CSE 503

Software Engineering

Winter 2021

Software Testing

January 29, 2021

Revisiting the Relationship Between Fault Detection, Test Adequacy Criteria, and Test Set Size

Yiqun T. Chen, Rahul Gopinath, Anita Tadakamalla, Michael D. Ernst,
Reid Holmes, Gordon Fraser, Paul Ammann, René Just

@yc_yc_yc_yc

Share your thoughts on this presentation and paper with #ASE2020

UNIVERSITY of WASHINGTON · CISPA · GEORGE MASON UNIVERSITY · UBC · UNIVERSITÄT PASSAU

How to assess the fault detection capacity of a test set?

Test set adequacy

Statement Coverage

```
double avg(double[] nums) {
  int n = nums.length;
  double sum = 0;
  for(int i=0; i<n; ++i) {
    sum += nums[i];
  }
  return sum * n;
}
```

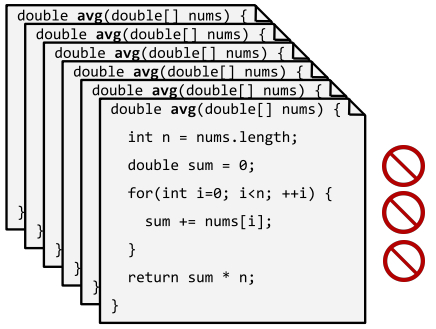
How to assess the fault detection capacity of a test set?

Test set adequacy

Statement Coverage

```
double avg(double[] nums) {
  int n = nums.length;
  double sum = 0;
  for(int i=0; i<n; ++i) {
    sum += nums[i];
  }
  return sum * n;
}
```

Mutation Score



How to assess the fault detection capacity of a test set?

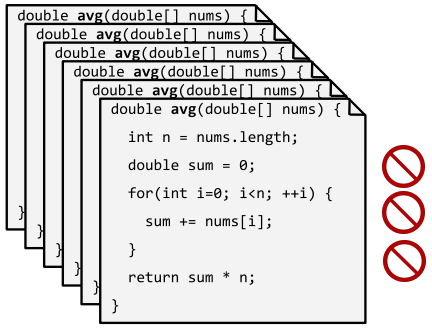
Test set adequacy

Test set size

Statement Coverage

Mutation Score

```
double avg(double[] nums) {
  int n = nums.length;
  double sum = 0;
  for(int i=0; i<n; ++i) {
    sum += nums[i];
  }
  return sum * n;
}
```



How to assess the fault detection capacity of a test set?

Test set adequacy

Test set size

Statement Coverage

Mutation Score

Is test set adequacy a good proxy for fault detection?

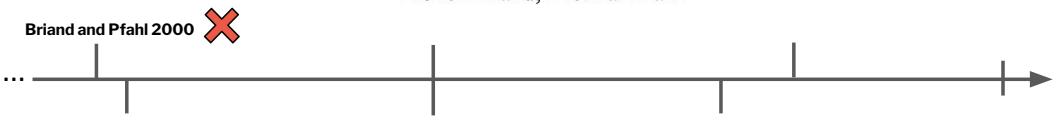
Is test set adequacy contributing beyond just size?

Which adequacy measure is the best?

Is test set adequacy correlated with fault detection?*

Using Simulation for Assessing the Real Impact of Test Coverage on Defect Coverage

Lionel Briand, Dietmar Pfahl



* Taking test set size into account

Is test set adequacy correlated with fault detection?*

The Influence of Size and Coverage on Test Suite Effectiveness

Akbar Siami Namin James H. Andrews

* Taking test set size into account

Is test set adequacy correlated with fault detection?*

Coverage Is Not Strongly Correlated with Test Suite Effectiveness

Laura Inozemtseva and Reid Holmes

* Taking test set size into account

Is test set adequacy correlated with fault detection?*

Code Coverage for Suite Evaluation by Developers

Rahul Gopinath Carlos Jensen Alex Groce

* Taking test set size into account

Is test set adequacy correlated with fault detection?*

Are Mutants a Valid Substitute for Real Faults in Software Testing?

René Just†, Darioush Jalali†, Laura Inozemtseva*, Michael D. Ernst†, Reid Holmes*, and Gordon Fraser‡

* Taking test set size into account

Is test set adequacy correlated with fault detection?*

Are Mutation Scores Correlated with Real Fault Detection?

A Large Scale Empirical study on the Relationship Between Mutants and Real Faults

Mike Papadakis Donghwan Shin

Shin Yoo Doo-Hwan Bae

* Taking test set size into account

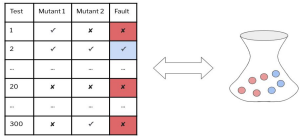
Is test set adequacy correlated with fault detection?*

And many other papers...!

* Taking test set size into account

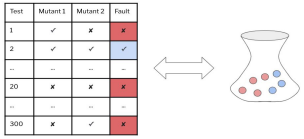
Outline

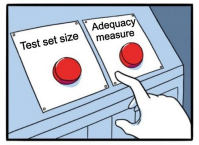
- Review of existing methods



Outline

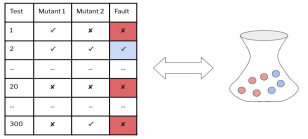
- Review of existing methods
- Ask the right (statistical) question

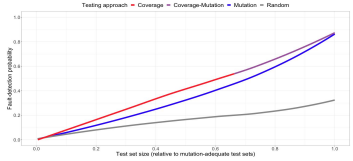




Outline

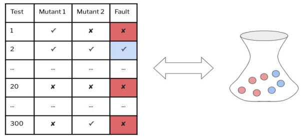
- Review of existing methods
- Ask the right (statistical) question
- Test adequacy measures are valid

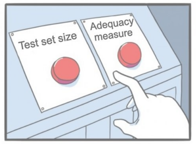


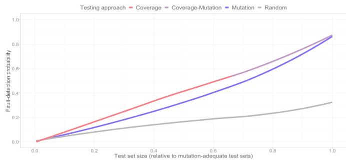


Outline

- Review of existing methods
- Ask the right (statistical) question
- Test adequacy measures are valid







One possible approach: Random selection

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

- **Random Selection**
  - Generate many test sets by **sampling** from an **existing pool**
  - Focus of our talk
- Alternatives DO exist

One possible approach: Random selection

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

- **Random Selection**
  - Generate many test sets by **sampling** from an **existing pool**
  - Focus of our talk
- Alternatives DO exist

One possible approach: Random selection

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

- **Random Selection**
  - Generate many test sets by **sampling** from an **existing pool**
  - Focus of our talk
- Alternatives DO exist

One possible approach: Random selection

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

- **Random Selection**
  - Generate many test sets by **sampling** from an **existing pool**
  - Focus of our talk
- Alternatives DO exist

Random Selection methodology

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

Test set 1

Test	Mutant 1	Mutant 2	Fault
2	✓	✓	✓
20	✘	✘	✘

Test set	Mutation score	Fault detection
1	1.0	✓

Sample n=2 tests from the test pool **without replacement**, and **analyze** the **results** for **different n**.

Random Selection methodology

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

Test set 1

Test	Mutant 1	Mutant 2	Fault
2	✓	✓	✓
20	✘	✘	✘

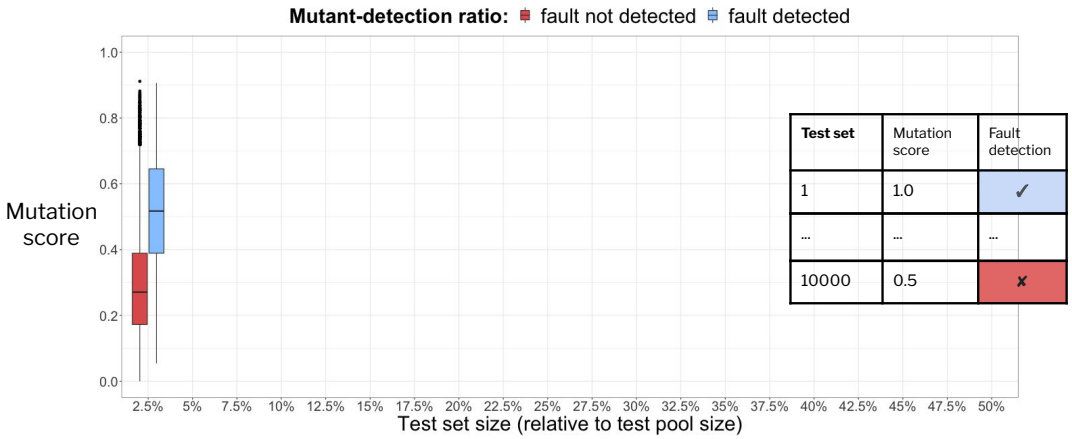
Test set 10000

Test	Mutant 1	Mutant 2	Fault
1	✘	✘	✘
300	✘	✓	✘

Test set	Mutation score	Fault detection
1	1.0	✓
...	...	...
10000	0.5	✘

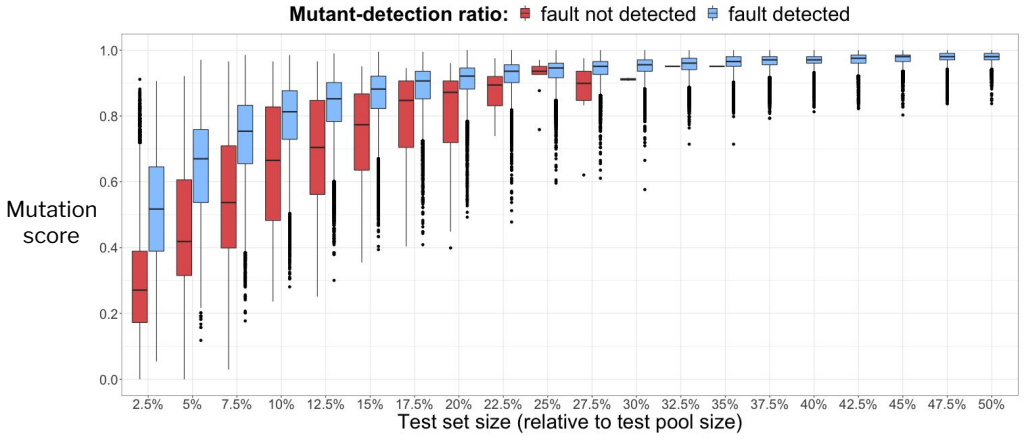
Sample n=2 tests from the test pool **without replacement**, and **analyze** the **results** for **different n**.

Case study: Closure-100 (Defects4J)

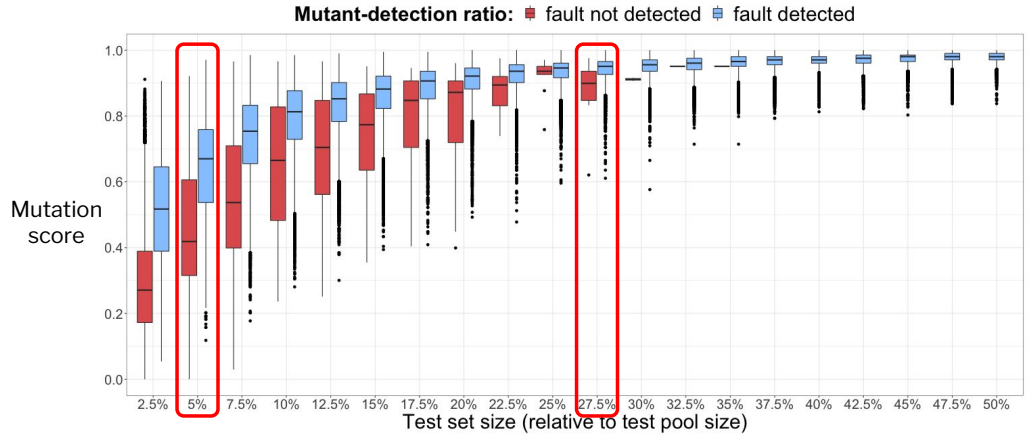


Test set	Mutation score	Fault detection
1	1.0	✓
...	...	...
10000	0.5	✘

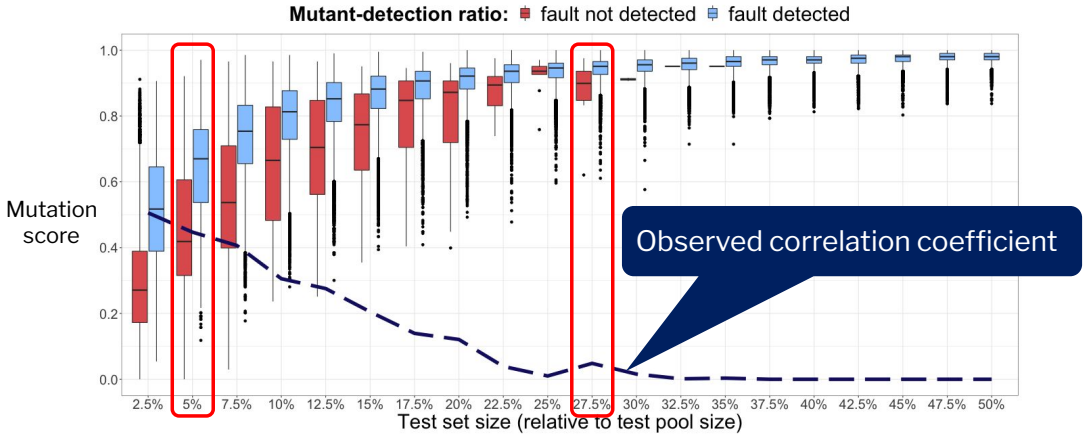
Case study: Closure-100 (Defects4J)



Case study: Closure-100 (Defects4J)

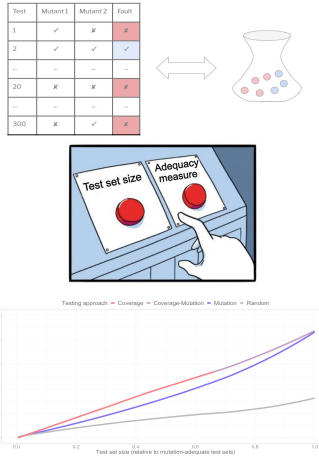


Case study: Closure-100 (Defects4J)



Outline

- Review of existing methods
- Ask the right (statistical) question
  - ill-posed question
  - mis-interpretation of correlation
- Test adequacy measures are valid



Random selection is prone to misleading conclusions!

An ill-posed question

Q: What are the **individual contributions** of **size and adequacy** to fault detection?

A: Impossible to answer when adequacy and size are **highly correlated**.

Random selection is prone to misleading conclusions!

An ill-posed question

Q: What are the **individual contributions** of **size and adequacy** to fault detection?

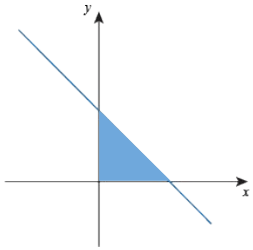
A: Impossible to answer when adequacy and size are **highly correlated**.

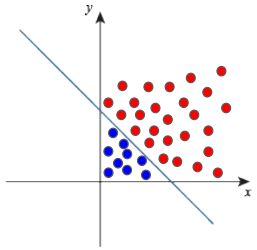
- Encode the same information
  - (Hypothetical) adequacy = size

100 x **size** + 0 x adequacy
=
0 x size + **100** x **adequacy**

A little digression

How would you compute the area under the curve?





A little digression

What's the probability of ...

A little digression

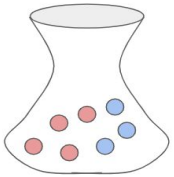
What's the probability of ...

A little digression

What's the probability of … observing two Hs and two Ts (regardless of order)?

A little digression

What's the probability of … selecting 1 blue ball, when selecting 2 balls (without replacement)?

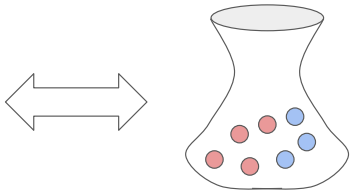


Why does Random Selection fall into this ill-posed question trap?

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
…	…	…	…
20	✘	✘	✘
…	…	…	…
300	✘	✓	✘

Why does Random Selection fall into this ill-posed question trap?

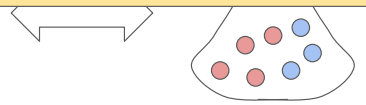
Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘



Why does Random Selection fall into this ill-posed question trap?

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

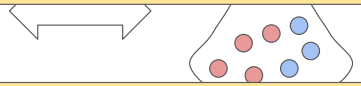
Probability of selecting a fault detecting test set
(1) is a **function** of **test set size**, and (2) has an **analytical form**



Why does Random Selection fall into this ill-posed question trap?

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

Probability of selecting a fault detecting test set
(1) is a **function** of **test set size**, and (2) has an **analytical form**



The same holds for **each mutant!**

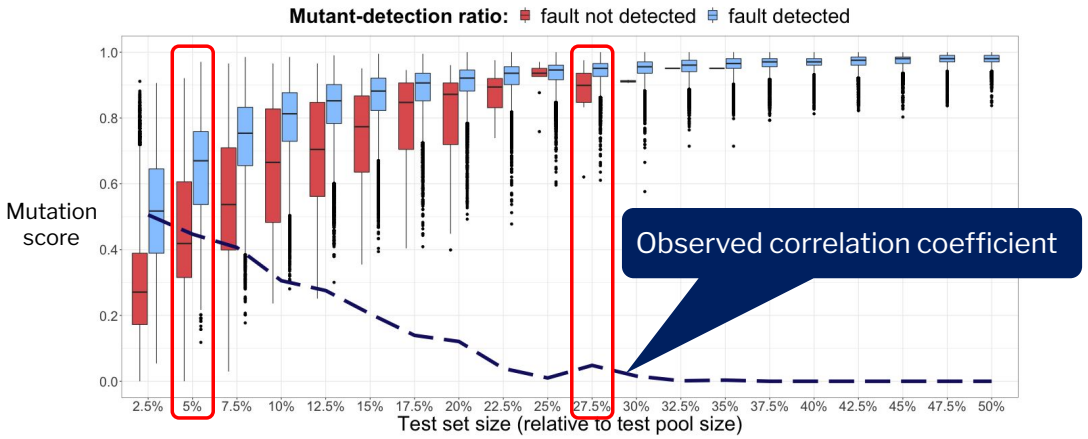
Random Selection implies the ill-posed question!

Larger test sets -> more fault detection

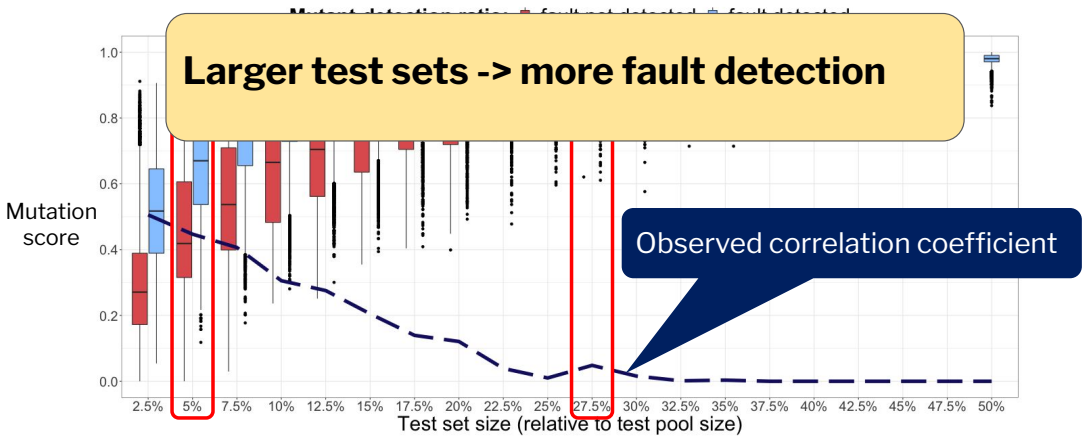
Larger test sets -> higher mutation score

High pairwise correlation as a result!

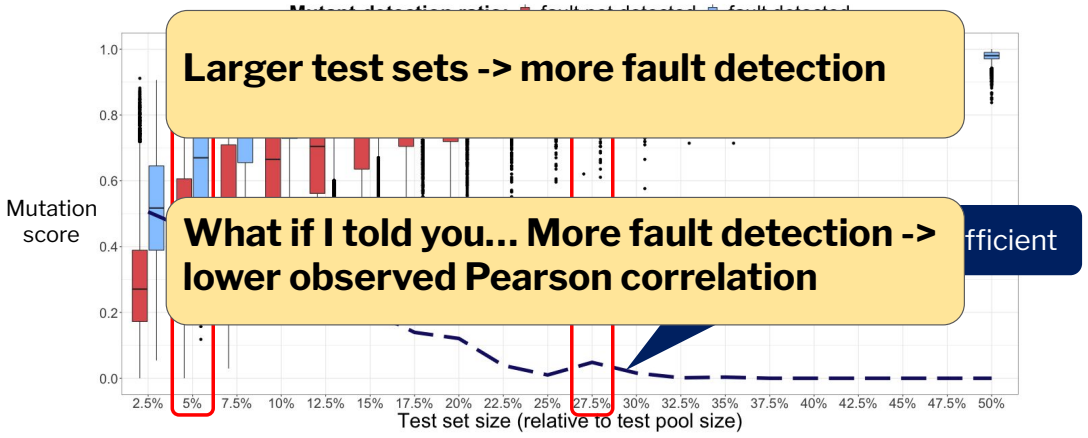
Revisit case study: mis-interpreted Pearson correlation



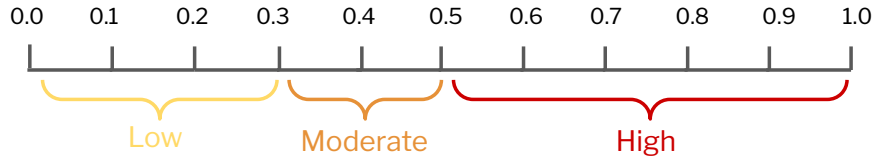
Revisit case study: mis-interpreted Pearson correlation



Revisit case study: mis-interpreted Pearson correlation

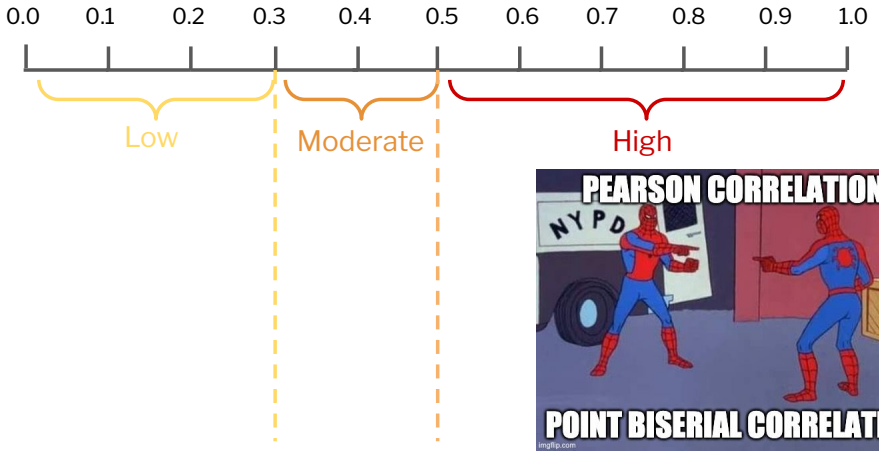


How we usually interpret Pearson correlation*

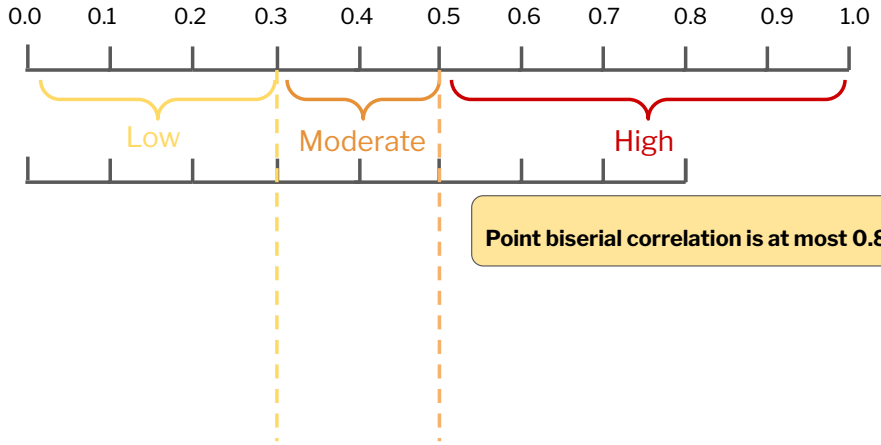


*Cohen (1988)

Fun Facts about Point Biserial Correlation



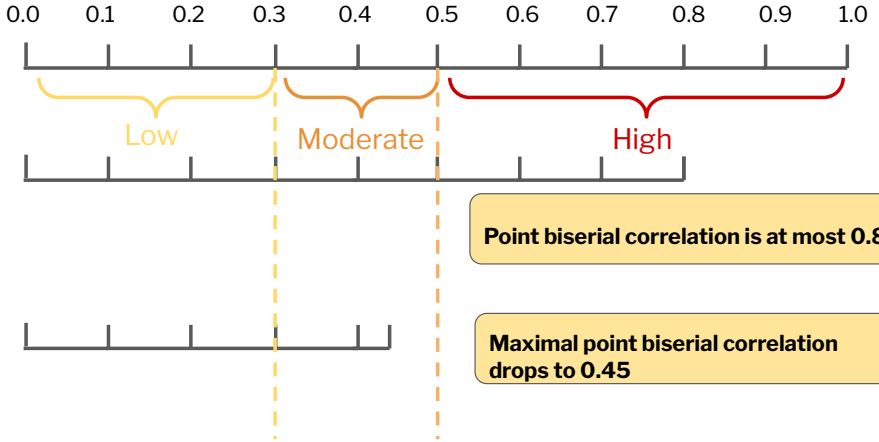
Fun Facts about Point Biserial Correlation



Fault detection
50.0%

Point biserial correlation is at most 0.8

Fun Facts about Point Biserial Correlation



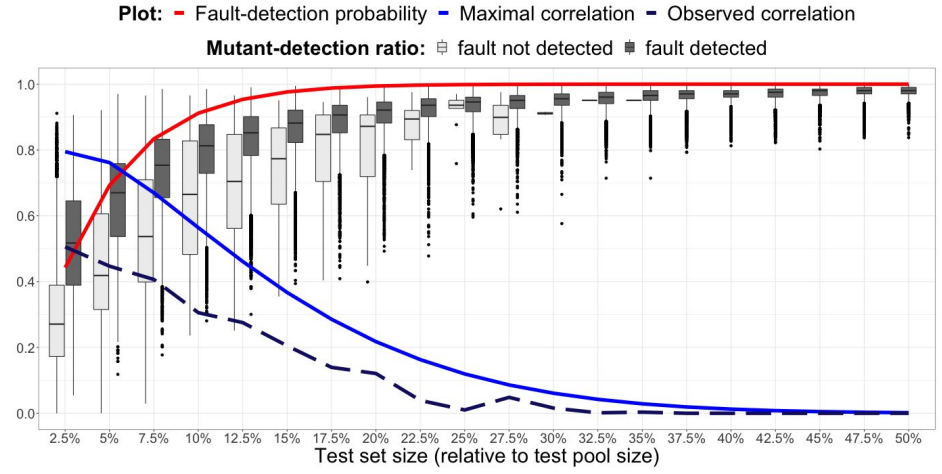
Fault detection
50.0%

Point biserial correlation is at most 0.8

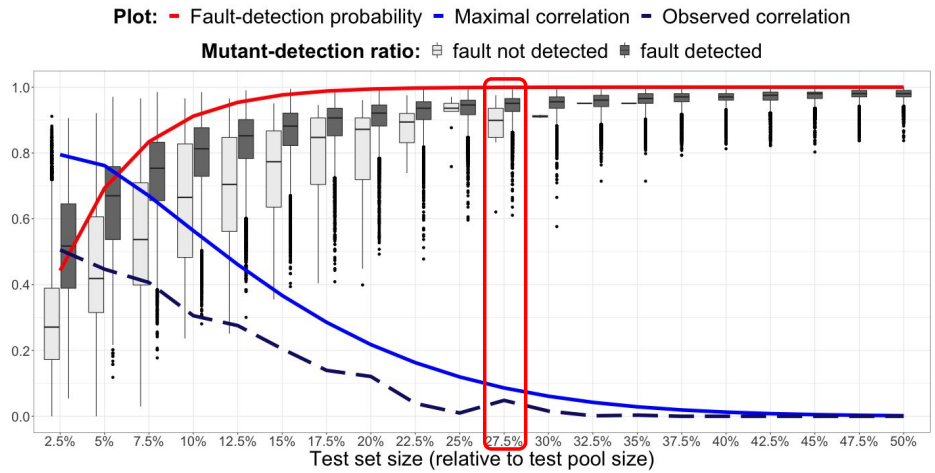
Fault detection
95.0%

Maximal point biserial correlation drops to 0.45

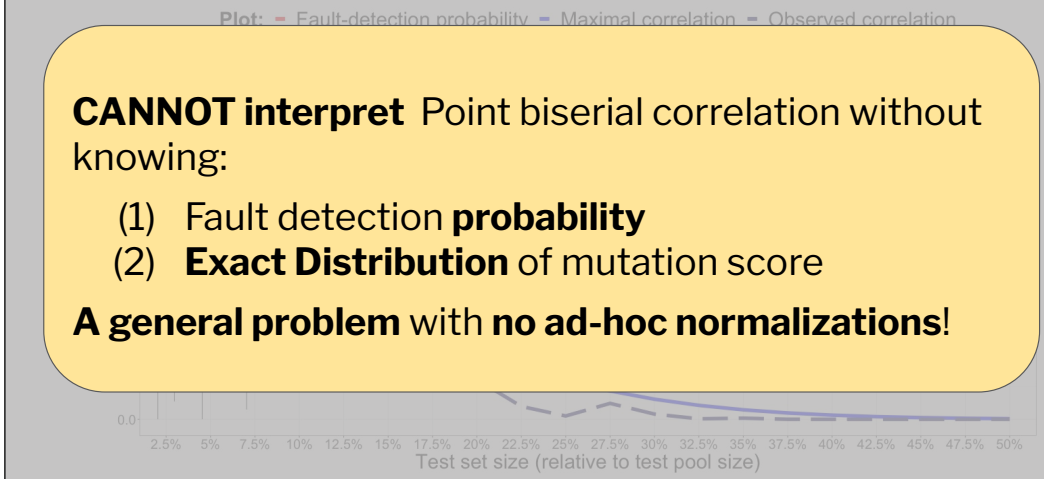
Random selection is prone to misleading conclusions!



Random selection is prone to misleading conclusions!



Random selection is prone to misleading conclusions!

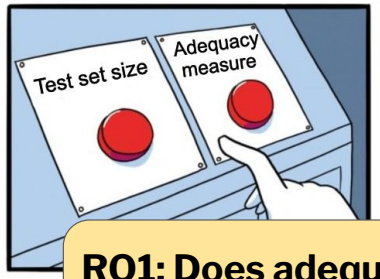


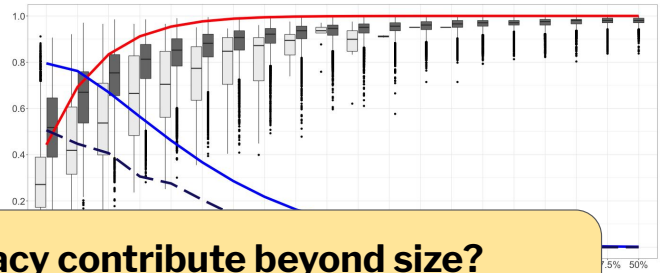
CANNOT interpret Point biserial correlation without knowing:

(1) Fault detection **probability**
(2) **Exact Distribution** of mutation score

A general problem with **no ad-hoc normalizations**!

What can we do to answer our research questions?





RQ1: Does adequacy contribute beyond size?

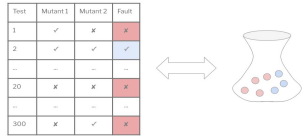
RQ2: Which adequacy measure is best?

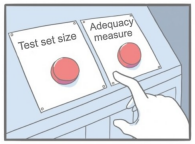
An ill-posed question
correlation doesn't fix that!

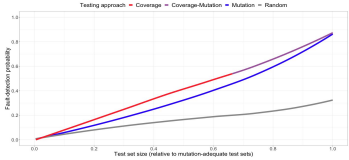
Class imbalance problem
correlation isn't what you think it is!

Outline

- Review of existing methods
- Ask the right (statistical) question
- Test adequacy measures are valid

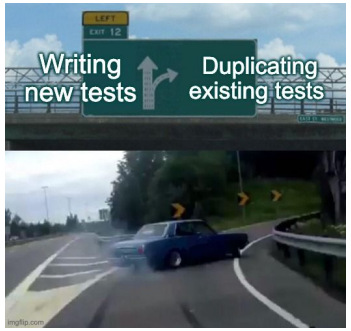






Random Selection is also conceptually flawed!

- Test set size is NOT a meaningful goal in practice!



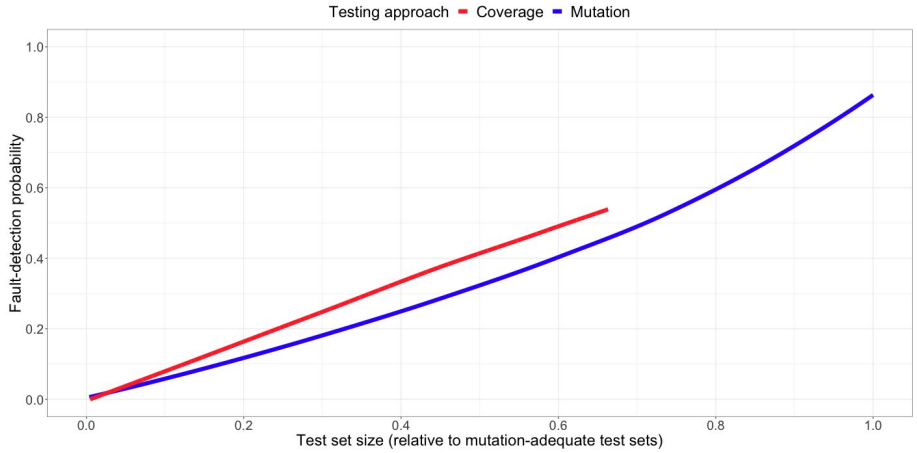
Alternative sets of experiments

- Address the conceptual issue
- Avoid the statistical pitfalls
- Account for test set size

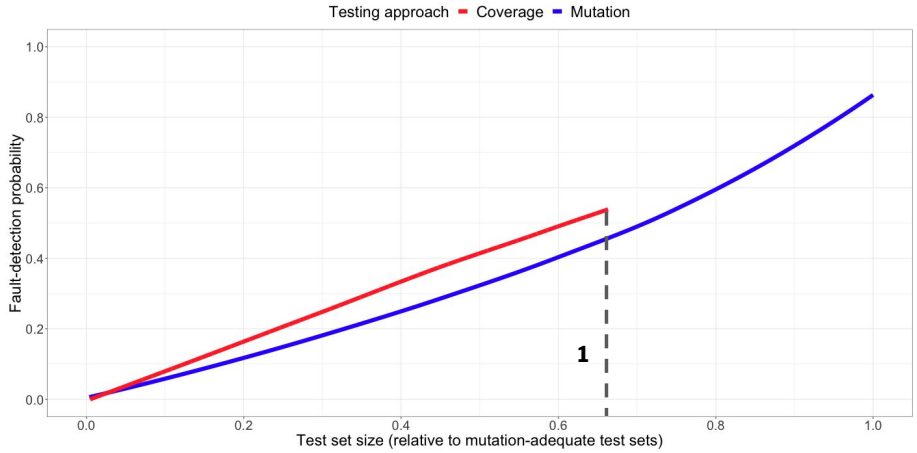
In a nutshell:

- Use adequacy-based testing to achieve a specified level (e.g., 80% coverage)

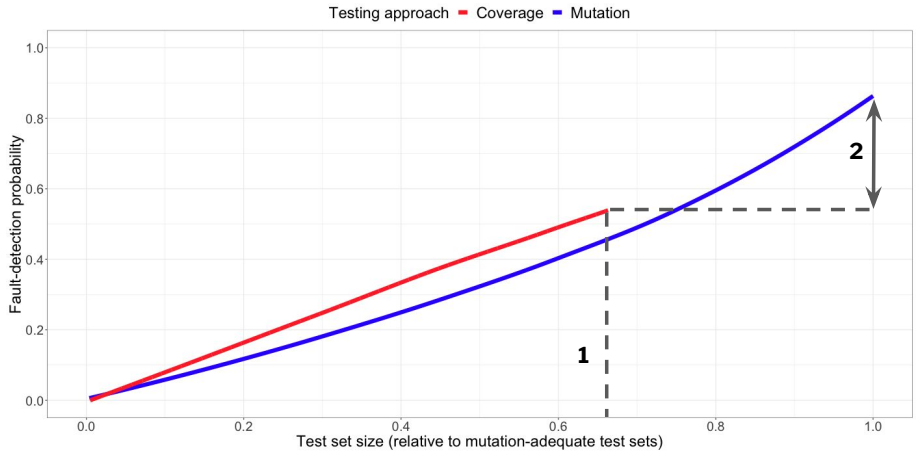
Statement coverage vs. Mutation score



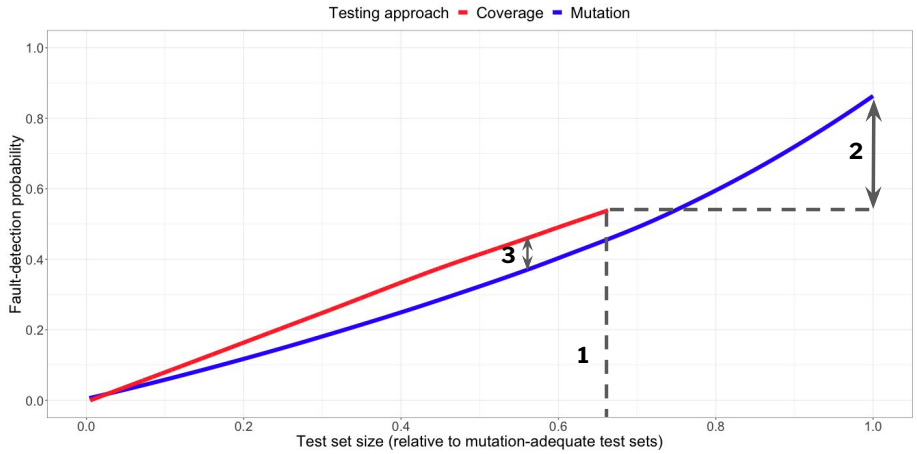
Statement coverage vs. Mutation score



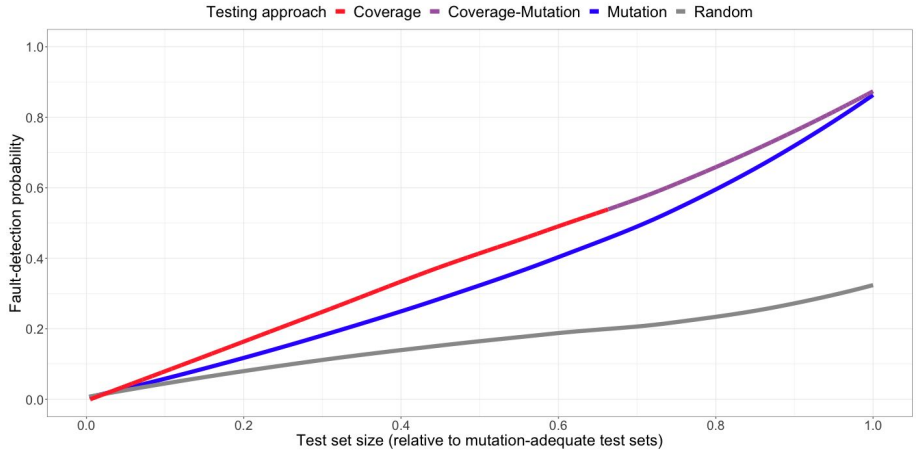
Statement coverage vs. Mutation score



Statement coverage vs. Mutation score



Statement coverage vs. Mutation score



(see also “State of Mutation Testing at Google”, Petrović and Ivanković (2018))

Conclusions

- Random selection is prone to misleading results.
- Mutation & coverage are VALID adequacy measures and contribute beyond just size.
- Want effective tests? Coverage + Mutation