

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

Thomas Bartz–Beielstein

Dortmund University, 44221 Dortmund, Germany,

Thomas.Bartz-Beielstein@udo.edu,

WWW home page: <http://ls11-www.cs.uni-dortmund.de/people/tom>

**Abstract.** *Evolutionary computation* (EC) is a relatively new discipline in computer science (Eiben & Smith, 2003). It tackles hard real-world optimization problems, e.g., problems from chemical engineering, airfoil optimization, or bioinformatics, where classical methods from mathematical optimization fail. Many theoretical results in this field are too abstract, they do not match with reality. To develop problem specific algorithms, experimentation is necessary. During the first phase of experimental research in EC (before 1980), which can be characterized as “foundation and development,” the comparison of different algorithms was mostly based on mean values, nearly no further statistics have been used. In the second phase, where EC “moved to mainstream” (1980-2000), classical statistical methods were introduced. There is a strong need to compare EC algorithms to mathematical optimization (main stream) methods. Adequate statistical tools for EC are developed in the third phase (since 2000). They should be able to cope with problems like small sample sizes, nonnormal distributions, noisy results, etc.

However—even if these tools are under development—they do not bridge the gap between the statistical significance of an experimental result and its scientific meaning. Based on Mayo’s *learning model* ( $NPT^*$ ) we will propose some ideas how to bridge this gap (Mayo, 1983, 1996). We will present plots of the *observed significance level* and discuss the *sequential parameter optimization* (SPO) approach. SPO is a heuristic, but implementable approach, which provides a framework for a sound statistical methodology in EC (Bartz-Beielstein, 2006).

## 1 Introduction

### 1.1 Experimental Research in Evolutionary Computation

One of the major goals in EC is to demonstrate that an algorithm,  $A$ , outperforms a related algorithm,  $B$ . Researchers suppose that  $A$  and  $B$  behave differently, because one algorithm has features the other lacks, e.g., an improved variation operator. Experimental hypotheses have implicitly the form “factor  $X$  produces result  $Y$ .” However, many experiments in EC do not test these hypotheses directly. A common way presenting results and drawing conclusions is to run algorithms  $A$  and  $B$  on a given set of problem instances and to compare their

results. These kind of experiments can be classified as *observation experiments*, because experimenters use default factor settings that are not varied. Observation experiments were predominating during the first two phases of experimental research in EC (until 2000). Enhanced experimental techniques, so-called *manipulation experiments*, vary several factors: experimenters demonstrate a relation between  $X$  and  $Y$ . Systematic approaches based on analysis of variance or regression techniques became popular in the last years. However—even if statistical tools for manipulating experiments are under development—they do not bridge the gap between the statistical significance of an experimental result and its scientific meaning.

Based on a standard situation in experimental research we propose a methodology to analyze the relationship between statistical significance and scientific import. Common to all observation and manipulation experiments is the need to compare two algorithms, a task that can be modeled in the framework of hypothesis testing. To test the hypothesis that algorithm  $A$  is better than  $B$ , first assume that they perform equally, i.e., there is no difference in means. Therefore we are facing a standard situation from statistics, the comparison of samples from two populations. Computer experiments can be designed, e.g., *common random numbers* (CRN) can be used. If the same number of runs with similar random seeds are performed, paired  $t$ -tests can be used to analyze the results. The reader is referred to Law & Kelton (2000) for a discussion of CRN and related variance-reducing techniques.

## 1.2 Paired $t$ -tests

The following standard test situation will be discussed in the remainder of this article. The  $j$ th paired difference

$$x_j = y_{1j} - y_{2j} \quad j = 1, \dots, n,$$

is used to define the test statistic

$$t_0 = \frac{\bar{x}}{S_x/\sqrt{n}},$$

where  $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$  denotes the sample mean, and

$$S_x = \sqrt{\sum_{j=1}^n \frac{(x_j - \bar{x})^2}{n-1}}, \quad (1)$$

is the *sample standard deviation of the differences*. Let

$$\theta = \mu_1 - \mu_2$$

denote the difference in means. The null hypothesis  $H : \mu_1 = \mu_2$ , or equivalently  $H : \theta = 0$ , would be not accepted if  $t_0 > t_{\alpha, n-1}$ . The paired  $t$ -test can be advantageous compared to the two-sample  $t$ -test due to its noise reduction properties. The reader is referred to the discussion in Montgomery (2001).

### 1.3 The New Experimentalism

In the following, we will introduce methods from the *new experimentalism* into evolutionary computation. The new experimentalism is an influential trend in recent philosophy of science that provides statistical methods to set up experiments, to test algorithms, and to learn from the resulting errors and successes. The new experimentalists are seeking a relatively secure basis for science, not in theory or observation but in experiment. To get the apparatus working for simulation studies is an active task. Sometimes the recognition of an oddity leads to new knowledge. Important representatives of the new experimentalism are Hacking (1983), Galison (1987), Gooding et al. (1989), Mayo (1996), and Franklin (2003). Deborah Mayo, whose work is in the epistemology of science and the philosophy of statistical inference, proposes a detailed way in which scientific claims are validated by experiment. A scientific claim can only be said to be supported by experiment if it passes a severe test. A claim would be unlikely to pass a severe test if it were false. Mayo developed methods to set up experiments that enable the experimenter, who has a detailed knowledge of the effects at work, to learn from error. Our presentation is based on Mayo (1983).

Severity is introduced in Section 2. Plots of the observed significance level are introduced as tools to derive metastatistical rules to test whether statistical significant results are scientifically relevant. Section 3 summarizes the sequential parameter optimization which defines a standard framework for experimental research in evolutionary computation.

## 2 Severity

The major goal introduced in Sec. 1.1 can be formulated in the context of Mayo's *objective theory of statistical testing*. To stay consistent with Mayo's seminal text, we assume a known standard deviation  $\sigma$  in the first part of this study and use the same notation as in Mayo (1983). Some quantity  $X$  is normally distributed, i.e.,  $X \sim \mathcal{N}(\theta, \sigma^2)$ . The goal is to test whether the value of  $\theta$  equals some value  $\theta_0$  or some greater value. Then the null ( $H$ ) and the alternative hypothesis ( $J$ ) read:

$$H : \theta = \theta_0 \text{ vs. } J : \theta > \theta_0 \quad (2)$$

The experimental test statistic  $S$  is the average of the  $n$  random variables  $X_i$ ,  $i = 1, \dots, n$ :

$$\text{Test Statistic } S = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}. \quad (3)$$

Under the assumption that the null is true, the experimental distribution of  $\bar{X}$  is  $\mathcal{N}(\theta_0, \sigma/\sqrt{n})$ . Let  $\alpha$  denote the *significance level* of the test. If  $\alpha$  is specified, the *critical value*  $d_\alpha$  can be determined, so that

$$P(\bar{X} \geq \sigma_0 + d_\alpha \sigma_{\bar{x}}; H) \leq \alpha \quad (4)$$

The corresponding critical values for  $\alpha = 0.01$ ,  $0.02$ , and  $0.05$  are  $d_\alpha = 2.3$ ,  $d_\alpha = 2$ , and  $d_\alpha = 1.6$ , respectively. The *test rule (RU)*, represented by  $T^+$ , maps  $\bar{X}$  into the critical region when  $\bar{X}$  is significantly far from the hypothesized average  $\theta_0$ , i.e.,

$$\text{Test Rule } T^+: \text{Reject } H : \theta = \theta_0 \text{ iff } \bar{X} \geq \theta_0 + d_\alpha \sigma_{\bar{x}} \quad (5)$$

The areas under the normal curve given by the distribution of  $\bar{X}$  under  $H$  illustrates the relationship between  $\alpha$  and the distances of  $\bar{X}$  from  $\theta_0$ . Similar to Mayo (1983), we present an example of an application of test  $T^+$ . It will be discussed in the context of EC.

## 2.1 An Example From Evolutionary Computation

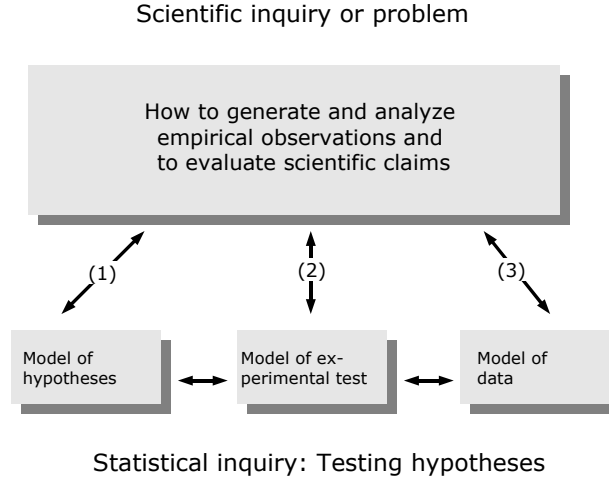
The flocking behavior of swarms and fish shoals was the main inspiration which led to the development of *particle swarm optimization* algorithms (Kennedy & Eberhart, 1995). Particle swarm optimization algorithms belong to the class of stochastic, population-based optimization algorithms. They exploit a population of individuals to probe the search space. In this context, the population is called a *swarm* and the individuals are called *particles*. Each particle moves with an adaptable velocity within the search space, and it retains in a memory the best position it has ever visited. Particle swarm optimization has been applied to numerous simulation and optimization problems in science and engineering (Kennedy & Eberhart, 2001; Parsopoulos & Vrahatis, 2002, 2004).

**Example 1 (Particle swarm size).** Analyzing a *particle swarm optimization algorithm* (PSO), we are interested in testing whether or not the swarm size has a significant influence on the performance of the algorithm. A minimization task, the 10-dimensional Rosenbrock function was chosen as a test function (Rosenbrock, 1960). Based on the parameterization in Shi & Eberhart (1999), the swarm sizes were set to 20 and 40. The corresponding settings will be referred to as run PSO(20) and PSO(40), respectively. The question is whether the increased swarm size improves the performance of the particle swarm optimization. Our inquiry can be formulated as the scientific claim

**Scientific Claim 1 (C).** Increasing the swarm size from 20 to 40 particles improves the algorithm’s performance.

As in Shi & Eberhart (1999), a random sample is drawn from each of the two populations. The average performance  $\bar{y}_1$  of  $n = 50$  runs of PSO(20) is 108.02, whereas the average performance  $\bar{y}_2$  of  $n = 50$  runs of PSO(40) is 56.29. The same number of function evaluations was used in both settings. The number of runs  $n$  is referred to as the *sample size*, and  $\bar{y}$  denotes the *sample mean*.  $\square$

A typical problem in this situation can be described as follows: The difference of the function values from the samples of the particle swarm optimizer may be observed to have an average performance  $\bar{X}$  that is larger than 0 even though the increased swarm size does not have the positive effect. Mayo (1983) states: “As such, the need for statistical considerations arises.” We present the situation from EC in the context of Mayo’s models of statistical testing, see Fig. 1.



**Fig. 1.** Models of statistical testing. Mayo (1983) develops a framework that permits a delinearization of the complex steps from raw data to scientific hypotheses. Primary questions arise when a substantive scientific question is broken down into several local hypotheses. Experimental models link primary questions based on the model of hypotheses to questions about the actual experiment. Data models describe how raw data are transformed before. Not the raw data, but these modeled data are passed to the experimental models. Mayo (1983) describes three major *metastatistical tasks*: “(1) relating the statistical hypotheses [...] and the results of testing them to scientific claims; (2) specifying the components of experimental test [...]; and, (3) ascertaining whether the assumptions of a model for the data of the experimental test are met by the empirical observations [...]”

- (A) *Statistical Hypothesis*. We will assume that the standard deviation is known, in our example  $\sigma = 160$ , see Bartz-Beielstein (2006). If  $\mathcal{C}$  is wrong, the distribution of  $X$  among particle swarm optimizers with an increased swarm size would not differ from particle swarm optimizers with 20 particles only. If  $\mathcal{C}$  is true, and the increased swarm size does have a positive effect,  $\theta$  will be larger than 0. These observations correspond to the following statistical hypotheses:

$$H : \theta = 0 \text{ vs. } J : \theta > 0. \quad (6)$$

- (B) *Experimental Test*. The vector  $y_i = (y_{i1}, \dots, y_{in})$  represents  $n$  observations from the  $i$ th configuration, and  $\bar{y}_i$  denotes the  $i$ th sample mean,  $i = 1, 2$ . The experimental test statistic is  $T = \bar{Y}_{12} = \bar{Y}_1 - \bar{Y}_2$ , and its distribution under  $H$  is  $\mathcal{N}(0, 2\sigma^2/n)$ . The *upper  $\alpha$  percentage point of the normal distribution* is denoted as  $z_\alpha$ , for example,  $z_{0.05} = 1.64$ , or  $z_{0.01} = 2.33$ . As the number

of observations was set to  $n = 50$ , it follows that the value of the *standard error* is  $\sigma_{\bar{x}} = \sigma_{\bar{y}_1 - \bar{y}_2} = 160\sqrt{2/50} = 32$ . The significance level of the test was  $\alpha = 0.01$ , thus  $z_\alpha = z_{0.01} = 2.33$ . So the *test rule RU* is

$$T : \text{Reject } H : \theta = 0 \text{ if } T = \bar{Y}_1 - \bar{Y}_2 \geq 0 + z_\alpha \cdot \sigma_{\bar{x}}.$$

(C) *Sample data.* The average performance  $\bar{y}_1$  of  $n = 50$  runs of PSO(20) is 108.02, whereas the average performance  $\bar{y}_2$  of  $n = 50$  runs of PSO(40) is 56.29. The difference  $\bar{x} = \bar{y}_1 - \bar{y}_2$  is 51.73. Since this value does not exceed 74.44, the test rule  $T^+$  does not reject  $H$ .

## 2.2 Empirical Scientific Inquiries and Statistical Models of NPT

The problem how to relate an empirical scientific inquiry to statistical models of NPT is a metastatistical problem. NPT can be interpreted as a means of deciding how to behave. To contrast her reformulation of NPT with this behavioristic model, Mayo (1983) introduces the term *learning model*, or simply NPT\*, for the former. NPT\* goes beyond NPT, it uses the distribution of the test statistic  $S$  (Eq. 3) to control error probabilities. Statistical tests are seen as “means of learning about variable phenomena on the basis of limited empirical data.” In the particle swarm example we are interested in learning if particle swarm optimizers with increased population sizes give rise to performance improvements, i.e., performance describable by  $\mathcal{N}(0, \sigma)$ , or by  $\mathcal{N}(\theta, \sigma)$ , with  $\theta > 0$ .

Mayo (1983) claims that NPT\* provides tools for specifying tests that “will very infrequently classify an observed difference as significant (and hence reject  $H$ ) when no discrepancy of scientific importance is detected, and very infrequently fail to do so (and so accept  $H$ ) when  $\theta$  is importantly very discrepant from  $\theta_0$ .”

## 2.3 Mayo’s Objective Interpretation of Rejecting a Hypothesis

NPT\* provides several tools for detecting whether the scientific relevance is misconstrued, e.g., if accidental effects such as measurement errors occur. A difference of 1 or 2 standard deviation units between  $\bar{X}$  and its mean  $\theta$  arises relatively often, so these difference can be easily confused with effects caused by real differences. Two types of misconstruals or misinterpretations can arise.

1. A test can be specified so that it will give rise to a  $\bar{x}$  that exceeds  $\theta_0$  by the required  $d_\alpha \sigma_{\bar{x}}$ , cf. Equation 5. Large sample sizes provide such sensitive tests, so that  $\theta_0 + d_\alpha \sigma_{\bar{x}}$  can be made so small that even the smallest differences appear meaningful.
2. A test can be specified so that it will give rise to a  $\bar{x}$  that does not exceed  $\theta_0$  by the required  $d_\alpha \sigma_{\bar{x}}$ . Decreasing  $\alpha$  values provide such insensitive tests, so that  $\theta_0 + d_\alpha \sigma_{\bar{x}}$  can be made so large that even the large differences appear meaningless.

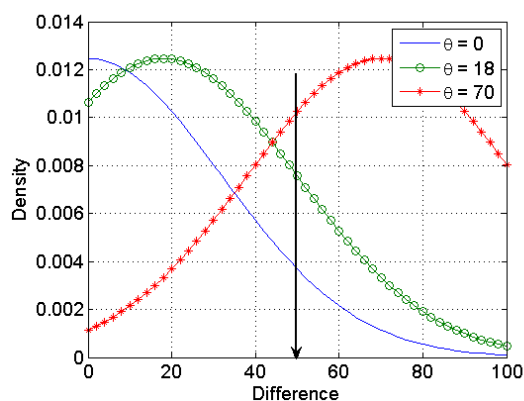
To avoid these misinterpretations, Mayo proposed considering the observed significance level.

## 2.4 The Observed Significance Level

The frequency relation between a rejection of the null hypothesis  $H$  and values of the difference in means,  $\theta$ , is important for the interpretation of the rejection. To interpret the rejection of  $H$ , Mayo introduces the *observed significance level*

$$\alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x}; \theta) \quad (7)$$

Hence,  $\hat{\alpha}(\theta)$  is the area under the normal curve to the right of the observed  $\bar{x}$ , as illustrated in Fig. 2. Note, that  $\hat{\alpha}$  is the frequency of an error of the first kind



**Fig. 2.** Observed difference and three hypothetical differences. Difference in means for  $n = 50$  samples and standard deviation  $\sigma = 160$ . The value from the test statistic  $\bar{x} = 51.73$  remains fixed for varying means  $\theta_i$  of different distributions associated with the null hypotheses  $H_i$ ,  $i = 1, 2, 3$ . The figure depicts the probability density functions of the associated normal distributions for three different means:  $\theta_1 = 0$ ,  $\theta_2 = 18$ , and  $\theta_3 = 70$ . To interpret the results, consider a hypothetical difference in means of  $\theta_2 = 18$ : The observed significance level  $\alpha(\bar{x}, \theta)$  is the area under the normal curve to the right of  $\bar{x}$ . The value  $\alpha(51.75, 18)$  is quite large and therefore not a good indication that the true difference in means is as large as  $\theta_2 = 18$ . This figure corresponds to Fig. 4.3 in Mayo (1983)

if we set  $\theta = \theta_0$ . A rejection of the null  $H : \theta = \theta_0$  is a good indicator that  $\theta > \theta_0$  if the observed significance level  $\hat{\alpha}$  is small. However, if some  $\theta$  values in excess of  $\theta_0$  are not deemed scientifically important, even small  $\hat{\alpha}$  values do not prevent such a rejection of the null from being misconstrued when relating it to the scientific claim  $C$ .

To relate the statistical result to the scientific import, Mayo proposes to define  $\theta_{\text{un}}$ :

$$\theta_{\text{un}} = \text{the largest scientifically unimportant value in excess of } \theta_0. \quad (8)$$

In many situations,  $\theta_{\text{un}}$  is not known exactly. Then, observing the values of  $\hat{\alpha}(\theta')$  for  $\theta' \in \Omega_J$  indicates if the construal is legitimate or illegitimate. If  $\hat{\alpha}(\theta')$  is large, then the statistical result is not a good indication that the scientific claim is true.

In addition to  $\theta_{\text{un}}$ , Mayo defined  $\theta^{\hat{\alpha}}$ , the *inversion of the observed significance level* function as:

$$\theta^{\hat{\alpha}} = \text{the value of } \theta \text{ in } \Omega \text{ for which } \alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = \hat{\alpha}. \quad (9)$$

**Example 2.** Consider a sample size of  $n = 50$ . If  $\theta_{\text{un}} = 30$ , then rejecting  $H$  cannot be taken as an indication that the scientific claim “PSO(40) outperforms PSO(20)” is true. Figure 3 illustrates this situation. The observed significance level  $\hat{\alpha}(30) = 0.25$  is not a strong indication that  $\theta$  exceeds 30. However, if the sample size is increased ( $n = 500$ ), then  $\hat{\alpha}(30) = 0.05$  is small. This example illustrates that  $\hat{\alpha}$  is a function of the sample size  $n$ .  $\square$

But are these results good indications that one is observing a difference  $\theta > 0$  that is also scientifically important? This problem is outside the domain of statistics. Its answer requires the specification of a scientifically important difference, a reasonable sample size, and an acceptable error of the first kind. The  $\hat{\alpha}$  function provides a nonsubjective tool for understanding the  $\theta$  values, a metastatistical rule that enables learning on the basis of a given  $RU$  rejection. As the examples demonstrate, NPT\* tools enable the experimenter to control error probabilities in an objective manner.

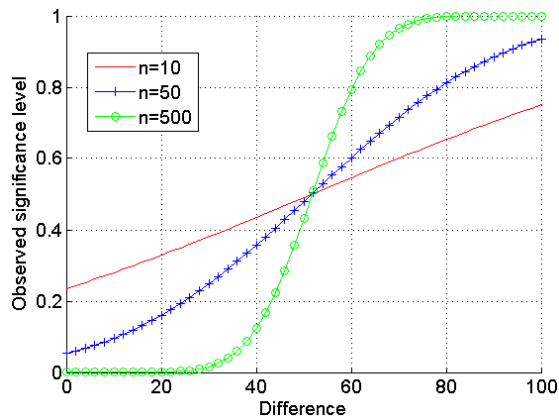
## 2.5 Monte Carlo Simulations and the Bootstrap

Mayo (1983) introduced the observed significance level under the assumptions that

1. the samples follow a normal distribution
2. the variance  $\sigma^2$  is known

In many real-world situations, these assumptions are not true. Results from performance comparisons are not normally distributed, e.g., many values are worse, but no value is better than the optimum. Bartz-Beielstein (2006) proposes a bootstrap approach to tackle this difficulties.

**Bootstrap** *Monte Carlo simulations* can be applied for known population distributions from which the samples are drawn and unknown sampling distributions of the test statistic, for example, the trimmed mean or the interquartile range. As *bootstrap* methods treat the sample as the population, they can be applied if the sampling distribution is unknown (Efron & Tibshirani, 1993). They require a representative sample of the population. Nowadays the bootstrap is considered a standard method in statistics (Mammen & Nandi, 2004). It has been successfully applied to solve problems that would be too complicated for classical statistical techniques and in situations where the classical techniques are not valid (Zoubir & Boashash, 1998).



**Fig. 3.** Plot of the observed significance level  $\alpha(\bar{x}, \theta)$  as a function of  $\theta$ , the possible true difference in means. Lower  $\hat{\alpha}$  values support the assumption that there is a difference as large as  $\theta$ . The measured difference is  $\bar{x} = 51.73$ , the standard deviation is  $\sigma = 160$ , cf. Example 1. Each point of the three curves shown here represents one single curve from Fig. 2. The observed significance value is value of area under the normal curve to the right of the observed difference  $\bar{x}$ . The values can be interpreted as follows: Regard  $n = 50$ . If the true difference is (a) 0, (b) 51.73, or (c) 100, then (a)  $H : \theta = 0$ , (b)  $H : \theta = 51.73$ , or (c)  $H : \theta = 100$  is wrongly rejected (a) 5%, (b) 50%, or (c) 95% of the time

The idea behind the bootstrap is similar to a method that is often applied in practice. Experiments are repeated to improve the estimate of an unknown parameter. If a representative sample is available, the bootstrap randomly reassigns the observations and recomputes the estimate. The bootstrap is a computationally intensive technique. Let  $\hat{\theta}$  be the estimate of an unknown parameter  $\theta$  that has been determined by calculating a statistic  $S$  from the sample:

$$\hat{\theta} = S = s(y_1, \dots, y_n).$$

By sampling with replacement,  $n_b$  bootstrap samples can be obtained. The bootstrap replicates of  $\hat{\theta}$

$$\hat{\theta}^{*b} = s(y^{*b}), \quad b = 1, \dots, n_b,$$

provide an estimate of the distribution of  $\hat{\theta}$ . The generic bootstrap procedure is described in Fig. 4.

We describe the basic bootstrap procedure to determine the observed significance level  $\alpha(\bar{x}, \theta)$ . It can be applied to generate plots of the observed significance, as shown in Fig. 3. Note that this procedure requires only two paired and representative samples,  $y_1$  and  $y_2$ .

Let  $y_1 = (y_{11}, \dots, y_{1n})^T$  and  $y_2 = (y_{21}, \dots, y_{2n})^T$  denote the random samples, and  $x = y_1 - y_2 = (y_{11} - y_{21}, \dots, y_{1n} - y_{2n})^T$  their difference vector. The procedure to obtain an estimate of the observed significance level  $\alpha(\bar{x}, \theta)$  for a

**Algorithm:** Generic Bootstrap

1. Calculate  $\hat{\theta}$  from a representative sample  $y = (y_1 \dots, y_n)$ .
2. To generate the bootstrap data sets  $y^{*b} = (y_1^{*b}, \dots, y_n^{*b})$  sample with replacement from the original sample.
3. Use the bootstrap sample  $y^{*b}$  to determine  $\hat{\theta}^{*b}$ .
4. Repeat steps 2 and 3  $n_b$  times.
5. Use this estimate of the distribution of  $\hat{\theta}$  to obtain the desired parameter, for example the mean.

**Fig. 4.** The generic bootstrap procedure

difference  $\theta_0$  under the null hypothesis  $H$  can be implemented as in the following example:

**Example 3 (Bootstrap).** Let  $y_1$  and  $y_2$  denote two vectors with representative samples from a population. If  $a \in \mathbb{R}$  and the vector  $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ , the *scalar-vector addition* is defined as

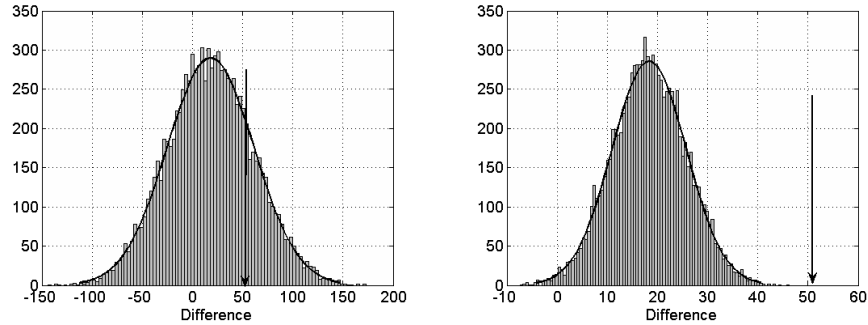
$$a + y = (y_1 + a, \dots, y_n + a)^T.$$

The bootstrap approach to generate the plots of the observed significance comprises the steps shown in Fig. 4. They can be detailed as follows:

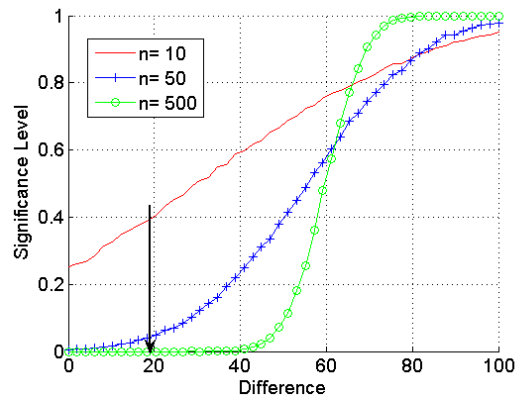
1. Calculate  $x = y_1 - y_2$ .
2. Determine  $\bar{x} = 1/n \sum_{j=1}^n (y_{1j} - y_{2j})$ .
3. Specify the lower bound  $a$  and the upper bound  $b$  for the plot.
4. Specify  $m$ , the number of points to be plotted in the interval  $[a, b]$ .
5. For  $i = 1$  to  $m$  do:
  - (a) Determine  $x_i = x - \bar{x} + w_i$  with  $w_i = a + i \times (b - a)/m$ .
  - (b) Generate  $n_b$  bootstrap sample sets  $x_i^{*b}$ ,  $b = 1, \dots, n_b$  from  $x_i$ .
  - (c) Determine the  $n_b$  mean values  $\bar{x}_i^{*b}$ .
  - (d) Determine  $n_i$ , that is, the number of times that  $\bar{x}_i^{*b} > \bar{x}$ .
  - (e) Determine the ratio  $r_i = n_i/n_b$ .

Finally, the  $m$  points  $(w_i, r_i)$  are plotted. The ratio  $r_i$  is a bootstrap estimate of the observed significance value  $\alpha(\bar{x}, w_i)$ .  $\square$

Histograms of the bootstrap replicates as shown in Fig. 5 are appropriate tools for examining the distribution of  $\hat{\theta}$ . Figure 6 depicts the result based on the bootstrap. It represents the same situation as shown in Fig. 3, without making any assumption on the underlying distribution. As the sample size is increased, i.e., from 50 to 500, the bootstrap and the true curve start to look increasingly similar.



**Fig. 5.** Histograms of the bootstrap samples. *Left:* 50 samples (repeats); *right:* 500 samples. These figures show histograms of the bootstrap samples that were generated at step 5 in Example 3. The difference  $\theta$  has the value 18.37. The *dash-dotted curves* show the superimposed normal density. The area to the right of  $\bar{x} = 51.73$  under the curve corresponds approximately with the observed significance level  $\alpha(51.73, 18.37)$ , i.e., the ratio  $r_i$

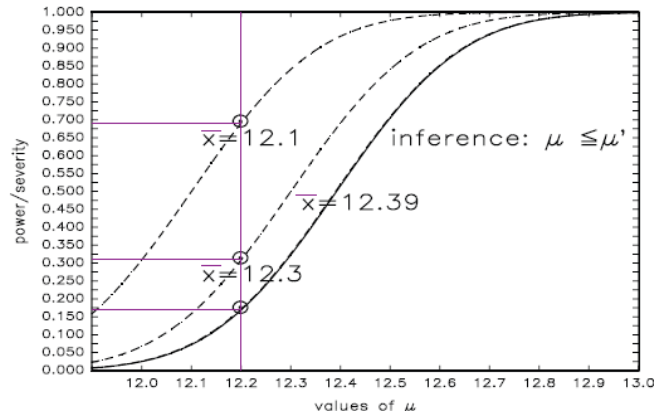


**Fig. 6.** This figure depicts the same situation as shown in Fig. 3. But, unlike in Fig. 3, no assumptions on the underlying distribution have been made. Samples of size  $n = 10$ , 50, and 500, respectively, have been drawn from a normal distribution. The bootstrap procedure described in Example 3 has been used to generate this plot. The curves look qualitatively similar to the curves from Fig. 3. As the number of samples increases, the differences between the exact and the bootstrap curves becomes smaller. The measured difference is 51.73,  $\sigma = 160$ , cf. Example 1. Regard  $n = 50$ : If the true difference is (a) 0, (b) 51.73, or (c) 100, then (a)  $H : \delta = 0$ , (b)  $H : \delta = 51.73$ , or (c)  $H : \delta = 100$  is (approximately) wrongly rejected (a) 1%, (b) 50%, or (c) 99% of the time

The (bootstrap) plots of the observed significance can be used to apply concepts developed in Mayo (1983) to important research goals in EC. They make no assumptions on the underlying distribution. Bartz-Beielstein (2006) presents further examples and guidelines how to interpret plots of the observed significance.

## 2.6 Plots of the Observed Significant and Mayo’s Severity Curves

Mayo & Spanos (2006) present severity curves, cf. Fig. 7. These curves show the same data as plots of the observed significance, but from a different perspective. Severity curves vary the observed sample values  $\bar{x}$ , whereas  $\bar{x}$  remains unchanged in plots of the observed significance. The former illustrate an enhanced (“fuzzy-fied”) power concept, the so-called attained or actual power, the latter vary the alternative hypotheses  $\theta'$  and the number of experiments to illustrate the effect on the error of the first kind.



**Fig. 7.** Mayo & Spanos (2006) define “the severity with which we claim  $\mu \leq \mu_1$  passes test  $T(\alpha)$  with data  $x_0$ ” in the case of an a statistically insignificant result, i.e., accept  $H$ , as  $\text{SEV}(\mu \leq \mu_1) = P(d(X) > d(x_0); \mu \leq \mu_1 \text{ false})$ . Here, the experimental result is varied, whereas plots of the observed significance level vary the number of experiments and the hypothesized differences

## 3 Sequential Parameter Optimization

*Sequential parameter optimization* describes an implementable but heuristic method for the comparison and analysis of computer programs. It comprehends the 12 steps described in Tab. 1.

During the *preexperimental planning phase* (S-1) the experimenter defines exactly what is to be studied and how the data are to be collected. The recognition and statement of the problem seems to be a rather obvious task. However,

**Table 1.** Sequential parameter optimization. This approach combines methods from computational statistics and exploratory data analysis to improve (tune) the performance of direct search algorithms. It is detailed in Bartz-Beielstein (2006).

Step	Action
(S-1)	Preexperimental planning
(S-2)	Scientific claim
(S-3)	Statistical hypothesis
(S-4)	Specification of the <ul style="list-style-type: none"> <li>(a) Optimization problem</li> <li>(b) Constraints</li> <li>(c) Initialization method</li> <li>(d) Termination method</li> <li>(e) Algorithm (important factors)</li> <li>(f) Initial experimental design</li> <li>(g) Performance measure</li> </ul>
(S-5)	Experimentation
(S-6)	Statistical modeling of data and prediction
(S-7)	Evaluation and visualization
(S-8)	Optimization
(S-9)	Termination: If the obtained solution is good enough, or the maximum number of iterations has been reached, go to step (S-11)
(S-10)	Design update and go to step (S-5)
(S-11)	Rejection/acceptance of the statistical hypothesis
(S-12)	Objective interpretation of the results from step (S-11)

in practice, it is not simple to formulate a generally accepted goal. *Discovery*, *comparison*, *conjecture* and *robustness* are only four possible scientific goals of an experiment. Furthermore, the experimenter should take the boundary conditions into account. Statistical methods like run length distributions provide suitable means to measure the performance and describe the qualitative behavior of optimization algorithms.

In step (S-2), the experimental goal should be formulated as a scientific claim, e.g., “Algorithm  $A$ , which uses a swarm size  $s$ , that is proportional to the problem dimension  $d$  outperforms algorithms that use a constant swarm size.”

A statistical hypothesis, such as “There is no difference in means comparing the performance of the two competing algorithms,” is formulated in the step (S-3) that follows.

Step (S-4) requires the specification of problem and algorithm specific parameter settings, the so-called *problem* and *algorithm designs*.

After that, the experiment is run (S-5). Preliminary (pilot) runs can give a rough estimate of the experimental error, run times, and the consistency of the

experimental design. Since we consider probabilistic search algorithms in our investigation, design points must be evaluated several times.

The experimental results provide the base for modeling and prediction in step (S-6). The model is fitted and a predictor is obtained for each response.

The model is evaluated in step (S-7). Several visualization techniques can be applied. Simple graphical methods from exploratory data analysis are often helpful. Histograms and scatterplots can be used to detect outliers. If the initial ranges for the designs were chosen improperly (e.g., very wide initial ranges), visualization of the predictor can guide the choice of more suitable (narrower) ranges in the next stage. Several techniques to assess the validity of the model have been proposed. If the predicted values are not accurate, the experimental setup has to be reconsidered. This includes the scientific goal, the ranges of the design variables, and the statistical model. New design points in promising subregions of the search space can be determined (S-8) if further experiments are necessary. Thus, a termination criterion has to be tested (S-9). If it is not fulfilled, new candidate design points can be generated (S-10). A new design point is selected if there is a high probability that the predicted output is below the current observed minimum and/or there is a large uncertainty in the predicted output. Otherwise, if the termination criterion is true, and the obtained solution is good enough, the final statistical evaluation (S-11) that summarizes the results is performed. A comparison between the first and the improved configuration should be performed. Techniques from exploratory data analysis can complement the analysis at this stage. Besides an investigation of the numerical values, such as mean, median, minimum, maximum,  $\min_{\text{boot}}$  and standard deviation, graphical presentations such as boxplots, histograms, and RLDs can be used to support the final statistical decision.

Finally, we have to decide whether the result is scientifically important (S-12), since the difference, although statistically significant, can be scientifically meaningless. An objective interpretation of rejecting or accepting the hypothesis from (S-2) should be presented here. Consequences that arise from this decision are discussed as well. The experimenter's skill plays an important role at this stage. The experimental setup should be reconsidered at this stage and questions like "Have suitable test functions or performance measures been chosen?" or "Did floor or ceiling effects occur?" must be answered. Test problems that are too easy may cause such ceiling effects.

## 4 Summary

We described the current situation of experimental research in EC. Several statistical tools that reflect the requirements of today's optimization practitioners are developed nowadays. However, nearly no tools that enable an interpretation of and learning from the scientific results exists. Mayo's models of statistical testing bridge this gap. We extended an approach introduced in Mayo (1983), so that it is applicable if the underlying distributions are unknown. An example from EC was presented to illustrate this approach.

## Bibliography

- Bartz-Beielstein, T. (2006). *Experimental Research in Evolutionary Computation—The New Experimentalism*. Berlin, Heidelberg, New York: Springer.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.
- Eiben, A. E. & Smith, J. E. (2003). *Introduction to Evolutionary Computing*. Berlin, Heidelberg, New York: Springer.
- Franklin, A. (2003). Experiment in physics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford CA: Stanford University. <http://plato.stanford.edu/archives/sum2003/entries/physics-experiment>. Cited 14 April 2004.
- Galison, P. (1987). *How Experiments End*. Chicago IL: The University of Chicago Press.
- Gooding, D., Pinch, T., & Schaffer, S. (1989). *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge, U.K.: Cambridge University Press.
- Hacking, I. (1983). *Representing and Intervening*. Cambridge, U.K.: Cambridge University Press.
- Kennedy, J. & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings IEEE International Conference on Neural Networks*, volume IV (pp. 1942–1948). Piscataway NJ: IEEE.
- Kennedy, J. & Eberhart, R. (2001). *Swarm Intelligence*. San Francisco CA: Morgan Kaufmann.
- Law, A. & Kelton, W. (2000). *Simulation Modeling and Analysis*. New York NY: McGraw-Hill, 3rd edition.
- Mammen, E. & Nandi, S. (2004). Bootstrap and resampling. In J. E. Gentle, W. Härdle, & Y. Mori (Eds.), *Handbook of Computational Statistics* (pp. 467–495). Berlin, Heidelberg, New York: Springer.
- Mayo, D. G. (1983). An objective theory of statistical testing. *Synthese*, 57, 297–340.
- Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago IL: The University of Chicago Press.
- Mayo, D. G. & Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *British Journal for the Philosophy of Science*, (pp. axl003).
- Montgomery, D. C. (2001). *Design and Analysis of Experiments*. New York NY: Wiley, 5th edition.
- Parsopoulos, K. & Vrahatis, M. (2002). Recent approaches to global optimization problems through particle swarm optimization. *Natural Computing*, 1(2–3), 235–306.
- Parsopoulos, K. E. & Vrahatis, M. N. (2004). On the computation of all global minimizers through particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, 8(3), 211–224.

- Rosenbrock, H. (1960). An automatic method for finding the greatest or least value of a function. *Computer Journal*, 3, 175–184.
- Shi, Y. & Eberhart, R. (1999). Empirical study of particle swarm optimization. In P. J. Angeline, Z. Michalewicz, M. Schoenauer, X. Yao, & A. Zalzala (Eds.), *Proceedings of the Congress of Evolutionary Computation*, volume 3 (pp. 1945–1950). Piscataway NJ: IEEE.
- Zoubir, A. M. & Boashash, B. (1998). The bootstrap and its application in signal processing. *Signal Processing Magazine, IEEE*, 15(1), 56–67.