+001	3.13 *	FDGA
f_9	1.485e-001 6.257e-002	5.128e-002 4.193e-003	4.651e-002 1.672e-002	4.650e-002 1.285e-002	11.33 *	FDGA
f_{10}	9.212e-002 6.105e-002	7.232e-002 2.138e-002	6.480e-002 2.180e-002	6.484e-002 2.402e-002	2.94 *	FDGA

* The value of t with 49 degrees of freedom is significant at $\alpha = 0.05$ by a one-tailed test.

Figure: [CAF04]

Example (Good practice?)

- We need tools to
 - Determine adequate number of function evaluations to avoid floor or ceiling effects
 - Determine the correct number of repeats
 - Determine suitable parameter settings for comparison
 - Determine suitable parameter settings for comparison
 - Draw meaningful conclusions
- Authors used
 - Pre-defined number of evaluations set to 200,000
 - 50 runs for each program
 - Population sizes 20 and 200
 - Crossover rate 0.1 in program A, but 0.173 in B
 - A outperforms B significantly in f_6 to f_{10}

- Today: Adequate statistical methods, but wrong scientific conclusions
- Tomorrow:
 - Consider scientific meaning
 - Severe testing as a basic concept

- Generally: Statistical tools to decide whether a is better than b are necessary
- Today: Sequential parameter optimization (SPO)
 - Heuristic, but implementable approach
 - Extension of classical approaches from statistical design of experiments (DOE)
 - Other (better) approaches possible
 - SPO uses plots of the observed significance

SPO Overview

- **Pre-experimental** planning
- **Scientific** thesis
- **Statistical** hypothesis
- Experimental **design**: Problem, constraints, start-/termination criteria, performance measure, program parameters
- **Experiments**
- Statistical **model** and prediction (DACE). Evaluation and visualization
- Solution good enough?
 - Yes: Goto step 15
 - No: Improve the design (optimization). Goto step 15
- **Acceptance/rejection** of the statistical hypothesis
- Objective **interpretation** of the results from the previous step

Tests and Significance

- Plots of the observed significance level based on [May83]
- Rejection of the null hypothesis $H : \theta = \theta_0$ by a test T^+ based on an observed average \bar{x}
- Alternative hypothesis $J : \theta > \theta_0$

Definition (Observed significance level)

The observed significance level is defined as

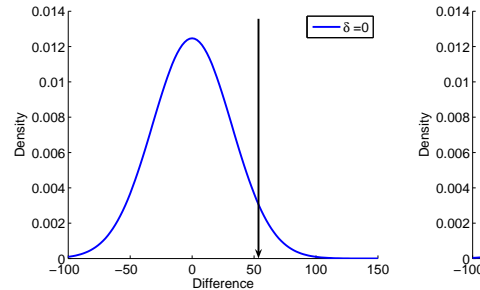
$$\alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x} | \theta) \quad (1)$$

Plots of the Observed Significance

- Observed significance level

$$\alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x} | \theta)$$

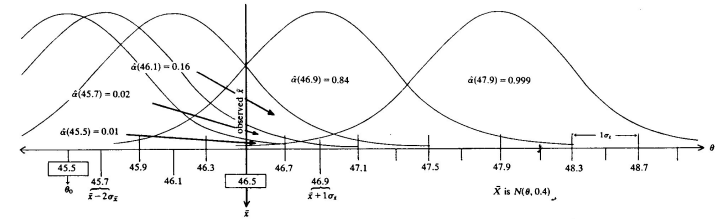
- Observed average $\bar{x} = 51.73$



- Rejection of the null hypothesis
- $H : \theta = \theta_0 = 0$
- by a test T^+ in favor of an alternative
- $J : \theta > \theta_0$
- Then $\hat{\alpha}(\theta) = 0.0530$
- Interpretation: Frequency of erroneously rejecting H ("there is a difference in means as large as θ_0 or larger") with such an \bar{x}

"An objective theory of statistical testing" [May83]

- Discriminate between legitimate and illegitimate construals of statistical results by considering the values of $\hat{\alpha}(\theta')$ for several θ' values
- [May83] defines θ_{un} the largest scientifically unimportant θ value in excess of θ_0
- But what if we do not know θ_{un} ?



OSL Plots

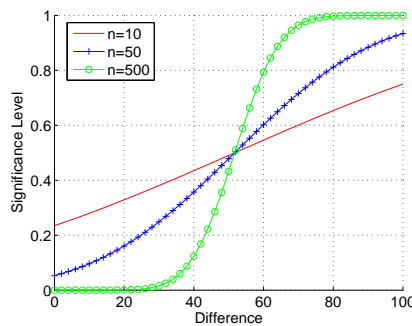
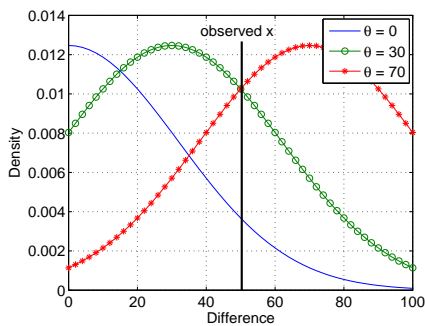
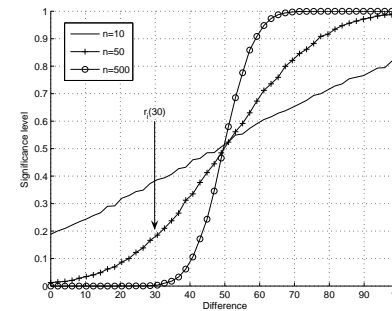


Figure: Plots of the observed difference. *Left*: This is similar to Fig. 4.3 in [May83]. Based on $n = 50$ experiments, a difference $\bar{x} = 51.3$ has been observed, $\hat{\alpha}(\theta)$ is the area to the right of the observed difference \bar{x} . *Right*: The $\hat{\alpha}(\theta)$ value is plotted for different n values.

OSL Plots and the Bootstrap







- Bootstrap procedure \Rightarrow no assumptions on the underlying distribution necessary
- Summary:
 - p value is not sufficient
 - OSL plots one tool to derive meta-statistical rules
 - Other tools needed

Figure: Same situation as above, bootstrap approach

Additional Information, Software



- Please check the WWW pages for [BB06]
<http://ls11-www.cs.uni-dortmund.de/people/tom/ExperimentalResearch.html>
 for papers, slides, software, etc.

-  **Thomas Bartz-Beielstein.**
Experimental Research in Evolutionary Computation—The New Experimentalism.
 Springer, Berlin, Heidelberg, New York, 2006.
-  **Kit Yan Chan, Emin Aydin, and Terry Fogarty.**
 An empirical study on the performance of factorial design based crossover on parametrical problems.
 In *Proceedings of the 2004 IEEE Congress on Evolutionary Computation*, pages 620–627, Portland, Oregon, 20-23 June 2004. IEEE Press.
-  **D. G. Mayo.**
 An objective theory of statistical testing.
Synthese, 57:297–340, 1983.
-  **D. C. Montgomery.**
Design and Analysis of Experiments.
 Wiley, New York NY, 5th edition, 2001.