CS 520
Theory and Practice of Software Engineering
Fall 2017

Experimental design and validity

October 12, 2017
Today

- The scientific method.
- Internal, external, and construct validity.
- Reasoning about two empirical studies.
- Paper discussion:
  Views on Internal and External Validity in Empirical Software Engineering
The scientific method

Question
The scientific method

Question → Observations
The scientific method

1. Question
2. Observations
3. Hypothesis
4. Predictions
5. Experiment
The scientific method

- Question
- Observations
- Hypothesis
- Experiment
- Predictions
The scientific method

1. Question
2. Observations
3. Hypothesis
4. Experiment
5. Predictions

Falsifiable
The scientific method

- Question
- Observations
- Hypothesis
- Predictions
- Experiment

- Falsifiable
- Repeatable
- Data collection and analysis!
The scientific method: common mistake

Question → Data collection → Observations → Hypothesis → Predictions → Data Analysis → Predictions
"If you torture the data long enough, it will confess."
(Ronald Harry Coase)
Internal, external, and construct validity

Internal validity

External validity

Construct validity
Internal, external, and construct validity

**Internal validity**
How well does the experimental design isolate the effect/variables that it studies (i.e., control for confounds)?

**External validity**
How well does the experimental design generalize to the real world (i.e., other populations, situations, etc.)?

**Construct validity**
How well does the experimental design measure what it is supposed to measure? Does it use the right metrics and collect the right measurements?
Internal validity: a classic example

**Internal validity**
How well does the experimental design isolate the effect/variables that it studies (i.e., control for confounds)?

**Classic example**
Murder rates and ice cream sales are highly positively correlated. Possible explanations?
Internal validity: a classic example

Internal validity
How well does the experimental design isolate the effect/variables that it studies (i.e., control for confounds)?

Classic example
Murder rates and ice cream sales are highly positively correlated. Possible explanations?

- Possibilities:
  - Resurrected zombies primarily feed off ice cream
  - Excessive ice cream consumption makes others jealous
Internal validity: a classic example

Internal validity
How well does the experimental design isolate the effect/variables that it studies (i.e., control for confounds)?

Classic example
Murder rates and ice cream sales are highly positively correlated. Possible explanations?

- Possibilities:
  - Resurrected zombies primarily feed off ice cream
  - Excessive ice cream consumption makes others jealous

Actually, the weather is a non-controlled confound!
Threats to validity: example experiment

Research question:
Does coffee consumption improve code quality?

Methodology
- I program on project 1 on Mondays with coffee.
- I program on project 2 on Fridays without coffee.
- Measure code quality in number of defects I encounter.
- Measure coffee consumption in dollars spent on coffee beans, as listed on my grocery-shopping receipt.
Threats to validity: example experiment

Research question:
Does coffee consumption improve code quality?

Methodology
● I program on project 1 on Mondays with coffee.
● I program on project 2 on Fridays without coffee.
● Measure code quality in number of defects I encounter.
● Measure coffee consumption in dollars spent on coffee beans, as listed on my grocery-shopping receipt.

What are threats to construct, internal, and external validity?
Another empirical study

Goal:
Studying the relationship between time spent on studying Java and success rate in completing coding assignment.

Methodology:
● 75 participants are randomly selected in front of LGRT.
● Each participant is given a high-level overview of the study.
● Each participant decides on how long to study before attempting to solve any coding assignment.
● Each participant solves as many coding assignments as possible in one hour (after studying).
Conclusion: Spending more time on learning Java makes you a worse Java programmer.
Overall results

Something is fishy… Is there a dead salmon in here somewhere?

Goal:
- Each group comes up with a testable hypothesis about the data.
- Each group comes up with 2 methodology questions.
Results per group (field of study)

This phenomenon is called: Simpson’s paradox.
Paper discussion

Views on Internal and External Validity in Empirical Software Engineering

High-level topics
- Internal validity
- External validity
- Construct validity

Open discussion
- Is there a tradeoff between internal and external validity?
- Should we maximize internal or external validity?
- How representative are students as developers?