An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Old Dominion Univ.

March 1998

Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing

Combining Representative and Directed Tests
Scenario
Experimental Design
Results

Open Questions & Future Directions
Goals:

- **Model Flexibility**
  - reliability estimates using either representative or directed test data
  - tolerance of “normal” variations in data
- **Improved Data Collection**
  - reduced noise
  - integrating multiple sources of information
1. Overview of Order Statistic Model
2. Experiment: combining representative and directed tests
Order Statistic Model

Assumptions
Testing as Biased Selection
Order Statistics and Testing

Combining Representative and Directed Tests
Scenario
Experimental Design
Results

Open Questions & Future Directions
Assumptions

Basic idea:

- The faults present in a program have operational failure rates randomly selected from a distribution $F$.
  - $F$ depends upon
    - program structure and semantics
    - operational input distribution
- Test methods tend to find the largest faults first.
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing
Combining Representative and Directed Tests
Scenario
Experimental Design
Results
Open Questions & Future Directions

Detailed assumptions

- Operational program failure rate between repairs is constant.
- Faults manifest independently.
- Detected faults are repaired perfectly.
- The testing process is biased towards early detection of faults with the largest failure rates.
- Faults in a program have failure rates $\phi$ with distribution $F(\phi)$.
- The program contains a finite number of faults.
Comparing Assumptions

Our assumptions are actually less restrictive than most existing RGMs

- Operational program failure rate between repairs is constant.
- Faults manifest independently.
- Detected faults are repaired perfectly.
- The testing process is biased towards early detection of faults with the largest failure rates.
- Faults in a program have failure rates $\phi$ with distribution $F(\phi)$.
- The program contains a finite number of faults.

- ▶ Operational program failure rate between repairs is constant.
- ▶ Faults manifest independently.
- ▶ Detected faults are repaired perfectly.
- ▶ The testing process is biased towards early detection of faults with the largest failure rates.
- ▶ Faults in a program have failure rates $\phi$ with distribution $F(\phi)$.
- ▶ The program contains a finite number of faults.
Comparing Assumptions

Our assumptions are actually less restrictive than most existing RGMs

- Operational program failure rate between repairs is constant.
- *Operational program failure rate between repairs is constant.*
- Faults manifest independently.
- *Faults manifest independently.*
Comparing Assumptions

Our assumptions are actually less restrictive than most existing RGMs

- Detected faults are repaired perfectly.
- *Detected faults are repaired perfectly and instantaneously.*
Comparing Assumptions

Our assumptions are actually less restrictive than most *existing RGMs*

- The testing process is biased towards early detection of faults with the largest failure rates.
- *The test process finds the faults in decreasing order of failure rate.*
Comparing Assumptions

Our assumptions are actually less restrictive than most existing RGMs

- Faults in a program have failure rates $\phi$ with distribution $F(\phi)$.
- Faults $f_i$ in a program have failure rates $\phi_i$ whose expected value is a monotonic non-increasing function $g_{\alpha,\beta}(i)$. 
Comparing Assumptions

Our assumptions are actually less restrictive than most existing RGMs

- The program contains a finite number of faults.
- The program contains a finite number of faults, or
- The program contains an infinite number of faults.
Testing as Biased Selection

- **Representative** testing tends to find largest (failure rate) faults first
  - But is not, as usually assumed, *guaranteed* to do so
  - Existing RGMs may be sensitive to these permutations
- **Directed** testing is often assumed to find faults in arbitrary order
  - No evidence for this assumption
  - Counter-intuitive
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Order Statistic Model Assumptions Testing as Biased Selection Order Statistics and Testing Combining Representative and Directed Tests Scenario Experimental Design Results Open Questions & Future Directions

Ordered Directed Testing Property

We propose the following as more likely:

For a given directed testing method, as we approach coverage of the method, the set of $k$ faults revealed will be the $k$ faults with the largest individual operational failure rates.

- expresses a trend, not a guarantee
- The data used in this study supports this property at more than 97% level of significance.
Order Statistics and Testing

- Suppose that a program contains \( n \) faults with failure rates \( \phi_i \).
- Sort these into ascending order.
  
  Let \( \phi_{j:n} \) denote the \( j^{\text{th}} \) smallest of these failure rates. 
  
  \( \phi_{j:n} \) is called the \( j^{\text{th}} \) order statistic of the set of \( n \) failure rates.
Order Statistic Distributions

- If the fault failure rates are governed by an underlying distribution with
  - probability density \( f(x) \) and
  - cumulative distribution \( F(x) \),
- then their order statistics are distributed according to

\[
f_{r:n}(x) = r \binom{n}{r} F_{r-1}(x) f(x) (1 - F(x))^{n-r}
\]
If testing has found the $k$ faults with the largest $\phi$’s, the program failure rate $\lambda$ can be estimated as

$$\lambda = 1 - \prod_{i=1}^{n-k} (1 - E(\phi_i:n))$$

or, for small $\phi$’s

$$\lambda = \sum_{i=1}^{n-k} E(\phi_i:n)$$
Why Order Statistics?

**Representative Testing:** Even if faults are found out of order, the explicit sorting of $\phi$’s eventually corrects this permutation.

**Directed Testing:**
- If a directed test method tends to capture all faults with sufficiently large $\phi$,
- then the sorted $\phi$’s can be fitted to the RGM at coverage.
- Suggests that testers should “climb a subsumption hierarchy”
Advantages of the OS Model

- Can be used with representative or directed methods
- Can be used with both
  - program failure rate data, and
  - fault failure rate data
- Provides an alternative to time-to-first-failure (TTFF) collection
  - TTFF is inherently noisy
  - and becomes increasingly expensive
- Automatically corrects minor permutations in detection order
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing
Brian Mitchell and Steven J Zeil

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing

Combining Representative and Directed Tests
Scenario Experimental Design Results

Open Questions & Future Directions

Typical TTFF Data
Combining Representative and Directed Tests

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing

Combining Representative and Directed Tests
Scenario
Experimental Design
Results

Open Questions & Future Directions
Purely Representative Testing

- Prior work with purely representative TTFF data suggested the OS model is competitive with existing models “on their own turf”.
- One study with purely directed fault failure rate data showed higher error in predictions than hoped.
- This study examines combination of the program and fault failure rate data.
Scenario

- Begin testing with representative tests, collecting TTFF data
- When TTFF exceeds 1000 executions, switch to directed testing
- In this study, used Knowledge-Driven Functional Testing (KDFT)
  - a form of equivalence partitioning
  - augmented with special cases drawn from an knowledge base of “expert” test info.
Experimental Design

- Launch-intercept code
- faults were isolated
- $10^6$ representative cases run
  - manifestations of each fault counted to estimate $\phi_i$
- KDFT test cases defined
  - 100 tests run under each case
  - manifestations of each fault counted to estimate prob of detection under directed test
The failure rate data collected this way is “better” than would be achieved under the scenario

- no order permutations
- lower noise in representative data
We therefore simulated 4 different debugging sequences:

- for test numbers $t = 1, 2, \ldots$
  - for each fault $f$
    - if $f$ had not manifested on a prior test and $\text{rand()} < \phi_f$
      - then
    - mark test $t$ as failed due to fault $f$

This procedure restores the expected exponential distribution on failure times.
After a run of 1000 tests without failures, a similar procedure was employed using the directed test rates in place of the representative failure rates.

- Simulated choosing up to 5 tests per KDFT category
- Beginning halfway through each simulated test run,
  - least-squares fits computed for Jelinski-Moranda (Musa basic), Musa/Okomoto logarithmic Poisson, and Order Statistic models
  - Predictions made of next time-to-failure.
Results

- average relative error in predictions
- parameter progression
Example of $T_{TFF}$ Data
Example of RGMS

![Graph showing failure rate vs. failure number for different methods: Representative, JM, and ML. The graph has a logarithmic scale on the y-axis and a linear scale on the x-axis. The data points are represented by diamonds, and the lines for each method are shown in different colors: Representative in blue, JM in green, and ML in cyan. The x-axis represents the failure number ranging from 0 to 35, and the y-axis represents the failure rate ranging from 1e-06 to 1. The graph shows a decreasing trend in failure rate as the failure number increases.]
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing

Combining Representative and Directed Tests

Scenario
Experimental Design

Results

Open Questions & Future Directions

Adding Directed Data

![Graph showing failure rate and failure number for representative and directed tests.](image)
Final Model Fits

An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing
Combining Representative and Directed Tests
Scenario Experimental Design Results
Open Questions & Future Directions

Final Model Fits

![Graph showing model fits for different failure rates and failure numbers.](graph.png)
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing
Combining Representative and Directed Tests
Scenario
Experimental Design
Results
Open Questions & Future Directions

Final Model Fits

Failure Rate vs. Failure Number

LSq Fit - Sequence 1

Representative
Directed
OS
JM
ML

0 5 10 15 20 25 30 35 40
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing

Combining Representative and Directed Tests
Scenario
Experimental Design
Results

Open Questions & Future Directions

Final Model Fits
An Experiment in Estimating Reliability Growth Under Both Representative and Directed Testing

Brian Mitchell and Steven J Zeil

Order Statistic Model
Assumptions
Testing as Biased Selection
Order Statistics and Testing
Combining Representative and Directed Tests
Scenario
Experimental Design
Results
Open Questions & Future Directions

Final Model Fits

![Graph showing failure rate versus failure number for different testing methods. The graph includes Representative, Directed, Order Statistic (OS), JM, and ML methods. The y-axis represents the failure rate on a logarithmic scale, ranging from 1e-06 to 1, and the x-axis represents the failure number ranging from 0 to 35. The graph shows the LSq Fit - Sequence 4 with different markers for each testing method.]
Predictive Accuracy

On all four debugging sequences, the Order Statistic model showed the lowest average relative error in predictions.

<table>
<thead>
<tr>
<th>Model</th>
<th>JM</th>
<th>ML</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>21.9155</td>
<td>18.43406</td>
<td>15.24542</td>
</tr>
<tr>
<td>Set 2</td>
<td>15.91654</td>
<td>8.871714</td>
<td>7.095273</td>
</tr>
<tr>
<td>Set 3</td>
<td>15.46484</td>
<td>13.69872</td>
<td>11.33527</td>
</tr>
<tr>
<td>Set 4</td>
<td>17.21274</td>
<td>13.63313</td>
<td>10.99466</td>
</tr>
</tbody>
</table>
Parameter Progression

- Debugging sequence 1 caused the most problems for all three models
  - includes an early out-of-sequence fault
- Overall, Musa logarithmic model appeared most sensitive to permutations in the order of detection
Open Questions & Future Directions

- Appropriate distribution $F$ for the $\phi$’s?
- Effects of different mixes of representative and directed testing?
- Effectiveness of various techniques for estimating $\phi$?
- Need more experience applying to realistic projects.