The credit for creating these slides belongs to Fall 2014 CS 521/621 students. Student names have been removed per FERPA regulations.
Coverage Is Not Strongly Correlated with Test Suite Effectiveness
Authors

Laura Inozemtseva and Reid Holmes
University of Waterloo
Waterloo, ON, Canada
Test Suites

What determines whether a test suite is good? Most developers assume that enough code coverage means more bugs will be caught.

Where did this assumption come from?
Correlation

Many previous studies have shown that code coverage and effectiveness of test suites are correlated.

The strength of correlation changes between each study.
Key Idea

Analyze previous studies to determine where the difference in results occur.

Introduces three new questions to determine the relationship between the effectiveness, size, and coverage of test suites.
Previous Studies

Most fail to include programs of larger size
All found some correlation between effectiveness and code coverage
Independent of test suite size
Very little agreement between strength of correlation
Research Questions

Is the effectiveness of a test suite correlated:

- with the number of test cases in the suite?
- with its coverage when the number of test cases in the suite is ignored?
- with its coverage when the number of test cases in the suite is held constant?
Methodology

Faulty version of programs
Generate test suites
Measure the coverage of suites
Determine the effectiveness of suites

Measure the coverage and effectiveness of suites to evaluate the relationship
Faulty Versions of the Programs

Select projects:
- Reasonably large
- In java
- Have large number of test methods

Generate faulty programs:
- Open source tool PIT
Generate Test Suites

Master suite
Randomly selecting a subset of the master suite
Different sizes
Measuring coverage

Open source tool CodeCover

- Statement
- Decision
- Modified condition
Measuring effectiveness

Mutation analysis

Two effectiveness measurements:

● Raw effectiveness measurement
● Normalized effectiveness measurement
Results

Is Size Correlated With Effectiveness?
Results

There is a moderate to very high correlation between the effectiveness of a test suite and the number of test methods it contains.
Results

Is Coverage Correlated With Effectiveness when Size Is Ignored?

Table 3: The Kendall $\tau$ correlation between normalized effectiveness and different types of coverage when suite size is ignored. All entries are significant at the 99.9% level.

<table>
<thead>
<tr>
<th>Project</th>
<th>Statement</th>
<th>Decision</th>
<th>Mod. Cond.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache POI</td>
<td>0.75</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>Closure</td>
<td>0.83</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>HSQLDB</td>
<td>-0.35</td>
<td>-0.35</td>
<td>-0.35</td>
</tr>
<tr>
<td>JFreeChart</td>
<td>0.50</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>Joda Time</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 4: The Kendall $\tau$ correlation between non-normalized effectiveness and different types of coverage when suite size is ignored. All entries are significant at the 99.9% level.

<table>
<thead>
<tr>
<th>Project</th>
<th>Statement</th>
<th>Decision</th>
<th>Mod. Cond.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache POI</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>Closure</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>HSQLDB</td>
<td>0.81</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>JFreeChart</td>
<td>0.91</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>Joda Time</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Results

There is a moderate to high correlation between the effectiveness and the coverage of a test suite when the influence of suite size is ignored.
Results

Is Coverage Correlated With Effectiveness when Size Is Fixed?
Results
Correlation between coverage and effectiveness drops with controlled suite size
Ranges from low to moderate
Not generally safe to assume correlation
Stronger correlation with non-normalized effectiveness measurement
Type of coverage had little influence on the strength of the relationship
Conclusions
Questions

These studies assume that any mutant not detected by the test suite is equivalent, why would this be?
Questions

Wouldn’t assuming that any mutant not detected by the master test suite alter the results?
Questions

According to the paper, a suite with more coverage will “almost certainly” kill more mutants; are there cases where this is not true?
Questions

What if mutation score is not a good metric for this problem? And what is the possible solution?
Questions

How can developers benefit from the results of this paper?
Sources


