METHODS FOR TESTING UNIFORMITY STATISTICS

KRISHNA PATEL AND ROBERT M. HIERONS
BRUNEL UNIVERSITY – BRUNEL SOFTWARE ENGINEERING LAB (BSEL)

DAVID CLARK AND HECTOR D. MENENDEZ
UNIVERSITY COLLEGE LONDON – CREST CENTRE
Definitions

• **Uniform Distribution**: A sample is said to adhere to a uniform distribution if every element in the sample has an equal chance of being randomly selected.

• **Uniformity Statistic**: A Uniformity Statistic is a means of measuring the extent to which a sample conforms to a uniform distribution.
  - The Uniformity Statistics considered in our research produce lower values for samples that adhere more strongly to a uniform distribution.
Problem Definition

• Uniformity Statistics have the oracle problem, because it is very difficult to predict the outcome.

• We investigated three different approaches for alleviating the oracle problem in uniformity statistics.
Intuition

• The standard deviation of a sample is a measure of the spread of values in that sample.
• Higher measures of standard deviations indicate that the values in the sample are more spread out, and thus the sample should adhere more strongly to a uniform distribution.
• Thus, the standard deviation is intrinsically linked to uniformity.
• All of our oracles are based on this observation.
Intuition Behind a Metamorphic Relation

Sample with Higher SD → Uniformity Statistic → Statistic Value (A)

Sample with Lower SD → Uniformity Statistic → Statistic Value (B)

Compare

B < A → Fail

A < B → Pass
Intuition Behind Regression Model Oracles (1)

• For each uniformity statistic, we performed a Regression Analysis to learn the precise nature of the relationship between the standard deviation and test statistic value.

• For a given test statistic, the Regression Analysis enabled us to derive a mathematical formula that accepts a standard deviation value as input and outputs a predicted test statistic value.
Intuition Behind Regression Model Oracles (2)

• Plot Statistic (Black) and Model (Grey), against standard deviation, based on 10000 samples.

• Applied one Mann-Whitney U Test per subject program to compare the statistic and model, and applied Benjamini-Hochberg correction to these tests. 14/18 of the statistics did not report a significant result.

• Most models are indistinguishable.
Intuition Behind Regression Model Oracles (3)

Sample → Uniformity Statistic → Statistic Value → Compare → Pass

Compare:
- Similar enough → Pass
- Too dissimilar → Fail

Fail → Model → Model Value
Intuition behind Metamorphic Regression

Model Oracles

Sample with Higher SD

Uniformity Statistic

Absolute Difference

Uniformity Statistic

Sample with Lower SD

Fail

Too dissimilar

Similar enough

Compare

Pass

Model

Absolute Difference

Model
Experimental Design – Subject Programs

- **Subject Programs:** 18 Uniformity Statistics – $D_n^+$, $D_n^-$, $V_n$, $W_n^2$, $U_n^2$, $C_n^+$, $C_n^-$, $C_n$, $K_n$, $T_1$, $T_2$, $T_1'$, $T_2'$, $G(n)$, $Q$, $S_n^{(m)}$, $A^*(n)$, $E_{m,n}$

- **Code Reuse:**
  - $V_n$ reuses $D_n^+$ and $D_n^-$
  - $U_n^2$ reuses $W_n^2$
  - $C_n$ reuses $C_n^+$ and $C_n^-$
  - $K_n$ reuses $C_n^+$ and $C_n^-$
  - $Q$ reuses $G(n)$
Experimental Design – Mutants

- Mutmut mutation testing tool.
- Removed equivalent mutants.
- Removed crashed mutants.
- 196 mutants in total.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Number Of Mutants</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{m,n}$</td>
<td>19</td>
</tr>
<tr>
<td>$G(n)$</td>
<td>14</td>
</tr>
<tr>
<td>$K_n$</td>
<td>1(+9+8)</td>
</tr>
<tr>
<td>$Q$</td>
<td>12(+14)</td>
</tr>
<tr>
<td>$S^{(m)}_n$</td>
<td>14</td>
</tr>
<tr>
<td>$T_1$</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Number Of Mutants</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T'_1$</td>
<td>14</td>
</tr>
<tr>
<td>$T_2$</td>
<td>14</td>
</tr>
<tr>
<td>$T'_2$</td>
<td>16</td>
</tr>
<tr>
<td>$U^2_n$</td>
<td>6(+20)</td>
</tr>
<tr>
<td>$V_n$</td>
<td>1(+7+5)</td>
</tr>
<tr>
<td>$W^2_n$</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Number Of Mutants</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A^*(n)$</td>
<td>24</td>
</tr>
<tr>
<td>$C_n$</td>
<td>0(+9+8)</td>
</tr>
<tr>
<td>$C^-_n$</td>
<td>9</td>
</tr>
<tr>
<td>$C^+_n$</td>
<td>8</td>
</tr>
<tr>
<td>$D^-_n$</td>
<td>7</td>
</tr>
<tr>
<td>$D^+_n$</td>
<td>5</td>
</tr>
</tbody>
</table>
Experimental Design – Test Suites

• **Mutation Testing Test Suites:**
  • We generated one test suite per oracle, by random testing.
  • These test suites consist of 100 test cases.
  • Test cases in these test suites could either deterministically report false positives, or deterministically not report false positives.
    • Metamorphic Regression Model Oracle had one such test case – this was replaced to prevent false positives from confounding the results.

• **False Positive Rate Test Suites:**
  • We generated one test suite per oracle, by random testing.
  • Each test suite consisted of 1000 test cases.
Results and Discussion – Mutation Score

- MR – 77/196, RMO – 159/196, and MRMO – 119/196
- Fisher’s Exact Tests + Benjamini-Hochberg Correction = Significant Difference
- MRMO is probably more effective than MR because of tightness
- RMO is probably more effective than MRMO because:
  - RMO was less aggressively tuned
  - MRMO is blind to faults that cause the same level of difference between the source and follow-up test case, whilst RMO is not
Results and Discussion – Failure Detection Rate

- RMO obtained an FDR of 100% for 137/159 killed mutants
- MR obtained an FDR of 100% in 52/77 killed mutants
- MRMO obtained an FDR of 100% for 40/119 killed mutants
- Mann-Whitney U Tests + Benjamini-Hochberg Correction = Significant
- Interesting: MR is more effective than MRMO in terms of FDR
Results and Discussion – False Positive Rate

• False positives arise from:
  • Statistics can make errors and this could result in false positives
  • The models used in the RMO and MRMO oracles could make inaccurate predictions

• MR reports 0 false positives in all subject programs

• The largest false positive rates that were observed for RMO and MRMO across all subject programs is:
  • MRMO: 0.40%
  • RMO: 0.40%
Future Work

• A Genetic Algorithm based test case selection methodology that attempts to maximise the difference between the statistic and the models for the RMO oracle.

• The RMO and MRMO oracles both require tuning before they can be used. A method that circumvents this requirement would improve the usability of these techniques.
Thank you for listening. Are there any questions?