An overview of basic and advanced statistic techniques for calibrating and comparing algorithms

Rubén Ruiz García

INSTITUTO TECNOLÓGICO DE INFORMÁTICA
APPLIED OPTIMIZATION SYSTEMS GROUP
DEPARTMENT OF APPLIED STATISTICS, OPERATIONS RESEARCH AND QUALITY
POLYTECHNIC UNIVERSITY OF VALENCIA, SPAIN

EMAA WORKSHOP, ICELAND
2006

http://www.upv.es/gio
Outline

- Motivation
- Preliminaries
  - Parametric vs. non-parametric
- Experimental design
  - Example
  - Analysis of the results: ANOVA
  - Checking ANOVA assumptions
  - Interactions
- Decision trees
- Conclusions
Motivation

- After two decades of publications and efforts (McGeoch, 1986) we still find the same shortcomings in algorithm experimentation and evaluation as ever.
- Often is difficult, if not impossible, to ascertain which algorithm is the best in a given domain from published results and comparisons.
- Just some examples taken from INFORMS Journal on Computing:
Searching for Good Multiple Recursive Random Number Generators via a Genetic Algorithm

No word about how parameters and operators have been selected

No statistical testing whatsoever
Barrage of tables with average values
A New Genetic Algorithm for the Quadratic Assignment Problem

Improper experimentation for fixing parameters and operators

No statistical testing at all.

After many experiments with moderately sized problems (30 ≤ n ≤ 64) we selected a population size of 100. The number of generations for the concentric tabu was set to max{20π, 1000}. The number of generations for the descent and the simple tabu was set to double these values. We noticed an improvement in the results of the algorithms when the population size is increased (and the number of generations is increased proportionally). However, in order to stay within reasonable run times, we opted to experiment with a fixed population size of 100.
Some key parameters set after running a handful of instances and comparing averages

<table>
<thead>
<tr>
<th>Problem</th>
<th>Best Known</th>
<th>No Genetic</th>
<th>TS/FF</th>
<th>Cohesive</th>
<th>Scrambled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>† ‡ * † ‡ * † ‡ * † ‡ * † ‡ *</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kna30a</td>
<td>88900</td>
<td>20 0 0.45</td>
<td>20 0 0.43</td>
<td>20 0 0.38</td>
<td>20 0 0.32</td>
</tr>
<tr>
<td>Kna33b</td>
<td>91420</td>
<td>20 0 0.44</td>
<td>20 0 0.48</td>
<td>20 0 0.33</td>
<td>20 0 0.31</td>
</tr>
<tr>
<td>Nug50</td>
<td>8124</td>
<td>20 0 0.49</td>
<td>20 0 0.47</td>
<td>20 0 0.37</td>
<td>20 0 0.33</td>
</tr>
<tr>
<td>Tho30</td>
<td>149956</td>
<td>20 0 0.49</td>
<td>20 0 0.46</td>
<td>20 0 0.35</td>
<td>20 0 0.33</td>
</tr>
<tr>
<td>Est32a</td>
<td>130</td>
<td>20 0 0.52</td>
<td>20 0 0.51</td>
<td>20 0 0.35</td>
<td>20 0 0.37</td>
</tr>
<tr>
<td>Est32b</td>
<td>168</td>
<td>20 0 0.43</td>
<td>20 0 0.42</td>
<td>20 0 0.39</td>
<td>20 0 0.30</td>
</tr>
<tr>
<td>Est32c</td>
<td>642</td>
<td>20 0 0.29</td>
<td>20 0 0.28</td>
<td>20 0 0.27</td>
<td>20 0 0.27</td>
</tr>
<tr>
<td>Est32d</td>
<td>200</td>
<td>20 0 0.34</td>
<td>20 0 0.33</td>
<td>20 0 0.28</td>
<td>20 0 0.28</td>
</tr>
<tr>
<td>Est32h</td>
<td>438</td>
<td>20 0 0.34</td>
<td>20 0 0.34</td>
<td>20 0 0.29</td>
<td>20 0 0.29</td>
</tr>
<tr>
<td>Sts36a</td>
<td>9526</td>
<td>6 0.114 0.96</td>
<td>8 0.069 0.91</td>
<td>19 0.005 0.55</td>
<td>16 0.021 0.65</td>
</tr>
<tr>
<td>Sts36b</td>
<td>15852</td>
<td>20 0 0.91</td>
<td>20 0 0.66</td>
<td>20 0 0.81</td>
<td>20 0 0.68</td>
</tr>
<tr>
<td>Sts36c</td>
<td>8239.11</td>
<td>1 0.107 0.96</td>
<td>14 0.024 0.69</td>
<td>14 0.036 0.59</td>
<td>18 0.010 0.66</td>
</tr>
<tr>
<td>Tho40</td>
<td>240516</td>
<td>3 0.034 1.34</td>
<td>3 0.025 1.52</td>
<td>5 0.010 0.98</td>
<td>5 0.015 0.91</td>
</tr>
<tr>
<td>Sio42</td>
<td>15812</td>
<td>20 0 1.66</td>
<td>20 0 0.66</td>
<td>20 0 1.15</td>
<td>20 0 1.20</td>
</tr>
<tr>
<td>Sio49</td>
<td>23396</td>
<td>10 0.032 2.89</td>
<td>14 0.022 2.61</td>
<td>17 0.009 2.13</td>
<td>18 0.007 2.16</td>
</tr>
<tr>
<td>Wlu50</td>
<td>48816</td>
<td>3 0.024 3.08</td>
<td>7 0.014 2.04</td>
<td>18 0.002 1.99</td>
<td>18 0.002 1.85</td>
</tr>
<tr>
<td>Sio56</td>
<td>34458</td>
<td>0 0.041 5.01</td>
<td>0 0.046 4.83</td>
<td>19 0.001 3.24</td>
<td>17 0.002 3.25</td>
</tr>
<tr>
<td>Sio64</td>
<td>48498</td>
<td>3 0.048 9.06</td>
<td>1 0.005 9.06</td>
<td>20 0 5.85</td>
<td>19 0.000 6.01</td>
</tr>
<tr>
<td>Esb64a</td>
<td>116</td>
<td>20 0 3.21</td>
<td>20 0 3.19</td>
<td>20 0 3.05</td>
<td>20 0 3.10</td>
</tr>
<tr>
<td>Sio72</td>
<td>66256</td>
<td>0 0.120 15.76</td>
<td>0 0.115 15.50</td>
<td>10 0.014 8.36</td>
<td>7 0.013 7.74</td>
</tr>
<tr>
<td>Sio81</td>
<td>90998</td>
<td>0 0.124 25.52</td>
<td>0 0.112 25.43</td>
<td>5 0.014 13.30</td>
<td>4 0.019 12.78</td>
</tr>
<tr>
<td>Sio99</td>
<td>115534</td>
<td>0 0.139 41.81</td>
<td>0 0.126 41.63</td>
<td>4 0.011 22.35</td>
<td>2 0.019 19.52</td>
</tr>
</tbody>
</table>

† Number of times out of 20 that best-known solution obtained,
‡ Percentage of average solution over the best-known solution,
* Time in minutes per run,
Comparison among algorithms done similarly !!!

<table>
<thead>
<tr>
<th>Problem</th>
<th>Best Known</th>
<th>Descent</th>
<th>Simple Tabu</th>
<th>Concentric Tabu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kra30a</td>
<td>88900</td>
<td>162</td>
<td>0.253</td>
<td>0.05</td>
</tr>
<tr>
<td>Kra30b</td>
<td>91420</td>
<td>124</td>
<td>0.037</td>
<td>0.05</td>
</tr>
<tr>
<td>Ngu30</td>
<td>6124</td>
<td>160</td>
<td>0.013</td>
<td>0.05</td>
</tr>
<tr>
<td>Tho30</td>
<td>149396</td>
<td>192</td>
<td>0.009</td>
<td>0.05</td>
</tr>
<tr>
<td>Esc32a</td>
<td>130</td>
<td>144</td>
<td>0.569</td>
<td>0.05</td>
</tr>
<tr>
<td>Esc32b</td>
<td>168</td>
<td>200</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Esc32c</td>
<td>649</td>
<td>200</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Esc32d</td>
<td>200</td>
<td>200</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Esc32h</td>
<td>438</td>
<td>200</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Sta56a</td>
<td>9526</td>
<td>49</td>
<td>0.246</td>
<td>0.08</td>
</tr>
<tr>
<td>Sta56b</td>
<td>13852</td>
<td>195</td>
<td>0.015</td>
<td>0.08</td>
</tr>
<tr>
<td>Sta56c</td>
<td>8239.11</td>
<td>73</td>
<td>0.142</td>
<td>0.08</td>
</tr>
<tr>
<td>Tho40</td>
<td>240516</td>
<td>4</td>
<td>0.089</td>
<td>0.13</td>
</tr>
<tr>
<td>Sko42</td>
<td>15612</td>
<td>173</td>
<td>0.014</td>
<td>0.16</td>
</tr>
<tr>
<td>Sko49</td>
<td>21336</td>
<td>2</td>
<td>0.107</td>
<td>0.28</td>
</tr>
<tr>
<td>W150</td>
<td>48816</td>
<td>26</td>
<td>0.038</td>
<td>0.25</td>
</tr>
<tr>
<td>Sko56</td>
<td>34458</td>
<td>63</td>
<td>0.054</td>
<td>0.42</td>
</tr>
<tr>
<td>Sko64</td>
<td>48496</td>
<td>69</td>
<td>0.051</td>
<td>0.73</td>
</tr>
<tr>
<td>Esc44a</td>
<td>116</td>
<td>200</td>
<td>0</td>
<td>0.40</td>
</tr>
<tr>
<td>Sko72</td>
<td>66256</td>
<td>1</td>
<td>0.112</td>
<td>0.93</td>
</tr>
<tr>
<td>Sko81</td>
<td>90906</td>
<td>0</td>
<td>0.087</td>
<td>1.44</td>
</tr>
<tr>
<td>Sko90</td>
<td>115394</td>
<td>3</td>
<td>0.139</td>
<td>2.31</td>
</tr>
<tr>
<td>Sko0a</td>
<td>152002</td>
<td>7</td>
<td>0.114</td>
<td>3.42</td>
</tr>
<tr>
<td>Sko0b</td>
<td>153890</td>
<td>6</td>
<td>0.096</td>
<td>3.47</td>
</tr>
<tr>
<td>Sko0c</td>
<td>147662</td>
<td>2</td>
<td>0.075</td>
<td>3.22</td>
</tr>
<tr>
<td>Sko0d</td>
<td>149517</td>
<td>0</td>
<td>0.137</td>
<td>3.45</td>
</tr>
<tr>
<td>Sko0e</td>
<td>149150</td>
<td>4</td>
<td>0.071</td>
<td>3.31</td>
</tr>
<tr>
<td>Sko0f</td>
<td>149036</td>
<td>1</td>
<td>0.148</td>
<td>3.55</td>
</tr>
<tr>
<td>W1100</td>
<td>273038</td>
<td>0</td>
<td>0.076</td>
<td>3.51</td>
</tr>
</tbody>
</table>

† Number of times out of the corresponding number of runs that best-known solutions obtained.
‡ Percentage of average solution over the best-known solution.
* Time in minutes per run.
Motivation

- Recent examples such as these can be found in many other OR journals where new algorithms and/or techniques are shown.
- Some areas, like for example routing and scheduling are even worse as statistical techniques (even simple paired tests) are scarcely used.
Motivation

- The same old questions:
  - Which design options should I use?
  - Why some options work better than others?
  - Is the performance similar for all types of instances?
  - Am I correctly calibrating my algorithm?
  - Is my algorithm better than competitors?

- ...are still answered incorrectly in most published work

- ...some of them are not even raised or dealt with at all
Motivation

- The result of this is well known (Hooker, 1994, 1995, among many others):
  - Questionable findings, questionable contribution
  - Results almost impossible to reproduce
  - Hardly any possible generalization
  - Vague reports on results
  - No insight on why the proposed methods work
  - No insight on how instance characteristics affect performance
  - No quantification of what parts of the proposed method are actually helping
  - No indication of interactions...
Motivation

- Clearly, we already know enough to put an end to all this
- There is plenty of published papers and reports where all these problems are addressed and where tools are given to avoid them (McGeoch, 1992; Barr et al., 1995; McGeoch, 1996; Rardin and Uzsoy, 2001, Bartz-Beielstein, 2003...
Motivation

- In this talk I will try to overview the basics of correct and sound statistical experimentation.
- It will not be by any means comprehensive...
- ...but it will be really applied with hands-on examples.
- We will skip some important issues.
- I will stress the usage of parametric statistics whenever possible.
- Towards the end I will briefly introduce some advanced statistical techniques.
Preliminaries

- What we usually want:
  - To know is this or that feature of the algorithm we are building is worthwhile \( \text{design} \)
  - To comprehend why something works and why doesn’t, specially when using different instances \( \text{analysis} \)
  - To convince everybody with sound results that our algorithm is better \( \text{comparison} \)
- This triad of questions can be answered with the same tools in a sound statistical way
Preliminaries

- We will work with samples (instances)
- But we want sound conclusions: generalization over a given population (all possible instances)
- Thus we need STATISTICAL INFERENCE
- Very important:
  - Descriptive statistics are nice but one should never infer from a median, average or percentile
  - Sadly, and as we have seen, this is exactly what we find in the literature: “the proposed algorithm is better than algorithm X because it gives better average results on some instances (out of a benchmark of 20)”
Preliminaries
Parametric vs. non-parametric

- As we know:
  - Parametric inferential tests do have some assumptions and requirements on your data
  - This is necessary so that the theoretical statistical models we adopt are appropriate for making inferences
  - Non-parametric tests are “distribution-free”

- Then, Why don’t we just use non-parametric tests?
Preliminaries

Parametric vs. non-parametric

- There are very, very few “completely assumption free” statistical tests
- Non-parametric tests can be too over conservative
  - The differences in the means have to be strong in order to find statistically significant differences
- This might not sound too bad... but digging a little bit more...
Preliminaries
Parametric vs. non-parametric

- We will be contrasting the following hypothesis:
  - $H_0 = \text{There are no differences in the response variable}$

- Truth table:

<table>
<thead>
<tr>
<th>Nature of $H_0$</th>
<th>Hypothesis testing over $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>No reject</td>
</tr>
<tr>
<td>False</td>
<td>Error Type II</td>
</tr>
</tbody>
</table>
Preliminaries
Parametric vs. non-parametric

- Power of a test: 1- _
  - Probability of rejecting $H_0$ when it’s false
  - The power increases, among other things with the sample size

- it’s very difficult to estimate a priori

- It is desired to have a low _, a low _ and a high power
Preliminaries
Parametric vs. non-parametric

- With all this in mind:
  - If the differences in the means are not strong enough the non-parametric tests have very little power
  - This means that we will be having high _:
    - The non-parametric tests tend to not accept $H_0$ when it’s false
    - You will be wrongly answering negatively to the triad of questions!!
Preliminaries
Parametric vs. non-parametric

- Parametric testing:
  - Robust: you really have to depart from the assumptions in order to find trouble
  - If sample is large enough (>100) CLT takes care of many things
  - If the sample is large, using non-parametric makes very little sense…
  - …but interestingly, many significance tests in non-parametric statistics are based on asymptotic (large samples) theory
Preliminaries
Parametric vs. non-parametric

- You really need large data samples...
  - If you really find that your algorithm is a mere 3% better than all other algorithms with very few samples then you have done something wrong or you cannot really generalize
  - Or if you have an algorithm that is a 300% better than all others in a small sample probably you do not need statistics
- ... therefore, after all this the question now is reversed:
  - “Why use non-parametric tests?”
Experimental design

- Among the basic techniques, experimental design can help us answer all the triad of questions
- All other basic questions can also be adequately answered
- Easy to understand, easy to use:

  DESIGN OF EXPERIMENTS (DOE)
Experimental design

- The experimental design is just a few guidelines to carry out the experiments so to obtain results as clearly and as efficiently as possible.
- There are many types of experiments and many associated techniques.
- In my opinion, one does not really need to go far in DOE before reaching our goals.
- Computer experimentation is a very easy environment as far as DOE goes (Bartz-Beielstein, 2003).
Experimental design

- Some special characteristics of computer experiments as far as DOE goes:
  - Reproducibility to the bit (re-using the random seed)
  - Malleable environment in most cases (input can be controlled)
  - A priori knowledge present most times
  - “Cheap” and fast data collection
  - Systematic errors in experimentation are unlikely to occur and easy to avoid
Experimental design

- Response variable: The aim of the experiment; characteristic that we want to study: percentage deviation from optima, time needed to a given solution/quality...

- Controlled Factor: variables, options, parameters that we CAN control and that might affect the response variable
  - Quantitative: Probability of crossover (levels)
  - Qualitative: Type of crossover (variants)
Experimental design

- Treatment: a given combination of the levels/variants of the different controlled factors
- Experience: the execution of a treatment and the associated resulting value of the response variable
- Replicate: when a given treatment is executed more than once
- Non controlled factor: All other factors (known or not) that we can NOT control
Experimental design

- The easiest design is called FULL FACTORIAL
  - All the combinations of levels of all factors are experimented
  - Powerful design
  - Easy analysis of the results
  - Exponential growth on the number of experiences as the number of factors and/or levels grows
  - The results are usually presented in a table
## Experimental design

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Factors</th>
<th>Replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Experimental design

- The order in which the treatments (experiences) are carried out should be RANDOMIZED.
- Probably this is not needed in computer algorithms but memory leaks and in general degradation of computer resources represent a very dangerous lurking variable.
- Lurking variables: non-controlled factors that affect controlled factors in a systematic and consistent way.
- This generates a non-controlled structure in the data, which kills the experimentation.
Experimental design

Example

- Example of a screening experiment
  - Design and calibration of an Iterated Greedy metaheuristic. Application to the permutation flowshop problem (Stützle, Pranzo and Ruiz, in preparation):

```plaintext
S0 = Construct_Initial_Secuence(); How to construct it?
S1 = Local_Search(S0); Do we need local search?

While NOT(TerminationCriterion()) do
  S2 = Partially_Destruct(S1); How to destruct? How much to destruct?
  S3 = Construct_Secuence(S2); How to reconstruct?
  S4 = Local_Search(S3); Do we need local search?
  If Acceptance_Criterion(S4, S1) then S1 = S4 How to accept?
```
Experimental design Example

- **Response variable:**
  - Minimization of the percentage deviation over the best solution known for a set of HARD instances

- **Controlled factors:**
  - Type of initialization (2 variants): heuristic and random
  - Type of destruction (2 variants): random and blocked
Experimental design Example

- Controlled factors (cont):
  - Type of reconstruction (2 variants): greedy and random
  - Application of local search (2 variants): no, yes
  - Acceptance criterion (2 variants): SA, descent
  - Iterations for acceptance (2 levels): 1, 5
  - Number of jobs to destruct (2 levels): 4, 6

- 7 factors at two levels: full factorial of 128 tests
Experimental design Example

- In this case is better to run a half fraction: 27-1=64 treatments: Fractional factorial experiment
  - Resolution VII: allows us to study interactions of three factors with ease

- Very important to consider:
  - 3 groups of instances, 10 instances each= 30 instances
  - All instances have 20 machines and differ in the number of jobs (50, 100 and 200)
  - 5 replicates per treatment

- 64 treatments · 30 instances · 5 replicates = 9600 data

- RANDOMIZE + USE VRT!!
## Experimental design

**Example**

- **Crucial:** Termination criteria set at a maximum elapsed CPU time that depends on the instance \((n \cdot m \cdot 30 \text{ ms})\)

<table>
<thead>
<tr>
<th>IG TEST</th>
<th>Algorithm Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg</td>
<td>Initialization</td>
</tr>
<tr>
<td>44</td>
<td>1</td>
</tr>
<tr>
<td>53</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>52</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
</tr>
</tbody>
</table>
Experimental design
Analysis of the results: ANOVA

- Sir Roland Fisher, 1930
- The ANOVA (analysis of variance) is one the most powerful statistical tools available
- The term ANOVA is a source of confusion: detects differences on means by analyzing the variance!
- The ANOVA is a statistical model where the variation in the response variable is partitioned into components that correspond to the different sources of variation (factors)
Experimental design
Analysis of the results: ANOVA

- Let’s study the results
- ANOVA TABLE

Analysis of Variance for RPD - Type III Sums of Squares

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F-Ratio</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAIN EFFECTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A:Acceptance_C</td>
<td>7,26506</td>
<td>1</td>
<td>7,26506</td>
<td>27,62</td>
<td>0,0000</td>
</tr>
<tr>
<td>B:Destruct</td>
<td>389,076</td>
<td>1</td>
<td>389,076</td>
<td>1479,08</td>
<td>0,0000</td>
</tr>
<tr>
<td>C:Destruction_T</td>
<td>50,0663</td>
<td>1</td>
<td>50,0663</td>
<td>190,33</td>
<td>0,0000</td>
</tr>
<tr>
<td>D:Initialization</td>
<td>60,7802</td>
<td>1</td>
<td>60,7802</td>
<td>231,06</td>
<td>0,0000</td>
</tr>
<tr>
<td>E:Iterations_Acc</td>
<td>393,743</td>
<td>1</td>
<td>393,743</td>
<td>1496,82</td>
<td>0,0000</td>
</tr>
<tr>
<td>F:LS</td>
<td>12444,9</td>
<td>1</td>
<td>12444,9</td>
<td>47309,62</td>
<td>0,0000</td>
</tr>
<tr>
<td>G:n</td>
<td>133,771</td>
<td>2</td>
<td>66,8856</td>
<td>254,27</td>
<td>0,0000</td>
</tr>
<tr>
<td>H:Reconstruction_T</td>
<td>4286,73</td>
<td>1</td>
<td>4286,73</td>
<td>16296,09</td>
<td>0,0000</td>
</tr>
<tr>
<td>I:replicate</td>
<td>0,402254</td>
<td>4</td>
<td>0,100563</td>
<td>0,38</td>
<td>0,8215</td>
</tr>
</tbody>
</table>
Experimental design
Checking ANOVA assumptions

- Before actually starting, we have to check the three main assumptions of the ANOVA: normality, homoscedasticity and independence of the residuals

- Checking normality:
  - Outlier analysis
  - Distribution fitting of the data to a normal distribution, Normal Probability plots...
  - Numerical tests are very strict and normally they will reject the hypothesis that the data comes from a normal distribution
Experimental design
Checking ANOVA assumptions

Normal Probability Plot

- Ooops!
  - Non normality
  - Studies support the fact that ANOVA is very robust wrt to normality
  - Still there is much that we can do
Experimental design
Checking ANOVA assumptions

- Sources of trouble regarding normality:
  - Presence of severe outliers
    - Outliers should not be eliminated as the environment is controlled. Check for bugs or other potential problems in the code
  - Factors or levels with excessive effect
    - There is no need to test what is evident
  - “Clustering”
    - Totally different behavior on the results depending on some levels or factors
Experimental design
Checking ANOVA assumptions

- According to the ANOVA table, the factor LS has a very large effect
- Means plot: a simple plot, usually along with confidence intervals suitable for multiple comparisons:

Means and 99,0 Percent Tukey HSD Intervals

![Diagram](image)
Experimental design
Checking ANOVA assumptions

Normal Probability Plot

- Much better now
- Many transformations possible
- I would not worry unless aberrant plot
Experimental design
Checking ANOVA assumptions

- Checking homocedasticity:
  - Study the dispersion of the residuals with respect to the levels of all factors
    - Some levels or factors might result in higher or lower variance
  - Study the dispersion of the residuals with respect to the values of the response variable
    - Probably higher or lower values of the response variable might result in higher or lower variance
Experimental design

Checking ANOVA assumptions

- No problem
- It has to be repeated for every factor
- Also for the response variable
Experimental design
Checking ANOVA assumptions

- Sources of trouble regarding homocedasticity:
  - A level of a factor resulting in more variance
    - Disregard the level in the experiment
  - More variance in the “hard” instances
    - Separated ANOVAs, one for each group of instances
  - Increased variance as response variable increases (decreases)
    - Properly select the response variable!
Experimental design
Checking ANOVA assumptions

- ANOVA is very sensitive to lack of independence
- Checking independence of the residual:
  - Plot of the dispersion of residuals over run number or time
    - We should expect the residual to be independent from time
  - Analyze the residual looking for self correlation patterns
    - The residual should be “white noise”
Experimental design
Checking ANOVA assumptions

Residual Plot for RPD

- No problem
- Controlled environment: no lurking variables
Experimental design
Checking ANOVA assumptions

- Sources of trouble regarding independence of the residual:
  - Residual bigger over time
    - Experiences run in batch mode, computer resources degrading over time
  - Structure in the residual
    - Randomization or “shuffling” of the experiences
    - ANOVA model NOT complete
Experimental design
Checking ANOVA assumptions

Means and 99.0 Percent Tukey HSD Intervals
Experimental design
Checking ANOVA assumptions

- Checking assumptions:
  - If the experiment is carried out with care...
  - if there are sufficient samples...
  - and if the technique is applied correctly...
  - ... there should not be any problem

- If everything else fails
  - Then use a non-parametric test and hope to obtain something!
Experimental design
Analysis of the results: ANOVA

- With large samples the p-value tends to be close to zero
  - If the sample size is large enough (infinity) any difference in the means of the factors, no matter how small, will be significant
- Real vs. Statistical significance (Montgomery and Runger, 2002)
  - Study factors until the improvement in the response variable is deemed small
Experimental design
Analysis of the results: ANOVA

Analysis of Variance for RPD - Type III Sums of Squares

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F-Ratio</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAIN EFFECTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A:Acceptance_C</td>
<td>1,14754</td>
<td>1</td>
<td>1,14754</td>
<td>10,23</td>
<td>0,0014</td>
</tr>
<tr>
<td>B:Destruct</td>
<td>33,0077</td>
<td>1</td>
<td>33,0077</td>
<td>294,17</td>
<td>0,0000</td>
</tr>
<tr>
<td>C:Destruction_T</td>
<td>0,264526</td>
<td>1</td>
<td>0,264526</td>
<td>2,36</td>
<td>0,1247</td>
</tr>
<tr>
<td>D:Initialization</td>
<td>0,0288163</td>
<td>1</td>
<td>0,0288163</td>
<td>0,26</td>
<td>0,6123</td>
</tr>
<tr>
<td>E:Iterations_Acc</td>
<td>155,4</td>
<td>1</td>
<td>155,4</td>
<td>1384,96</td>
<td>0,0000</td>
</tr>
<tr>
<td>F:n</td>
<td>120,115</td>
<td>2</td>
<td>60,0573</td>
<td>535,25</td>
<td>0,0000</td>
</tr>
<tr>
<td>G:Reconstruction_T</td>
<td>137,248</td>
<td>1</td>
<td>137,248</td>
<td>1223,20</td>
<td>0,0000</td>
</tr>
</tbody>
</table>

- Examine the factors by F-Ratio value:
  - Iterations_Acc, Reconstruction_T, n, Destruct, Acceptance_C
Experimental design
Analysis of the results: ANOVA

Means and 99,0 Percent Tukey HSD Intervals

- Iterations_Acc
- Reconstruction_T
Experimental design
Interactions

- A very interesting feature of the ANOVA is that one can study interactions
- For algorithm design, the most interesting interactions are those between the design options and the characteristics of the instances
- “Short experiments”, “One factor at a time” or even modern racing algorithms (Birattari et al., 2002) do not allow the study of interactions
Experimental design
Interactions

- Let us work with another example (Ruiz et al., in press at C&OR, Thijs and Ruiz, in preparation)
- Heuristics and genetic algorithms for realistic scheduling problems
- 10 controlled factors depicting different characteristics of the instances
- Very large datasets and comprehensive experiments: we want to know why algorithms work
## Experimental design

### Interactions

<table>
<thead>
<tr>
<th>Factor</th>
<th>Small (9,216)</th>
<th>Large (3,072)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of jobs</td>
<td>$n$ 5,7,9,11,13,15</td>
<td>50,100</td>
</tr>
<tr>
<td>Number of stages</td>
<td>$m$ 2, 3</td>
<td>4, 8</td>
</tr>
<tr>
<td>Number of unrelated parallel machines per stage</td>
<td>$mi$ 1, 3</td>
<td>2, 4</td>
</tr>
<tr>
<td>Distribution of the release dates for the machines</td>
<td>$rm$ 0, $\mathbb{U}[1,200]$</td>
<td>0, $\mathbb{U}[1,200]$</td>
</tr>
<tr>
<td>Probability for a job to skip a stage</td>
<td>$F$ 0%, 50%</td>
<td>0%, 50%</td>
</tr>
<tr>
<td>Probability for a machine to be eligible</td>
<td>$E$ 50%, 100%</td>
<td>50%, 100%</td>
</tr>
<tr>
<td>Distribution of the setup times as a percentage of the processing times</td>
<td>$S$ $\mathbb{U}[25,74], \mathbb{U}[75,125]$</td>
<td>$\mathbb{U}[25,74], \mathbb{U}[75,125]$</td>
</tr>
<tr>
<td>Probability for the setup time to be anticipatory %</td>
<td>$A$ $\mathbb{U}[0,50], \mathbb{U}[50,100]$</td>
<td>$\mathbb{U}[0,50], \mathbb{U}[50,100]$</td>
</tr>
<tr>
<td>Distribution of the lag times</td>
<td>$lag$ $\mathbb{U}[1,99], \mathbb{U}[.99,99]$</td>
<td>$\mathbb{U}[1,99], \mathbb{U}[.99,99]$</td>
</tr>
<tr>
<td>Number of preceding jobs</td>
<td>$P$ 0, $\mathbb{U}[1,3]$</td>
<td>0, $\mathbb{U}[1,5]$</td>
</tr>
</tbody>
</table>
Experimental design

Interactions

- Example of a very strong 2-factor interaction:
Experimental design

Interactions

Example of a very strong 3-factor interaction:

![Graph showing interactions and 99.9 Percent LSD Intervals]

**Graph Details:**
- **x-axis:** n values range from 5 to 15.
- **y-axis:** AURPD values range from 0 to 0.03.
- **Legend:**
  - **BGA**
  - **SGAM**
  - **SGAR**
  - **SGA**
- **Comparisons:**
  - **NO PRECEDENCE CONSTRAINTS**
  - **PRECEDENCE CONSTRAINTS**
Experimental design
Interactions

Another example of 2-factor interaction
Decision trees

- In some cases, the nature of the data that we obtain does not allow for a parametric analysis no matter the number of samples.
- A clear example comes from categorized response variables.
- Non-parametric tests (Wilcoxon, Kruskal-Wallis) are very limited as regards the study of interactions.
- Decision trees and Automatic Interaction Detection (AID) tools are non-parametric and at the same time perfect for interaction study.
Decision trees

- AID (Morgan and Sonquist, 1963) recursively bisects experimental data according to one factor into mutually exclusive and exhaustive sets that describe the response variable in the best way. AID works on an interval scaled response variable and maximizes the sum of squares between groups by means of an F statistic.

- We use an improved version called Exhaustive CHAID from Biggs et al. (1991) that allows multi-way splits and significance testing. The result is a decision tree.
Decision trees

- Decision trees are very common in social and health sciences
- I have not seen them applied to algorithm design and calibration
- An example of categorical variable
  - Analysis of the performance of a MIP model on the previous dataset of 9,216 instances. Three different possible results:
    - 0: Optimum solution found within the time limit
    - 1: Time limit reached, solution found
    - 2: Time limit reached, no solution found
Decision trees

- First clumsy attempt: a table with averages

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$m_i$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>Opt</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>Av Time</td>
<td>0.32</td>
<td>2.06</td>
<td>10.47</td>
</tr>
<tr>
<td></td>
<td>%Limit</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>Opt</td>
<td>83.85</td>
<td>85.16</td>
<td>75.26</td>
</tr>
<tr>
<td></td>
<td>Av Time</td>
<td>60.58</td>
<td>99.33</td>
<td>18.31</td>
</tr>
<tr>
<td></td>
<td>%Limit</td>
<td>16.15</td>
<td>14.84</td>
<td>24.74</td>
</tr>
<tr>
<td>9</td>
<td>Opt</td>
<td>60.16</td>
<td>65.36</td>
<td>48.44</td>
</tr>
<tr>
<td></td>
<td>Av Time</td>
<td>124.30</td>
<td>89.95</td>
<td>51.38</td>
</tr>
<tr>
<td></td>
<td>%Limit</td>
<td>39.84</td>
<td>34.64</td>
<td>38.54</td>
</tr>
<tr>
<td>11</td>
<td>Opt</td>
<td>35.68</td>
<td>34.11</td>
<td>28.91</td>
</tr>
<tr>
<td></td>
<td>Av Time</td>
<td>106.81</td>
<td>125.49</td>
<td>140.87</td>
</tr>
<tr>
<td></td>
<td>%Limit</td>
<td>51.56</td>
<td>65.89</td>
<td>45.31</td>
</tr>
<tr>
<td>13</td>
<td>Opt</td>
<td>14.06</td>
<td>20.31</td>
<td>8.85</td>
</tr>
<tr>
<td></td>
<td>Av Time</td>
<td>254.17</td>
<td>146.95</td>
<td>230.03</td>
</tr>
<tr>
<td></td>
<td>%Limit</td>
<td>61.98</td>
<td>73.44</td>
<td>63.54</td>
</tr>
<tr>
<td>15</td>
<td>Opt</td>
<td>1.82</td>
<td>12.24</td>
<td>1.56</td>
</tr>
<tr>
<td></td>
<td>Av Time</td>
<td>492.76</td>
<td>176.77</td>
<td>246.60</td>
</tr>
<tr>
<td></td>
<td>%Limit</td>
<td>71.61</td>
<td>72.40</td>
<td>67.45</td>
</tr>
</tbody>
</table>
Decision trees are much more informative.
Decision trees

- CHAID needs large data samples and many replicates in order to be usable
- It looses power when there are many categories and results are difficult to analyze
- Not a common option in software. SPSS Decision Tree
- Interesting alternative for rank valued results in algorithm comparison
Decision trees

- After analyzing the tree many conclusions on the performance of the model can be obtained
  - This allowed us to detect weak spots that required further modeling
  - We gained a deep understanding of the model and how it could be improved
  - All the conclusions drawn are supported by a sound statistical procedure
Conclusions

- Even today we find inadequate analysis and testing of algorithms
- Parametric statistics pose an interesting and powerful alternative to non-parametric methods
- The DOE procedure and the posterior analysis by means of ANOVA techniques represent a very powerful approach that can be used for comparing performance of different algorithms and to calibrate methods
Conclusions

- Of particular interest is the study of interactions
- Insight on why algorithms work and how different features are affected by the input
- CHAID and decision trees: powerful non-parametric alternative for categorical response variables
- Sound statistical experimentation is a MUST
An overview of basic and advanced statistic techniques for calibrating and comparing algorithms

Rubén Ruiz García

INSTITUTO TECNOLÓGICO DE INFORMÁTICA
APPLIED OPTIMIZATION SYSTEMS GROUP
DEPARTMENT OF APPLIED STATISTICS, OPERATIONS RESEARCH AND QUALITY
POLYTECHNIC UNIVERSITY OF VALENCIA, SPAIN

EMAA WORKSHOP, ICELAND
2006

http://www.upv.es/gio