Softwaretechnik / Software-Engineering

Lecture 16: Testing

Prof. Dr. Andreas Podelski, Dr. Bernd Westphal
Albert-Ludwigs-Universität Freiburg, Germany

Introduction and Vocabulary

Limits of Software Testing

Glass-Box Testing

Statement-, branch-, term- coverage.

Other Approaches

Model-based testing,
Runtime verification.

Software quality assurance in a larger scope.

Program Verification

Partial and total correctness,
Proof System PD.

Review

VL 15
VL 16
VL 17
VL 18...

Recall: Test Case

Definition. A test case \( T \) is a pair \((I_n, S_{\mathrm{oll}})\) consisting of

- a description \( I_n \) of sets of finite input sequences,
- a description \( S_{\mathrm{oll}} \) of expected outcomes,

and an interpretation \( \llbracket \cdot \rrbracket \) of these descriptions.

A test execution \( \pi \), i.e. \((\pi_0, \ldots, \pi_n) \downarrow \Sigma \) in \( I_n \) for some \( n \in \mathbb{N}_0 \), is called

- successful (or positive) if it discovered an error, i.e., if \( \pi \not\in \llbracket S_{\mathrm{oll}} \rrbracket \).
- unsuccessful (or negative) if it did not discover an error, i.e., if \( \pi \in \llbracket S_{\mathrm{oll}} \rrbracket \).

Software Examination (in Particular Testing)

In each examination, there are two paths from the specification to results:

- the production path (using model, source code, executable, etc.), and
- the examination path (using requirements specifications).

A check can only discover errors on exactly one of the paths.
If difference detected:

- examination result is positive.
What is not on the paths, is not checked; crucial:

- specification and comparison.

Recall: Checking procedure shows no error
reports error
artefact has error

<table>
<thead>
<tr>
<th>Specification</th>
<th>Implement</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>/</td>
<td></td>
<td>/</td>
</tr>
<tr>
<td>✘</td>
<td>/</td>
<td>✘</td>
</tr>
</tbody>
</table>

(true positive, true negative, false positive, false negative)

(Ludewig and Lichter, 2013)
Why Can’t We Show The Absence of Errors (in General)?

Recall: “Software testing can be used to show the presence of bugs, but never to show their absence!” (E. W. Dijkstra, 1970)

Consider a simple pocket calculator for adding 8-digit decimals:

\[
\begin{array}{cccc}
12345678 & + & 27 & = \\
7890456 & + & 1234567 & = \\
\end{array}
\]

• Requirement: If the display shows \(x\), \(+\), and \(y\), then after pressing \(=\), the sum of \(x\) and \(y\) is displayed if \(x + y\) has at most 8 digits, otherwise “-E-” is displayed.

• With 8 digits, both \(x\) and \(y\) range over \([0, 10^8 - 1]\).

• Thus there are \(10^{16}\) possible input pairs \((x, y)\) to be considered for exhaustive testing!

• And if we restart the pocket calculator for each test, we do not know anything about problems with sequences of inputs...

   (Local variables may not be re-initialised properly, for example.)

Observation: Software Usually Has Many Inputs

• Example: Simple Pocket Calculator. With ten thousand different test cases (that’s a lot!), \(9,999,999,999,990,000\) of the \(10^{16}\) possible inputs remain uncovered.

   In other words: Only \(0.0000000001\%\) of the possible inputs are covered, \(99.99999999\%\) not touched.

   In diagrams:

   (red: untested, blue: tested)

Conclusion

• Question 1: If we cannot consider all test cases, are there clever choices of test cases?
• Software is (in general) not continuous.

• Consider a continuous function, e.g. \( \varepsilon, \delta \). For sufficiently small \( \varepsilon \)-environments of an input, the outputs differ only by a small amount \( \delta \).

• For software, adjacent inputs may yield arbitrarily distant output values.

Vocabulary:
• Range error: multiple "neighbouring" inputs trigger the error.
• Point error: an isolated input value triggers the error.

• For Software, we can (in general) not conclude from some values to others:

• For example, if a bridge endures a single car of 1000 kg, we strongly expect the bridge to hold a single person of 100 kg.

• If the pocket calculator is correct for 12345678 + 27, we can (in general) not expect anything on the other numbers.

Conclusion Cont'd

• Question 1: If we cannot consider all test cases, are there clever choices of test cases?

• Question 2: If we cannot conclude from few test cases to all inputs, when should we stop testing?

When To Stop Testing?

• There need to be defined criteria for when to stop testing; project planning should consider these criteria (and previous experience).

• Possible "testing is done" criteria:
  • all (previously) specified test cases have been executed with negative result, (Special case: All test cases resulting from a certain strategy, like maximal statement coverage (→ in a minute) have been executed.)
  • testing effort time sums up to \( x \) (hours, days, weeks),
  • testing effort sums up to \( y \) (any other useful unit),
  • \( n \) errors have been discovered,
  • no error has been discovered during the last \( z \) hours (days, weeks) of testing.

Values for \( x, y, n, z \) are fixed based on experience, estimation, budget, etc.

• Of course: not all criteria are equally reasonable or compatible with each testing approach.

Another Criterion

• Another possible "testing is done" criterion:

  • The average cost per error discovery exceeds a defined threshold \( c \).

    \[
    \text{cost per discovered error} = \frac{\text{cost}}{\text{number of discovered errors}}
    \]

  • Value for \( c \) is again fixed based on experience, estimation, budget, etc.

Limits of Software Testing

• Software examination paths
  • Is exhaustive testing feasible?

• Range- vs. point errors

• When To Stop Testing?

• Choosing Test Cases

• Requirements on test cases

• The natural habitat of many errors

• Test Oracle

• Glass-Box Testing

• Statement coverage

• Branch and term coverage

• Conclusions from coverage measures

• Model-Based Testing

• Testing in the Development Process
Choosing Test Cases

A test case is a good test case if discovers — with high probability — an unknown error.

An ideal test case (Soll) would be:

• of low redundancy, i.e. it does not test what other test cases also test.
• error sensitive, i.e. has high probability to detect an error, (Probability should at least be greater than 0.)
• representative, i.e. represent a whole class of inputs, (i.e., software S passes (In, Soll) if and only S behaves well for all In′ from the class)

The wish for representative test cases is particularly problematic:

• In general, we do not know which inputs lie in an equivalence class wrt. a certain error. (Recall: point errors.)
• Yet there is a large body on literature on how to construct representative test cases, assuming we know the equivalence classes.
Still, it is perfectly reasonable to test representatives of equivalence classes induced by the specification, e.g.

• valid and invalid inputs (to check whether input validation works at all),
• different classes of inputs considered in the requirements, like "buy water", "buy soft-drink", "buy tea" vs. "buy beverage".

Recall: strive to have at least one test case per feature.

Lion and Error Hunting

Hunting errors in software is (basically) the same.

Some traditional popular belief on software error habitat:

• Software errors (in contrast to lions?) (seem to) enjoy range boundaries, e.g.
  • 0, 1, 27 if software works on inputs from [0, 27],
  • -1, 28 for error handling,
  • $-2^{31}-1, 2^{31}$ on 32-bit architectures,
  • boundaries of arrays (first, last element),
  • boundaries of loops (first, last iteration),
  • special cases of the problem (empty list, use-case without actor, . . . ),
  • special cases of the programming language semantics,
  • complex implementations.

→ Good idea: for each test case, note down why it has been chosen.
  For example, "demonstrate that error handling is not completely broken".

Where Do We Get The "Soll"-Values From?

Recall: A test case is a pair (In, Soll) with proper expected (or "soll") values.

• In an ideal world, all "soll"-values are defined by the (formal) requirements specification and effectively pre-computable.
• In the this world,
  • the formal requirements specification may only reflectively describe acceptable results without giving a procedure to compute the results.
  • there may not be a formal requirements specification, e.g.
    • "the game objects should be rendered properly",
    • "the compiler must translate the program correctly",
    • "the notification message should appear on a proper screen position",
    • "the data must be available for at least 10 days",
    • etc.

Then: need another instance to decide whether the observation is acceptable.

• The testing community prefers to call any instance which decides whether results are acceptable a (test) oracle.

I'd prefer not to call automatic derivation of "soll"-values from a formal specification an "oracle". . .

("person or agency considered to provide wise and insightful [ ... ] prophetic predictions or precognition of the future, inspired by the gods. " says Wikipedia)

Glass-Box Testing: Coverage
Statement/branch coverage canonically extends to test suite coverage, and test suite coverage also looks critical. In programming languages with short-circuit evaluation, it is never evaluated.

In the following, we assume that execution of a (possibly infinite) set $S$ is a finite description of a (possibly infinite) set $S$.

In programming languages with short-circuit evaluation, it is never evaluated. In the program, if and only if $\text{term } x, y, z$ then $\text{if condition}$, and $\text{if condition}$.

Theorem Coverage

$\text{Term Coverage}$
Conclusions from Coverage Measures

• Assume, test suite \( T \) tests software \( S \) for the following property \( \phi \):
  
  • pre-condition: \( p \),
  • post-condition: \( q \),

and \( S \) passes (!) \( T \), and the execution achieves 100 % statement / branch / term coverage.

What does this tell us about \( S \)? Or: what can we conclude from coverage measures?

• 100 % statement coverage:
  
  • "there is no statement, which necessarily violates \( \phi \)"
  
  (Still, there may be many, many computation paths which violate \( \phi \), and which just have not been touched by \( T \).)

  • "there is no unreachable statement"

• 100 % branch (term) coverage:
  
  • "there is no single branch (term) which necessarily causes violations of \( \phi \)"

  In other words: "for each condition (term), there is one computation path satisfying \( \phi \) where the condition (term) evaluates to true, and one for false."

  • "there is no unused condition (term)"

Not more (→ exercises)!

That's definitely something, but not as much as "100 %" may sound like. . .

Coverage Measures in Certification

• (Seems that) DO-178B, "Software Considerations in Airborne Systems and Equipment Certification" (which deals with the safety of software used in certain airborne systems) requires that certain coverage measures are reached, in particular something similar to term coverage (MC/DC coverage).

  (Next to development process requirements, reviews, unit testing, etc.)

• If not required, ask: what is the effort / gain ratio?

  (Average effort to detect an error; term coverage needs high effort.)

• Currently, the standard moves towards accepting certain verification or static analysis tools to support (or even replace?) some testing obligations.

Limits of Software Testing

• Software examination paths

  • Is exhaustive testing feasible?

  • Range- vs. point errors

When To Stop Testing?

• Choosing Test Cases

  • Requirements on test cases

  • The natural habitat of many errors

  • Test Oracle

  • Glass-Box Testing

  • Statement coverage

  • Branch and term coverage

  • Conclusions from coverage measures

  • Model-Based Testing

Testing in the Development Process

• Statistical Tests

  • Classical statistical testing is another approach to deal with

  • in practice not exhaustively testable huge input space,

  • tester bias.

  (People tend to choose "good-will" inputs and disregard corner-cases; recall: the developer is not a good tester.)

  Procedure:

  • Randomly (!) choose test cases \( T_1, \ldots, T_n \) for test suite \( T \).

  • Execute test suite \( T \).

  • If an error is found:

    • good, we certainly know there is an error,

    • if no error is found:

      • refuse hypothesis "program is not correct" with a certain significance niveau.

      (Significance niveau may be unsatisfactory with small test suites.)

      (And: Needs stochastical assumptions on error distribution and truly random test cases.)

Statistical Testing: Discussion

• (Ludewig and Lichter, 2013) name the following objections against statistical testing:

  • In particular for interactive software, the primary requirement is often no failures are experienced by the "typical user".

  Statistical testing (in general) may also cover a lot of "untypical user behaviours" unless (sophisticated) user-models are used.

  • Statistical testing needs a method to compute "soll"-values for the randomly chosen inputs.

  That is easy for requirement "does not crash", but can be difficult in general.

  • There is a high risk for not finding point or small-range errors.

  If they live in their "natural habitat", carefully crafted test cases would probably uncover them.

  Findings in the literature can at best be called inconclusive.
Testing in the Development Process

• The final implementation is running on (prototype) hardware which is connected by its standard input/output interface (e.g. CAN-bus).

Model-Based Testing

• Hardware-in-the-loop

Conclusions

• true
• q

Hardware coverage term and Branch

• SOFT
• ¬ q

Glass-Box Testing

• send C50
• send E1

Oracle Test

• send pSOFT
• ←−−

The final implementation is examined using a separate computer to simulate other system components.

Software-in-the-loop

• pSOFT
• −→←−

Software examination paths

• send C50
• 1 q

Limits of Software Testing

• send C50
• 5 q

Oracle

• chg-C50
• ¬ q

Check whether for 'drink_ready', more interaction is necessary.

Is exhaustive testing feasible?

• SOFT
• ¬ 4 q

Software examination paths

• send C50
• 2 q

Check whether for 'have_c50', 'have_c100', 'have_c150'?

• send C50
• 3 q

Check for 'have_e1' by

• send C50
• 1 q

Check for 'drink_ready', more interaction is necessary.

• send C50
• 1 q

Check for 'have_c50', 'have_c100', 'have_c150'?

• send C50
• 3 q

Check for 'idle', 'have_c50', 'have_c100', 'have_c150'?

• send C50
• 5 q

Check for 'have_e1' by

• send C50
• 2 q

Check for 'drink_ready', more interaction is necessary.
Planning
Preparation
Execution
Evaluation
Analysis

Test Plan
Test Cases
Test Directions
Test Gear
Test Protocol
Test Report

(Ludewig and Lichter, 2013)

• Test Gear: (may need to be developed in the project!)

  test driver—A software module used to invoke a module under test and, often, provide test inputs, control and monitor execution, and report test results.

  Synonym: test harness.

  IEEE 610.12 (1990)

  stub—
  (1) A skeletal or special-purpose implementation of a software module, used to develop or test a module that calls or is otherwise dependent on it.
  (2) A computer program statement substituting for the body of a software module that is or will be defined elsewhere.

  IEEE 610.12 (1990)

• Roles:
tester and developer should be different persons!

Tell Them What You’ve Told Them...

• A check can only discover errors on exactly one path.

• Software testing is challenging because
  • typically huge input space,
  • software is non-continuous,

• Define criteria for “testing done” (like coverage, or cost per error).

• There is a vast amount of literature on how to choose test cases.

  A good starting point:
  • at least one test case per feature,
  • corner-cases, extremal values,
  • error handling, etc.

• Glass-box testing
  • considers the control flow graph,
  • defines coverage measures.

• Other approaches:
  • statistical testing,
  • model-based testing,
  • runtime verification (assert!)

• Process: tester and developer should be different persons.

References


