Five Days of Empirical Software Engineering: The PASED Experience

Massimiliano Di Penta
Giuliano Antoniol
Daniel M. Germán
Yann-Gaël Guéhéneuc
Bram Adams
Motivation

Empirical background important for graduate students

Courses on statistics insufficient to provide such a background

(Most) University curricula could not afford to have specific courses

Some exceptions (there are others for sure!):

- Easterbrook’s CSC2130: Empirical Research Methods in Software Engineering at the University of Toronto (2009)
- Dewayne E. Perry course, Univ. of Texas at Austin
So students need that!

1. *Can my boyfriend come along?*

2. *I'm not your boyfriend! You totally are. I'm casually dating a number of people.*

3. *But you spend twice as much time with me as with anyone else. I'm a clear outlier.*

4. *Your math is irrefutable. Face it—I'm your statistically significant other.*
General Info
About the School

• Ecole Polytechnique de Montréal, June 2011
• Funded by MITACS
  • low fee for students $250 all included
• 44 participants, 9 countries and 25 different institutions
• More on http://pased.soccerlab.polymtl.ca
Learning Objectives

1. plan and conduct software engineering experiments with human subjects and collect related data

2. plan and conduct software engineering studies involving the mining of data from (un)structured software repositories

3. build prediction and classification models from the collected data, and to use these models
Challenges

Choosing Topics

Dealing with heterogeneous participants

Combining theory and practice
What Topics?
What Topics?

Experiment Design

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lab 1</strong></td>
<td>Sys A</td>
<td>Sys A</td>
<td>Sys B</td>
<td>Sys B</td>
</tr>
<tr>
<td><strong>Lab 2</strong></td>
<td>Sys B</td>
<td>Sys B</td>
<td>Sys A</td>
<td>Sys A</td>
</tr>
</tbody>
</table>

Planning the study
What Topics?

Planning the study

Getting the data

Experiment Design

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab 1</td>
<td>Sys A</td>
<td>Sys A</td>
<td>Sys B</td>
<td>Sys B</td>
</tr>
<tr>
<td>Lab 2</td>
<td>Sys B</td>
<td>Sys B</td>
<td>Sys A</td>
<td>Sys A</td>
</tr>
</tbody>
</table>

Mining Software Repositories

- software complexity
- fault prediction
- effort estimation
- change propagation
- change patterns
- evolution visualization

Knowledge inference
Data visualization
Association clustering
What Topics?

Experiment Design

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab 1</td>
<td>Sys A</td>
<td>Sys A</td>
<td>Sys B</td>
<td>Sys B</td>
</tr>
<tr>
<td>Lab 2</td>
<td>Sys B</td>
<td>Sys B</td>
<td>Sys A</td>
<td>Sys A</td>
</tr>
</tbody>
</table>

Getting the data

Statistical Analyses

- Wilcoxon test:
  - Significant difference in Exp II (p-value=0.01).
  - Significant differences on the whole set (p-value<0.01).

- Effect size:
  - Exp b 0.63 (medium)
  - Exp II 1.26 (large)
  - Exp III ~ 0 (low)

Overall, we can reject $H_0$.

Planning the study

Mining Software Repositories

- software complexity
- fault prediction
- effort estimation
- change propagation
- change patterns
- evolution visualization

Analyzing results
Only 5 days....

...that's too much!
The Approach

- **Learn by example** and by doing format
  - Experiment design principles and statistics introduced by presenting cases from studies in literature
- **Practical application** of theoretical concepts during labs
- Course material and laboratory packages available online, including course videos
## School Content

<table>
<thead>
<tr>
<th>AM</th>
<th>Mining Software Archives</th>
<th>Exp. Design</th>
<th>Text mining</th>
<th>Statistical Analysis</th>
<th>Predictor Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>Keynote</td>
<td>Keynote</td>
<td>Keynote</td>
<td>Keynote</td>
<td>Keynote</td>
</tr>
<tr>
<td>PM</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
</tr>
<tr>
<td>PM</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
<td>Hands on lab</td>
</tr>
</tbody>
</table>
Learning by doing...
Running example I

• **Use of UML Stereotypes in comprehension and maintenance tasks**


  • Filippo Ricca, Massimiliano Di Penta, Marco Torchiano, Paolo Tonella, Mariano Ceccato: The Role of Experience and Ability in Comprehension Tasks Supported by UML Stereotypes. ICSE 2007: 375-384

• In the following briefly referred as “**Conallen**”
## Experiment Design

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lab 1</strong></td>
<td>Sys A</td>
<td>Sys A</td>
<td>Sys B</td>
<td>Sys B</td>
</tr>
<tr>
<td><strong>Lab 2</strong></td>
<td>Sys B</td>
<td>Sys B</td>
<td>Sys A</td>
<td>Sys A</td>
</tr>
</tbody>
</table>
### Data format: example

<table>
<thead>
<tr>
<th>Exp</th>
<th>Subject</th>
<th>System</th>
<th>Method</th>
<th>Lab</th>
<th>Ability</th>
<th>Precision</th>
<th>Recall</th>
<th>FMeasure</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T1</td>
<td>WFMS</td>
<td>Conallen</td>
<td>1 h</td>
<td>0.59</td>
<td>0.6</td>
<td>0.55</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T1</td>
<td>Claros</td>
<td>UML</td>
<td>2 h</td>
<td>0.79</td>
<td>0.78</td>
<td>0.77</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T10</td>
<td>Claros</td>
<td>Conallen</td>
<td>1 l</td>
<td>0.64</td>
<td>0.54</td>
<td>0.58</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T10</td>
<td>WFMS</td>
<td>UML</td>
<td>2 l</td>
<td>0.82</td>
<td>0.73</td>
<td>0.77</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T11</td>
<td>WFMS</td>
<td>Conallen</td>
<td>1 l</td>
<td>0.79</td>
<td>0.65</td>
<td>0.7</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T11</td>
<td>Claros</td>
<td>UML</td>
<td>2 l</td>
<td>0.76</td>
<td>0.8</td>
<td>0.74</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T12</td>
<td>Claros</td>
<td>Conallen</td>
<td>1 h</td>
<td>0.78</td>
<td>0.92</td>
<td>0.82</td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T12</td>
<td>WFMS</td>
<td>UML</td>
<td>2 h</td>
<td>0.48</td>
<td>0.54</td>
<td>0.47</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T13</td>
<td>WFMS</td>
<td>UML</td>
<td>1 l</td>
<td>0.38</td>
<td>0.2</td>
<td>0.25</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T13</td>
<td>Claros</td>
<td>Conallen</td>
<td>2 l</td>
<td>0.67</td>
<td>0.57</td>
<td>0.61</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T2</td>
<td>WFMS</td>
<td>Conallen</td>
<td>1 h</td>
<td>0.63</td>
<td>0.63</td>
<td>0.61</td>
<td>118</td>
<td></td>
</tr>
</tbody>
</table>
Boxplots: Conallen
Paired analysis: example

> wilcox.test(F.Conallen, F.UML, paired = TRUE,
  alternative = "greater")

Wilcoxon signed rank test with continuity correction

data: F.Conallen and F.UML

V = 138, p-value = 0.04354

alternative hypothesis: true location shift is greater than 0

- Must have data in a paired format
- or you can use a proper R script
- Need to remove subjects that took part to one lab only
  - For parametric statistics just replace `wilcox.test` with `t.test`
Hands on Labs

- **Mining software repository challenge**: extract interesting facts from git
- **Experiment design**: groups working together on designing a study
- **Data analysis**: text mining, analyze working data sets of previous experiments and build bug predictors
  - Working data sets from previous experiments, PROMISE data sets
- **Tools**: R and Weka
#UNPAIRED ANALYSIS

#Analysis of single experiment
#Mann-Whitney
attach(tbn)
wilcox.test(Correct[Fit=="yes"],Correct[Fit=="no"],paired=FALSE,alternative="greater")

attach(ttrento)
wilcox.test(Correct[Fit=="yes"],Correct[Fit=="no"],paired=FALSE,alternative="greater")

attach(tphd)
wilcox.test(Correct[Fit=="yes"],Correct[Fit=="no"],paired=FALSE,alternative="greater")

#All data
attach(t)
wilcox.test(Correct[Fit=="yes"],Correct[Fit=="no"],paired=FALSE,alternative="greater")

#Exercises:
# 1) perform a two-tailed test
# 2) can t-test be applied instead of Wilcoxon test? test for data normality using the Wilk-Shapiro test
# 3) repeat the analysis using the t-test?
# 4) repeat the analysis for the time dependent variable
Calibrating Courses to Participants’ Profiles

**Statistical Analyses**
- Excellent: 1
- Good: 21
- Basic: 23
- None: 1

**Empirical Sw Engineering**
- Excellent: 4
- Good: 20
- Basic: 18
- None: 2

**Mining Sw repositories**
- Excellent: 4
- Good: 16
- Basic: 24
- None: 1

**Machine learning**
- Excellent: 1
- Good: 12
- Basic: 25
- None: 7
Feedbacks

Guidelines on what not to do

Longer Labs

Tutorials on tools

How to write empirical papers
Acknowledgments

- Lecturers (other than the paper authors):
  - Ahmed E. Hassan, Queen’s University, Canada
  - Andrian Marcus, Wayne State University, USA
- Keynote Speakers
  - Gail Murphy, University of British Columbia, USA
  - Prem Devanbu, UC Davis, USA
  - Alain Picard, Benchmark Consulting Services Inc., Canada
  - Maria Codipiero, Peter Colligan, Kal Murtaia, SAP, Canada
  - Marc-André Decoste, Google Montréal, Canada
- Student volunteers
- The attendees!
- MITACS ([http://www.mitacs.ca/](http://www.mitacs.ca/))
Conclusions
Challenges

- Choosing Topics
- Dealing with heterogeneous participants
- Combining theory and practice

What Topics?

- Experiment Design
- Getting the data
- Statistical Analyses

The Labs

- **Mining software repository challenge**: extract interesting facts from git
- **Experiment design**: groups working together on designing a study
- **Data analysis**: text mining, analyze working data sets of previous experiments and build bug predictors
  - Working data sets from previous experiments, PROMISE data sets
  - Tools: R and Weka

Feedbacks

- **Guidelines on what not to do**
- **Longer Labs**

- **Tutorials on tools**
- **How to write empirical**